

Physical Principles of Remote Sensing

Third Edition

Fully updated and with significant new treatments of photography, laser profiling and image processing, the third edition of this popular textbook covers a broad range of remote sensing applications and techniques across the Earth, environmental and planetary sciences. It focusses on physical and mathematical principles, giving students a deeper understanding of remote sensing systems and their possibilities, while remaining accessible to those with less mathematical training by providing less technical summaries of quantitative topics.

Key features

- Boxed examples and additional photos engage students and show them how the theory relates to real-world applications.
- Numerous colour images bring the subject to life.
- Section summaries, review questions and additional problems allow students to check their understanding of key concepts and practice handling real data for themselves.
- Review questions link out to supplementary online material, which includes freely available software, practical exercises, and animations.

W.G. Rees is a Senior Lecturer at the Scott Polar Research Institute, University of Cambridge, where he has taught and researched in the field of remote sensing for over 20 years. He has been active in developing and applying remote sensing methods to the mapping and monitoring of the polar regions, having conducted fieldwork in arctic regions of Europe and Asia, and in Svalbard. For the past few years he has been joint coordinator of PPS Arctic, a major programme to investigate the characteristics and behaviour of the arctic treeline as part of the International Polar Year, and he is also a member of the ISPRS (International Society for Photogrammetry and Remote Sensing) working group on LiDAR. Dr Rees has published several books on remote sensing, including the first and second editions of *Physical Principles of Remote Sensing* (1990, 2001, Cambridge University Press), *The Remote Sensing Data Book* (1999, Cambridge University Press) and *Remote Sensing of Glaciers* (with P. Pellikka, 2010, Taylor & Francis). He was made a Fellow of the Institute of Physics in 1996 and is a member of the Photogrammetry and Remote Sensing Society.

“this is a dummy quote this is a dummy quote”

- Reviewer 1, somewhere

“this is a dummy quote this is a dummy quote”

- Reviewer 2, somewhere

“this is a dummy quote this is a dummy quote”

- Reviewer 3, somewhere

Physical Principles of Remote Sensing

THIRD EDITION

W. G. REES

Scott Polar Research Institute
University of Cambridge



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press

The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9781107004733

© W. G. Rees 2012

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2012

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data

Rees, Gareth, 1959–

Physical principles of remote sensing / Gareth Rees. – 3rd ed.

p. cm.

ISBN 978-1-107-00473-3 (Hardback) – ISBN 978-0-521-18116-7 (Paperback)

1. Remote sensing. I. Title.

G70.4.R44 2012

621.36'78–dc23

2012018197

ISBN 978-1-107-00473-3 Hardback

ISBN 978-0-521-18116-7 Paperback

Additional resources for this publication at www.cambridge.org/rees
Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to
in this publication, and does not guarantee that any content on such
websites is, or will remain, accurate or appropriate.

*For Christine
as always*

CONTENTS

Preface	<i>page</i>	xiii
Acknowledgements	xvi	
1 Introduction	1	
1.1 A short history of remote sensing	2	
1.2 Applications of remote sensing	5	
1.3 A systems view of remote sensing	6	
1.4 Further reading, and how to obtain data	9	
2 Electromagnetic waves in free space	11	
2.1 Electromagnetic waves	11	
2.2 Polarisation	15	
2.3 Spectra and the Fourier transform	19	
2.4 The Doppler effect	24	
2.5 Describing angular distributions of radiation	25	
2.6 Thermal radiation	28	
2.6.1 Characteristics of solar radiation	34	
2.7 Diffraction	36	
Review questions	40	
Problems	40	
3 Interaction of electromagnetic radiation with matter	42	
3.1 Propagation through homogeneous materials	43	
3.1.1 Complex dielectric constants: absorption	44	
3.1.2 Dielectric constants and refractive indices of real materials	45	
3.1.3 Dispersion	49	
3.2 Plane boundaries	52	
3.3 Scattering from rough surfaces	56	
3.3.1 Description of surface scattering	57	
3.3.2 Simple models of surface scattering	59	
3.3.3 The Rayleigh roughness criterion	62	
3.3.4 Models for microwave backscatter	64	
3.4 Absorption and scattering by particles	73	
3.4.1 Very small particles ($\chi \ll 1$)	74	
3.4.2 Larger particles	76	
3.4.3 Absorption and scattering by atoms and molecules	78	
3.5 The radiative transfer equation	84	
3.5.1 Propagation through an absorbing medium	85	
3.5.2 Propagation through an absorbing and emitting medium	86	

3.5.3 A simple model of scattering and absorption: the two-stream approximation	89
3.5.4 Scattering, absorption and emission	93
3.6 Interaction of electromagnetic radiation with real materials	94
3.6.1 Visible and near-infrared region	95
3.6.2 Emissivities in the thermal infrared region	99
3.6.3 Emissivities in the microwave region	100
3.6.4 Effect of cloud and snow on microwave radiation	102
3.6.5 Microwave backscattering coefficients	103
3.6.6 Modelling microwave backscattering: case study of a snowpack	105
Review questions	107
Problems	108
4 Interaction of electromagnetic radiation with the Earth's atmosphere	110
4.1 Composition and structure of the gaseous atmosphere	110
4.2 Molecular absorption and scattering in the atmosphere	114
4.3 Particles in the atmosphere: aerosols	119
4.4 Fog and cloud	121
4.5 Rain and snow	125
4.6 The ionosphere	128
4.7 Atmospheric turbulence	131
Review questions	133
Problems	133
5 Photographic systems	135
5.1 Photographic film	135
5.1.1 Performance of photographic film: speed, contrast and spatial resolution	136
5.1.2 Digital photography	139
5.2 Photographic optics	141
5.2.1 Lens distortion	144
5.3 Photogrammetry and stereogrammetry	147
5.3.1 Relief displacement	149
5.3.2 Stereophotography	152
5.4 Atmospheric propagation	156
5.5 Some instruments	158
5.6 Applications of aerial and space photography	162
Review questions	162
Problems	163
6 Electro-optical systems	164
6.1 Visible and near-infrared imaging systems	164
6.1.1 Detectors	164
6.1.2 Imaging	167
6.1.3 Spatial resolution	170

6.1.4 Spectral resolution	171
6.1.5 Atmospheric propagation and correction	172
6.2 Types of VNIR imager	175
6.2.1 Very high resolution imagers	176
6.2.2 High resolution imagers	176
6.2.3 Medium resolution imagers	179
6.2.4 Low resolution imagers	179
6.2.5 Ocean colour imagers	180
6.2.6 Hyperspectral imagers	181
6.2.7 Geostationary imagers	182
6.3 Major applications of VNIR images	184
6.4 Thermal infrared imagers	188
6.4.1 Detectors	188
6.4.2 Thermal infrared imaging	189
6.4.3 Spatial resolution	189
6.4.4 Spectral resolution and sensitivity	190
6.4.5 Atmospheric propagation and correction	191
6.5 Types of TIR imager	194
6.5.1 High resolution TIR imager	194
6.5.2 Medium resolution TIR imager	194
6.5.3 Geostationary TIR imager	196
6.6 Major applications of thermal infrared images	197
6.6.1 Earth surface temperature	198
6.6.2 Thermal inertia	199
6.6.3 Cloud detection and monitoring	204
6.7 Atmospheric sounding	206
6.7.1 Temperature profiling from observations at nadir	207
6.7.2 Profiling of gas concentrations at nadir	210
6.7.3 Backscatter observations at nadir	211
6.7.4 Limb-sounding observations	211
6.7.5 Spectral resolution for atmospheric sounding observations	213
6.8 Some profiling instruments	216
Review questions	220
Problems	221
7 Passive microwave systems	223
7.1 Antenna theory	223
7.1.1 Angular response and spatial resolution	223
7.1.2 Sensitivity	229
7.1.3 Scanning radiometers	229
7.2 Applications of passive microwave radiometry	232
7.2.1 Oceanographic applications	232
7.2.2 Land surface applications	235
7.3 Atmospheric correction of passive microwave imagery	239
7.4 Examples: the SSMIS and the MSMR	241

Contents

7.5 Atmospheric sounding using passive microwave observations	243
Review questions	247
Problems	248
8 Ranging systems	250
8.1 Laser profiling	250
8.1.1 Scanning laser profilers	253
8.1.2 Waveform-resolving laser profiling	255
8.1.3 Atmospheric correction of laser profiler data	255
8.1.4 Applications of laser profiling	257
8.2 Radar altimetry	261
8.2.1 Simple model of the waveform	261
8.2.2 Effect of the Earth's curvature	265
8.2.3 Effect of coherence: range accuracy	266
8.2.4 Response from a rough surface	267
8.2.5 Applications of radar altimetry	269
8.2.6 Atmospheric and ionospheric correction of radar altimeter data	273
8.2.7 Example: the Envisat RA-2 radar altimeter	275
8.3 Other ranging systems	277
Review questions	278
Problems	278
9 Scattering systems	281
9.1 LiDAR	281
9.2 The radar equation	282
9.3 Microwave scatterometry	285
9.3.1 Applications of microwave scatterometry	287
9.3.2 Example: ASCAT	291
9.4 Real-aperture imaging radar	292
9.4.1 Image distortions	294
9.4.2 Instruments and applications	296
9.5 Synthetic aperture radar	297
9.5.1 More exact treatment of the azimuth resolution	300
9.5.2 Speckle	301
9.5.3 Distortions of SAR images	304
9.5.4 Limitations imposed by ambiguity	306
9.5.5 SAR interferometry	307
9.5.6 Major applications of radar imaging	312
9.5.7 Example: Radarsat-2	314
Review questions	316
Problems	317
10 Platforms for remote sensing	318
10.1 Aircraft	318
10.2 Satellites	320

10.2.1	Launch of satellites	321
10.3.	Description of the satellite orbit	325
10.3.1	Effects of the Earth's asphericity	329
10.3.2	Special orbits	331
10.4	Satellite station-keeping and orbital manoeuvres	343
	Review questions	346
	Problems	346
11	Data processing	348
11.1	Transmission and storage of data	348
11.2	Image processing	351
11.2.1	Preprocessing	352
11.2.2	Image enhancement	360
11.2.3	Band transformations	371
11.3	Image classification	380
11.3.1	Density slicing and pseudocolour display	381
11.3.2	Multispectral classification	381
11.3.3	Hyperspectral classification	385
11.3.4	Advanced classification methods	386
11.3.5	Sub-pixel classification	388
11.3.6	Texture classification	389
11.3.7	Error matrices and classification accuracy	390
11.4	Image segmentation and detection of geometrical features	394
11.4.1	Segmentation	394
11.4.2	Detecting shapes	395
11.5	Geographic information systems	402
11.6	Image formats and data compression	405
11.6.1	Image compression	405
11.6.2	Image formats for remote sensing	406
	Review questions	409
	Problems	410
	Appendix Data tables	412
A.1	Physical constants	412
A.2	Units	412
A.3	Illuminance at the Earth's surface	413
A.4	Properties of the Sun and Earth	414
A.5	Position of the Sun	414
	References	417
	Index	426
	See colour plates section between pages xxx and xxx	

PREFACE

There are many books that explain the subject of remote sensing to those whose backgrounds are primarily in the environmental sciences. This is an entirely reasonable fact, since they continue to be the main users of remotely sensed data. However, as the subject grows in importance, the need for a significant number of people to understand not only what remote sensing systems do, but how they work, will grow with it. This was already happening in 1990, when the first edition of *Physical Principles of Remote Sensing* appeared, and since then increasing numbers of physical scientists, engineers and mathematicians have moved into the field of environmental remote sensing. It is mainly for such readers that this book, like its previous editions, has been written. That is to say, the reader for whom I have imagined myself to be writing is educated to a reasonable standard (although not necessarily to first degree level) in physics, with a commensurate mathematical background. I have however found it impossible to be strictly consistent about this, because of the wide range of disciplines within and beyond physics from which the material has been drawn, and I trust that readers will be understanding when they find the treatment either too simple or over their heads.

This book attempts to follow a logical progression, more or less following the flow of information from the remotely sensed object to the user of the data. The first four chapters lay the general foundations. Chapter 1 sets the subject in context. Chapter 2 is a non-rigorous treatment of electromagnetic wave propagation in free space, which can be regarded as a compendium of necessary results. It will represent, I hope, mostly revision to most readers, although it assumes little or no previous knowledge of Fourier transforms or of Fraunhofer diffraction theory. Chapter 3 discusses the interaction of electromagnetic radiation with smooth and rough surfaces and with inhomogeneous materials such as soil and snow, and Chapter 4 discusses the interaction of radiation with the atmosphere and ionosphere. By this stage of the book, our information is, as it were, travelling upwards towards the sensor. Chapters 5 to 9 discuss the sensors themselves, beginning with the more familiar passive sensors and going on to consider active systems. These chapters explain, so far as is consistent with the level of the book, the functioning of the sensors, important operational constraints, and some of the more important applications derived from them. These chapters also include brief descriptions of real instruments on existing or forthcoming satellite missions. The platforms on which the sensors are supported are discussed in Chapter 10. After a short discussion of remote sensing from aircraft, the chapter devotes itself to satellite orbits. Finally, Chapter 11 presents an introduction to the data processing aspects of remote sensing, particularly digital image processing and analysis. An appendix contains tables of data frequently needed in remote sensing. A short list of problems or exercises is included at the end of most chapters. Most of these problems are reasonably straightforward (I have tried to indicate which are for ‘enthusiasts’), designed to extend and consolidate the reader’s understanding of the material. Some problems will require material from more than one chapter for

their solution. The problems are of a more or less ‘academic’ format, many of them having originated as exercises for students.

It will perhaps be useful to indicate those features of the book that have been preserved from the second edition and those that are new. The underlying rationale has not changed. It has still been my intention to keep the book as short as possible, consistent with clarity, although this edition is significantly longer than the second because it includes new material. In particular, the treatment of stereophotography (Chapter 5), laser profiling (Chapter 8), synthetic aperture radar interferometry (Chapter 9), and digital image processing (Chapter 11) have all been significantly expanded. As before, the aim has been to teach principles of remote sensing rather than to present a lot of technical or engineering detail. However, the inclusion of brief discussions of real sensor systems or surveys of types of sensor, introduced in the second edition, has been continued, expanded and updated, particularly in Chapter 6 which deals with visible-wavelength and infrared systems. The book’s bibliography has been brought up to date, although I have still attempted to keep it short enough so as not to overwhelm the reader with an enormous list of references. Some selection and omission has therefore been necessary, and I hope my colleagues will forgive me if my selection does not tally with theirs. One particular goal in compiling the bibliography has been to include enough recent references to allow the reader efficiently to find his or her way into the modern literature. As in the first edition, I have deliberately avoided the rigorous consistency in the use of *symbols* that demands that a given symbol be used to represent only a single physical quantity. Because of the wide scope of remote sensing, this would lead to an unforgivably confusing proliferation of symbols with many sub- and superscripts. Consistency of symbols is therefore confined to sections of the text that deal with a single topic, except for a few ‘universal’ symbols such as h for the Planck constant and ω for angular frequency, which are used throughout the book. SI units are used consistently, although a table in the appendix gives equivalents for some common non-SI units.

There are some other important changes in this edition. The number of illustrations has increased by about a third, with much greater use of colour. This is especially valuable in the understanding of satellite imagery. As aids to understanding the ideas presented in the book, each chapter apart from the introductory Chapter 1 now includes more or less non-technical summaries and a set of review questions. The summaries, which are presented in boxes at the end of each major section, can be read consecutively as a 12 000 word outline of the whole book, but are probably most usefully read where they have been placed, at the end of each section, where they can be used to check the reader’s understanding of the main points without getting lost in mathematical detail. The review questions can also be used as an aid to comprehension of the material: when teaching university courses based on some of this material I have found it helpful to ask students to prepare five-minute responses to such review questions, which they can then present to the class.

One difficulty that I found in the first two editions of the book was that of clearly explaining the behaviour of some time-dependent phenomena. Examples of this include the propagation of electromagnetic radiation, the orbit of a satellite around the Earth, and the manner in which the pulse of radiation from a radar altimeter interacts with a rough surface, although there are others. I have developed some computer animations to illustrate such phenomena, and these are accessible on the book’s website – another innovation for this edition. They are indicated by a red ‘A’ in the margin. The website also

provides hints and solutions to the problems, useful links, and a repository of computer programs that can be used to explore some of the ideas in the book. The problems included at the end of each chapter are supplemented, in the website, by suggested practical exercises. Particular emphasis has been placed on exercises designed to consolidate the reader's understanding of the main ideas in Chapter 11. It is expected that all of the web-based material will evolve over time.

This book arose from a course of undergraduate lectures delivered first at the Scott Polar Research Institute, and later at the Cavendish Laboratory, both in the University of Cambridge. I am grateful to both departments for letting me try out my ideas. Many people are owed thanks for their contributions to the writing of the first edition. It is difficult to single out individuals, but I particularly wish to thank Andrew Cliff, Bernard Devereux, Michael Gorman, Christine Rees and Michael Rycroft. Subsequently I have developed some of the concepts in the book for teaching to audiences with a wider range of background knowledge than the book's intended readership, including undergraduates in the Department of Geography at the University of Cambridge and at the Geography Faculty of Moscow State University, and children as young as six years old. Much of the credit for any improvements that I may have made to the book since the first edition lies with the constructive criticisms of the users and reviewers of the first edition and of the many graduate and undergraduate students who I have had the pleasure of working with since 1990, as well as with my professional colleagues. In particular, I thank Neil Arnold, Olga Tutubalina and Sophie Weeks. As always, Cambridge University Press has provided advice and encouragement whenever it was needed.

ACKNOWLEDGEMENTS

Permission to reproduce copyright and other material from the following sources is gratefully acknowledged: Colorado University, Dundee Satellite Receiving Station, the European Space Agency, GeoEye, International Space Company Kosmotras, the Japanese Aerospace Exploration Agency, Jim Doty, the National Aeronautics and Space Administration (USA), the National Geospace Intelligence Agency (USA), the National Oceanic and Atmospheric Administration (USA), the National Snow and Ice Data Center (USA), the Natural Environment Research Council (UK), Nauchnyy Tsentr Operativnogo Monitoringa Zemli (Research Centre for Earth Operative Monitoring, Russia), the Royal Society, the University of Alabama in Huntsville, and the University of Cambridge. Some illustrations are derived from public-domain internet resources. I have been particularly grateful for the fact that the United States Copyright Law ensures that a huge amount of material produced by NASA, NOAA and other organisations can be reproduced without infringing copyright.

The text, all of the animations and many of the illustrations were prepared using free software, principally OpenOffice (for text processing), Zotero (for management of references), ImageJ and MultiSpec (for image processing), QGIS (for geographic information system processing), Plot and Veusz (for preparing graphs) and GNU Octave (for many tasks).

1

Introduction

'Remote sensing' is, broadly but logically speaking, the collection of information about an object without making physical contact with it. (The term was coined by Evelyn Pruitt of the US Office of Naval Research in the 1950s.) This is a simple definition, but too vague to be really useful (Campbell 2008), so for the purpose of this book we restrict it by confining our attention to the Earth's surface and atmosphere, viewed from above using electromagnetic radiation. This narrower definition excludes such techniques as seismic, geomagnetic and sonar investigations, as well as (for example) medical and planetary imaging, all of which could otherwise reasonably be described as remote sensing, but it does include a broad and reasonably coherent set of techniques, nowadays often described by the alternative name of *Earth observation*. These techniques, which now have a huge range of applications in the 'civilian' sphere as well as their obvious military uses, make use of information impressed in some way on electromagnetic radiation ranging from ultraviolet to radio frequencies.

One important casualty of our restricted definition of remote sensing is the use of spaceborne methods of measuring the Earth's gravitational field. Although observations from artificial Earth satellites have been used since the 1970s to measure the Earth's gravity, our current (at the time of writing, in 2012) ability in this regard is a remarkable indication of the level of space technology. This is the GRACE (Gravity Recovery and Climate Experiment) mission, launched in 2002. Two satellites, each with a mass of around half a tonne, follow the same orbit 500 km above the Earth's surface. They are approximately 220 km apart, and the distance between them is constantly monitored with an accuracy of 10 µm. This distance changes as the satellites cross regions of different gravitational field strength. The GRACE system is sensitive enough to respond to changes in groundwater in a large river basin. Data from the GRACE mission are described in Chapter 8.

1.1

A short history of remote sensing

The origins of remote sensing can plausibly be traced back to the fourth century BC and Aristotle's *camera obscura* (or, at least, the instrument described by Aristotle in his *Problems* but perhaps known even earlier). Although significant developments in the theory of optics

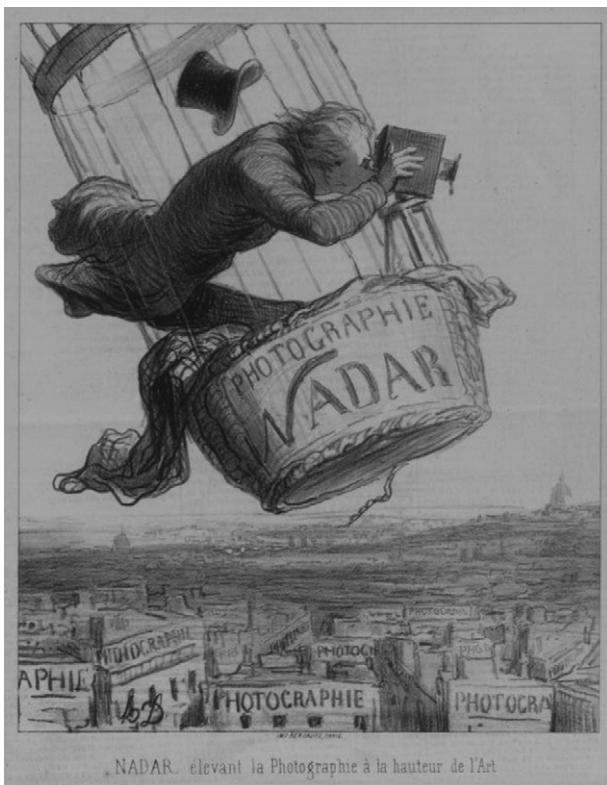


Figure 1.1. ‘Nadar raising Photography to the level of Art’. Lithograph by Honoré Daumier, 1863. (Source: Brooklyn Museum, via Wikipedia. http://en.wikipedia.org/wiki/File:Brooklyn_Museum_-_Nadar_Élevant_la_Photographie_à_la_Hauteur_de_l'Art_-_Honoré_Daumier.jpg)

began to be made in the seventeenth century, and glass lenses were known much earlier than this, the first real advance towards our modern conception of remote sensing came in the first half of the nineteenth century with the invention of photography. For the first time, it became possible to record an image permanently and objectively. Also during the nineteenth century, forms of electromagnetic radiation were discovered beyond the visible part of the spectrum – infrared radiation by Herschel, ultraviolet by Ritter, and radio waves by Hertz – and in 1863 Maxwell developed the electromagnetic theory on which so much of our understanding of these phenomena depends.

Airborne photography followed almost immediately on the discovery of the photographic method. The first aerial photograph, unfortunately no longer in existence, was probably made in 1858 by Gaspard-Félix Tournachon, from a balloon at an altitude of about 80 m. Tournachon, born in 1820, used the pseudonym ‘Nadar’ and was one of the earliest developers of photography as art (Figure 1.1). He was also the inventor of the crowd control barrier. Kites were also soon used, and by 1890 the usefulness of aerial photography was so far recognised that Arthur Batut had published a textbook on the subject (Batut 1890) (Figure 1.2).

1.1 A short history of remote sensing



Figure 1.2. Labruguière, photographed from a kite by Arthur Batut in 1889. (Source: Wikipedia. http://en.wikipedia.org/wiki/File:Labruguière_photographed_by_Arthur_Batut%27s_kite_in_1889.jpg)

The next step towards what we now recognise as remote sensing was taken with the development of practicable aeroplanes in the early twentieth century. Again, the potential applications were quickly recognised and aerial photographs were recorded from aeroplanes from 1909. Airborne photography was used during the First World War for military reconnaissance, and during the period between the two World Wars civilian uses of this technique began to be developed, notably in cartography, geology, agriculture and forestry. Cameras, film and aircraft underwent significant improvements, and stereographic mapping attained an advanced state of development. The modern descendants of these applications are discussed in Chapter 5. Also during this period, John Logie Baird, the inventor of television, performed early work on the development of airborne scanning systems capable of transmitting images to the ground. This work (its modern developments are discussed in Chapter 6) was highly confidential, having been carried out on behalf of the French Air Ministry. It was ended by the war and forgotten about until 1985 (Burns 2000).

The Second World War brought substantial developments to remote sensing. Photographic reconnaissance reached a high state of development – the German invasion of Britain, planned for September 1940, was forestalled by the observation of concentrations of ships along the English Channel. Infrared-sensitive instruments and radar systems were developed. In particular, the Plan Position Indicator used by night bombers was an imaging radar that presented the operator with a ‘map’ of the terrain, and thus represented the ancestor of the imaging radar systems discussed in Chapter 9.

By the 1950s, false-colour infrared film, originally developed for military use, was finding applications in vegetation mapping, and high resolution imaging radars were being developed. As these developments continued through the 1960s, sensors began to be placed in space. This was originally part of the programme to observe the Moon, but the advantages of applying the same techniques to observation of the Earth were soon recognised, and the first multispectral spaceborne imagery of the Earth was acquired from

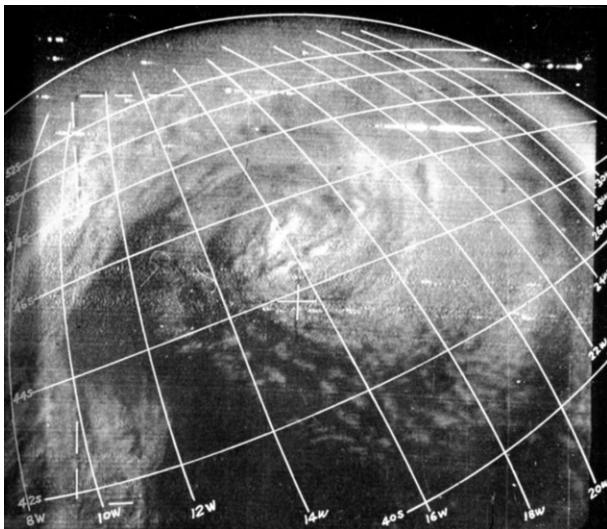


Figure 1.3. The first image transmitted to Earth electronically from a satellite. TIROS-1 (the name is an acronym for Television Infrared Observation Satellite) was the first successful weather satellite. (Source: NASA via Wikipedia. <http://en.wikipedia.org/wiki/File:TIROS-1-Earth.png>)

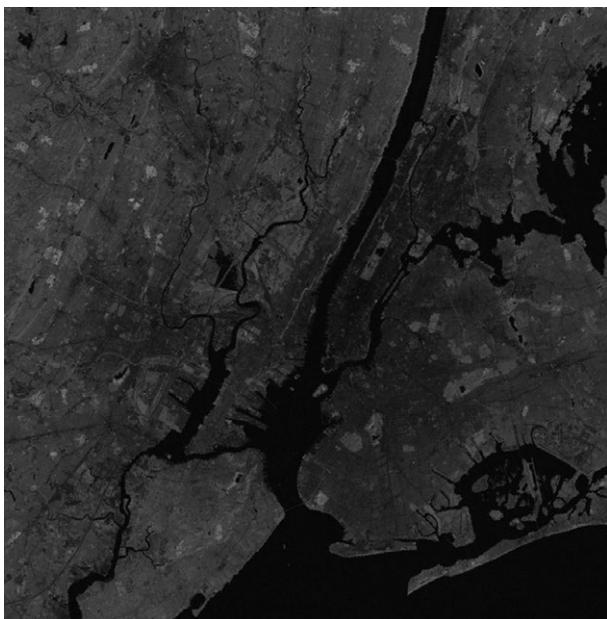


Figure 1.4. Extract of ERTS (Landsat-1) Multispectral Scanner image of New York, 10 October 1972.

1.2 Applications of remote sensing

Apollo 6. Although there were earlier unmanned remote sensing satellites (Figure 1.3), the opening of the modern era of spaceborne remote sensing ought probably to be dated to July 1972 with the successful operation of ERTS, the Earth Resources Technology Satellite, by the US National Aeronautics and Space Administration (NASA) (Figure 1.4). ERTS was renamed Landsat-1, and the Landsat programme is still continuing – at the time of writing (2012), Landsat-5 and Landsat-7 were both still operational.

Since the launch of ERTS in 1972, the number and diversity of spaceborne and airborne remote sensing systems has grown dramatically. A larger range of variables can be measured, and consistent and systematic datasets can be constructed for progressively longer periods of time. The explosive growth in the quantity of data being generated has been matched by growth in the availability of computing resources and the facilities for data storage and transmission. Since 2005, the availability of the *Google Earth* program has greatly increased public exposure to and use of remote sensing data.

1.2

Applications of remote sensing

The enormous growth in the availability of remotely sensed data over the last four decades has been matched by a fall in the real cost of the data. Nevertheless, it is still clear that use of the data must offer some tangible advantages to justify the cost of acquiring and analysing them. These advantages derive from a number of characteristics of remote sensing. Probably the most important of these is that data can be gathered from a large area of the Earth's surface, or a large volume of the atmosphere, in a short space of time, so that a virtually instantaneous 'snapshot' can be obtained. For example, scanners carried on geostationary meteorological satellites such as METEOSAT can acquire an image of approximately one-quarter of the Earth's surface in less than half an hour. When this aspect is combined with the fact that airborne or spaceborne systems can acquire data from locations that would be difficult (slow, expensive, dangerous, politically inconvenient ...) to measure *in situ*, the potential power of remote sensing becomes apparent. Of course, further advantages derive from the fact that most remote sensing systems generate calibrated digital data that can be fed straight into and analysed by a computer.

Remote sensing finds a very wide range of applications, naturally including the area of military reconnaissance in which many of the techniques had their origins. In the non-military sphere, most applications can loosely be categorised as 'environmental', and we can distinguish a range of environmental variables that can be measured. In the atmosphere, these include temperature, precipitation, the distribution and type of clouds, wind velocities, and the concentrations of gases such as water vapour, carbon dioxide, ozone etc. Over land surfaces, we can measure tectonic motion, topography, temperature, albedo (reflectance) and soil moisture content, and determine the nature of the land cover in considerable detail, for example by characterising the type of vegetation and its state of health or by mapping man-made features such as roads and towns. Over ocean surfaces, we can measure the temperature, topography (from which the Earth's gravitational field, as well as ocean tides and currents, can be inferred), wind velocity, wave energy spectra and colour (which is often related to biological productivity by plankton). The 'cryosphere', that part of the Earth's surface covered by snow and ice, can also be studied, giving data on the distribution, condition and dynamical behaviour of snow, sea ice, icebergs, glaciers and ice sheets.

This list of measurable variables, while not complete, is large enough to indicate that there is a correspondingly large number of disciplines to which remote sensing data can be applied. While by no means exhaustive, a list of applications could include the following disciplines: agriculture and crop monitoring, archaeology, bathymetry, cartography, civil engineering, climatology, coastal erosion, disaster monitoring and prediction, forestry, geology, geomorphology, glaciology, meteorology, oceanography, pollution monitoring, snow resources, soil characterisation, urban mapping, and water resource mapping and monitoring. It is not really possible to present a detailed cost–benefit analysis in this introduction. The development, insertion into orbit, and operation of a large remote sensing satellite costs typically a couple of billion euros. Perhaps it is sufficient to point out that the data available from remote sensing, particularly from spaceborne observations, can often not be obtained in any other way, that our current understanding of the global climate system is very largely based on spaceborne observations, and that the use of remotely sensed data for disaster warning has already saved many thousands of human lives.

As implied in Section 1.1, the era of systematic observation of the Earth from space is approaching middle age when seen from the perspective of a human lifespan. Landsat images have been collected continuously for over 40 years, radar altimeter data (which can be used to study changes in sea level, amongst many other applications) have been collected for over 20 years, and other examples can easily be found. These time-scales are long enough to form the basis of increasingly reliable measurements of change in many areas, not least of which is global change.

1.3

A systems view of remote sensing

We stated above, rather briefly, that remote sensing involves the collection of information, carried by electromagnetic radiation, about the Earth's surface or atmosphere. Let us try to expand this statement a little.

First, where does the radiation come from? One major classification of remote sensing systems is into the passive systems, which detect naturally occurring radiation, and the active systems, which emit radiation and analyse what is sent back to them. The passive systems can be further subdivided into those that detect radiation emitted by the Sun (this radiation consists mostly of ultraviolet, visible and near-infrared radiation), and those that detect the thermal radiation that is emitted by all objects that are not at absolute zero (i.e. all objects). For objects at typical terrestrial temperatures, this thermal emission occurs mostly in the infrared part of the spectrum, at wavelengths of the order of $10\text{ }\mu\text{m}$ (the so-called thermal infrared region), although measurable quantities of radiation also occur at longer wavelengths, as far as the microwave part of the spectrum. Active systems can, in principle, use any type of electromagnetic radiation. In practice, however, they are restricted by the transparency of the Earth's atmosphere. This is shown schematically in Figure 1.5. Chapter 4 presents a detailed discussion of the interaction of electromagnetic radiation with the atmosphere.

Figure 1.5 shows that there are three main ‘windows’ in the atmosphere. The first of these includes the visible and near-infrared (VNIR) parts of the spectrum, between wavelengths of about $0.3\text{ }\mu\text{m}$ and $3\text{ }\mu\text{m}$, although it does also contain a number of opaque

1.3 A systems view of remote sensing

Table 1.1. A simple taxonomy of remote sensing systems, excluding sounding instruments. The numbers in parentheses refer to the chapters of this book

		Active systems	
Passive systems		Ranging	Imaging
VNIR	Aerial photography (5) Electro-optical systems (6)	Laser profiler (8)	
TIR	TIR imager (6)		
Microwave	Passive microwave radiometer (7)	Radar altimeter (8) Ground-penetrating radar (8)	Microwave scatterometer (9) Imaging radar (9)

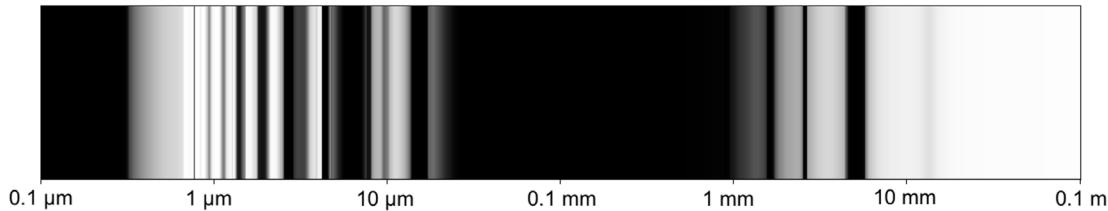


Figure 1.5. Transparency of the Earth's atmosphere as a function of wavelength (schematic). Black regions are opaque, white regions transparent.

regions. The second is a rather narrow region between about $8\text{ }\mu\text{m}$ and $15\text{ }\mu\text{m}$, in which is found the bulk of the thermal infrared (TIR) radiation from objects at typical terrestrial temperatures. The third more or less corresponds to the microwave region, between wavelengths of a few millimetres and a few metres. Thus we can expect that any active system designed to penetrate the Earth's atmosphere will operate in one of these three 'window' regions.

Table 1.1 summarises the main types of remote sensing system on the basis of the classifications we have just outlined. Sounding instruments, intended to profile some property of the atmosphere such as its temperature or chemical composition, generally operate on the boundaries between transparency and opacity of the atmosphere.

The sensor, whether it is part of a passive or an active instrument, detects electromagnetic radiation after it has interacted with or been emitted by the 'target' material. In what way can this radiation contain useful information about the target? There are essentially only two variables to describe the radiation that is received: *how much* radiation is detected (we may need to qualify this with a statement about the polarisation of the radiation), and *when* does it arrive? The time-structure of the detected radiation is obviously only relevant in the case of active systems where the time-structure of the emitted radiation can be controlled. In this case it is possible to determine the distance from the sensor to the target, and this is the principle behind various ranging systems such as the laser profiler, lidar, radar altimeter and other types of radar system. In all other

cases, the only information we have is the quantity of radiation received at the sensor. If the radiation arises from thermal emission, the quantity is characteristic of the temperature of the target material and its emissivity, a property that describes its efficiency at emitting thermal radiation. Otherwise (in the cases of both passive and active systems that measure reflected radiation) the amount of radiation that is received is determined by the amount illuminating the target material, and the target's reflectivity. Thus we can see that the information about a target material that is directly observable from remote sensing observations is actually rather limited: we can measure its range, its reflectivity, and a combination of its temperature and emissivity. However, these can be measured at different times, over a range of wavelengths and, sometimes, in different polarisation states, and this increase in the diversity of the variables at our disposal is responsible for the large range of indirect observables that was sketched out in Section 1.2.

The foregoing discussion has not included the effects of the Earth's atmosphere, except to point out that atmospheric opacity limits the scope for observing the Earth's surface to the two main atmospheric 'windows'. In fact, as almost any electromagnetic wave propagates through the atmosphere, its characteristics will be somewhat modified. This modification may be troublesome, requiring correction, or advantageous depending on whether we are more interested in studying the Earth's surface or the atmosphere itself. In general, we can say that if the observation is made at a wavelength at which the atmosphere is opaque, the measured signal will be characteristic of the atmosphere, whereas if the atmosphere is transparent, the data will be characteristic of the surface below.

Once the data have been collected by the sensor, they must be retrieved and analysed. In many, though not all, cases, the data will form an image, by which we mean a two-dimensional representation of the two-dimensional distribution of radiation intensity. Since the second edition of this book appeared in 2001, the concepts of digital imaging and digital images have become much more generally familiar as a result of the huge increase in the popularity of digital photography and the use of *Google Earth* and similar web-based resources. The images with which we have to deal in remote sensing are normally digital, so they can conveniently be analysed by computer, and need not be confined to the visible part of the electromagnetic spectrum. For example, an image might represent the radar reflectivity in one or more frequencies or polarisation states, or the thermal emission, as well as the visible or near-infrared reflectivity. Image processing forms an integral part of remote sensing. Typically, this involves several steps. The first is to correct the image so that it has a known geometrical correspondence to the Earth's surface and a known calibration, with atmospheric propagation effects removed. At this stage, the image may also be enhanced in various ways, for example by suppressing noise, to increase its intelligibility. The major goal of image processing, however, is the extraction of useful information from the sensor data, based on the brightness values of the image (probably in a number of spectral bands, at a number of different dates, in different polarisation states etc.) and also on the spatial context. Using the analogy of a colour photograph, we can say that information can be extracted on the basis of colour, texture, shape and spatial context. In the majority of cases, it is necessary or at least desirable to 'train' the process of extracting information from the image using data from known locations. The process can therefore be seen as one of extrapolation from areas that are already known, for example on the basis of field work, to much wider areas. The

1.4 Further reading, and how to obtain data

extrapolation need not be confined to the spatial domain, however, and the analysis of time-series of images for change detection is also an important application of remote sensing.

1.4

Further reading, and how to obtain data

The field of remote sensing is now well served with textbooks, and the interested (or puzzled) reader should be able to find alternative treatments of most of the topics discussed in this book. While what follows is a personal list and makes no pretence to completeness, recent general textbooks include the fourth edition of Campbell's *Introduction to Remote Sensing* (Campbell 2008) and the sixth edition of Lillesand, Kiefer and Chipman's *Remote Sensing and Image Interpretation* (Lillesand, Kiefer and Chipman 2008), while the *SAGE Handbook of Remote Sensing* (Warner, Nellis and Foody 2009) provides exceptional breadth of treatment. Somewhat more detailed or more specialised treatments are given by, for example, the second edition of Jensen's *Remote Sensing of the Environment* (Jensen 2006), Jones and Vaughan's *Remote Sensing of Vegetation* (Jones and Vaughan 2010), Purkis and Klemas's *Remote Sensing and Global Environmental Change* (Purkis and Klemas 2011), Liang's *Quantitative Remote Sensing of Land Surfaces* (Liang 2004) and the second edition of Elachi and van Zyl's *Introduction to the Physics and Techniques of Remote Sensing* (Elachi and van Zyl 2006).

Scientific journals also represent an important source of information. Articles in the scientific literature are usually aimed at specialists, but the more general reader can often also extract a useful understanding from them, and the journals sometimes also publish review articles. In this book I have provided references to both books and journal articles. The principal English language journals in remote sensing are the *Canadian Journal of Remote Sensing*, *Computers & Geosciences*, *Geophysical Research Letters*, *IEEE Transactions on Geoscience and Remote Sensing*, the *International Journal of Remote Sensing*, *Photogrammetric Engineering and Remote Sensing*, and *Remote Sensing of Environment*.

Finally, a few remarks about the Internet may be useful. This can represent a very powerful means of obtaining up-to-date information of all sorts, for example on the operational status of a particular remote sensing satellite, or the latest results from a research group, or access to remote sensing data (indeed, some of the illustrations used in this book have been obtained in this way) or the software needed to process it. As anyone who has grappled with the Internet will know only too well, the problem is usually to locate the information one needs. The well-known search engines can be extremely helpful, as can the collections of links assembled by public-spirited individuals and organisations. The website of this book is located at www.cambridge.org/9781107004733 and it includes a collection of links to some other useful websites including online catalogues from which satellite data may be located.

A question that inevitably arises when one starts to consider obtaining satellite or other remotely sensed imagery is – what does it cost? There is no single, simple answer to this question. As a rough guide, though, one may say that the coarser the spatial resolution of the imagery the more likely it is to be freely available, and the finer the spatial resolution the more likely it is to be rather expensive, by which one might mean a thousand pounds or equivalent. As an example, imagery from a particular commercial spaceborne

instrument, having a spatial resolution of around 1 m, cost US\$10–20 per square kilometre in late 2010, while MODIS imagery, with a spatial resolution of 250–1000 m, was freely available. Images collected by the Landsat series of satellites, having spatial resolutions of 15–80 m, were also freely available at the time of writing. The trend over the last decade or so has been towards the free availability of satellite imagery.

As noted in Section 1.3, quantitative processing of spatial data is an integral part of remote sensing. Some of the mathematical principles of image processing are discussed in Chapter 11. There is of course a wide range of computer software for manipulating and processing image data, often very powerful and flexible. However, such software can also manifest some disadvantages. It may be expensive, demanding of computing resources (memory, processing power and so on), able to run only on specific computing platforms, dependent on its own specific file formats, and so on. In contrast to this approach is the existence of free software, especially when developed in an open-source environment. The website of this book has links to sources of free software for image processing and for geographic information systems (GIS), a related technology that is discussed briefly in Chapter 11. In addition, readers of this book might wish to develop their own software for carrying out image processing operations. Again, there are several suitable possibilities among programming languages. Of freely available programming languages, two of the most useful are *R* and *GNU Octave*, both available since the early 1990s. *GNU Octave* is designed particularly for manipulating matrices, and since images share many properties with matrices (see Chapter 11) it is particularly well suited to processing images. *GNU Octave* runs under most computer operating systems. The book's website has a number of *Octave* programs that the reader can download and use to explore some of the topics discussed in the text, and readers are encouraged to submit their own programs.

2

Electromagnetic waves in free space

In Chapter 1 we noted that electromagnetic radiation is fundamental to remote sensing as we have defined it: the information about the sensed object is carried by this radiation. We therefore need to develop an understanding of the essential characteristics of this radiation and of how it interacts with its surroundings. This is a large topic and it is covered in this chapter and the next two. In this chapter we consider electromagnetic radiation in its *simplest* form, when it is propagating in (travelling through) a vacuum, usually termed ‘free space’. This is practically useful, because for much of its journey towards the sensor the radiation is propagating in a medium that approximates to free space, and it also allows us to develop some of the essential ideas that describe electromagnetic radiation without too much confusing detail.

A particularly important part of this chapter deals with thermal radiation. As we noted in Chapter 1, most passive remote sensing systems detect thermal radiation (in the infrared or microwave regions) or they detect reflected solar radiation. Solar radiation is itself, as explained in this chapter, essentially just another form of thermal radiation, so by developing an understanding of thermal radiation we are able to describe many of the characteristics of the radiation detected by passive systems.

2.1

Electromagnetic waves

It is assumed that the reader is more or less familiar with the theory of electromagnetism, and that Sections 2.1 and 2.2 can be regarded as a refresher. If not, the range of suitable physics textbooks is very wide. Although it originally dates from 1964, Volume 2 of the *Feynman Lectures on Physics* (Feynman, Leighton and Sands 2005) still offers one of the most illuminating approaches.

James Clerk Maxwell (Figure 2.1) unified the laws of electricity and magnetism in the 1860s, and proposed that light was a form of electromagnetic radiation. One form in which Maxwell’s equations can be written, for free space, is

$$\nabla \cdot \mathbf{E} = 0, \tag{2.1.1}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{2.1.2}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{2.1.3}$$

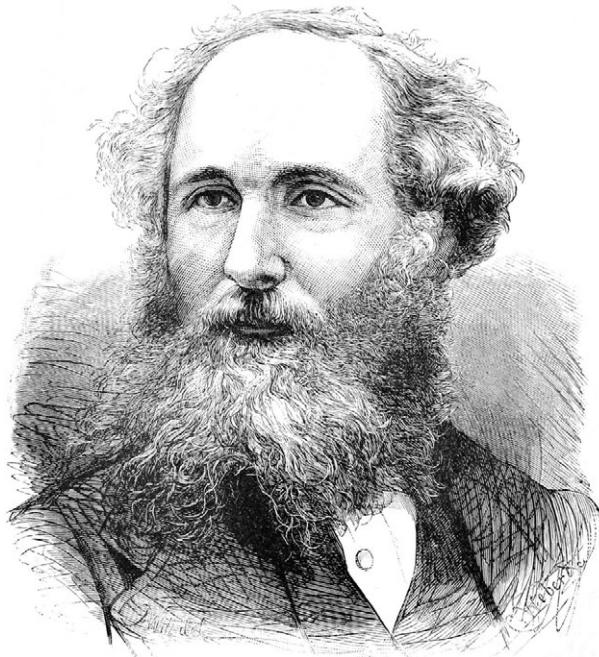


Figure 2.1. James Clerk Maxwell (1831–79) as a young man. (Source: Wikipedia. <http://en.wikipedia.org/wiki/File:YoungJamesClerkMaxwell.jpg>)

$$\nabla \times \mathbf{B} = \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t}. \quad (2.1.4)$$

(This is a particularly compact form, using the notation of differential operators in vector calculus.) In these expressions, \mathbf{E} and \mathbf{B} are the electric and magnetic field vectors, respectively, of the wave, and ϵ_0 and μ_0 are the *electric permittivity* and the *magnetic permeability* of free space.

It can easily be confirmed that the plane wave

$$E_x = E_0 \cos(\omega t - kz), \quad (2.2)$$

$$E_y = 0,$$

$$E_z = 0,$$

$$B_x = 0,$$

$$B_y = \frac{E_0}{c} \cos(\omega t - kz), \quad (2.3)$$

$$B_z = 0,$$

satisfies Equations (2.1.1) to (2.1.4), provided that the wave speed

$$c = \frac{\omega}{k} = \frac{1}{\sqrt{\epsilon_0 \mu_0}}. \quad (2.4)$$

2.1 Electromagnetic waves

c is the speed of light, and of all electromagnetic waves, in free space. It has a value of $299\,792\,458\text{ m s}^{-1}$. (This value is very well determined, and in fact now defines the metre in terms of the second. Values of important constants such as c are given in the appendix.) This plane wave is visualised in Figure 2.3.

Note that we have used the *angular frequency* ω and the *wavenumber* k , rather than the more familiar cyclic frequency f and wavelength λ . The former are usually more useful, and we shall use them often. They are related to frequency and wavelength respectively by

$$\omega = 2\pi f \quad (2.5)$$

and

$$k = \frac{2\pi}{\lambda}. \quad (2.6)$$

In principle, the frequency of an electromagnetic wave can take any value, and the whole range of possible frequencies is called the *electromagnetic spectrum*. Different regions of the spectrum are conventionally given names such as light, radio waves, ultra-violet radiation and so on, usually referring to the manner in which the radiation is generated or detected. The electromagnetic spectrum is shown schematically in Figure 2.2.

In addition to the wavelength λ and the frequency f , Figure 2.2 shows the *photon energy*. This is a concept of quantum physics, in which electromagnetic radiation behaves like a particle rather than (or as well as) a wave. The particle is called a photon, and its energy is given by

$$E = hf, \quad (2.7)$$

where h is the Planck constant (this was originally a proposal of Einstein's). In practice the SI unit of energy, the joule, is inconveniently large to represent photon energies, and it is common to use the *electronvolt* ($1\text{ eV} \approx 1.6 \times 10^{-19}\text{ J}$) instead. This unit is convenient when we consider the interaction of electromagnetic radiation with atoms and molecules (Section 3.4.1).

Returning to the electromagnetic wave specified by Equations (2.2) and (2.3), E_0 is the *amplitude* of the electric field, and E_0/c is the amplitude of the magnetic field, although since these two amplitudes are related by the factor c it is common simply to refer to E_0 as the amplitude of the wave. The wave carries energy in its direction of propagation, which is the positive z -direction, and the *flux density* (power crossing unit area normal to the propagation direction) is given by

$$F = \frac{E_0^2}{2Z_0}, \quad (2.8)$$

where Z_0 is the *impedance of free space*, defined by

$$Z_0 = \sqrt{\frac{\mu_0}{\epsilon_0}}. \quad (2.9)$$

It has a value of approximately 377Ω .

Electromagnetic waves in free space

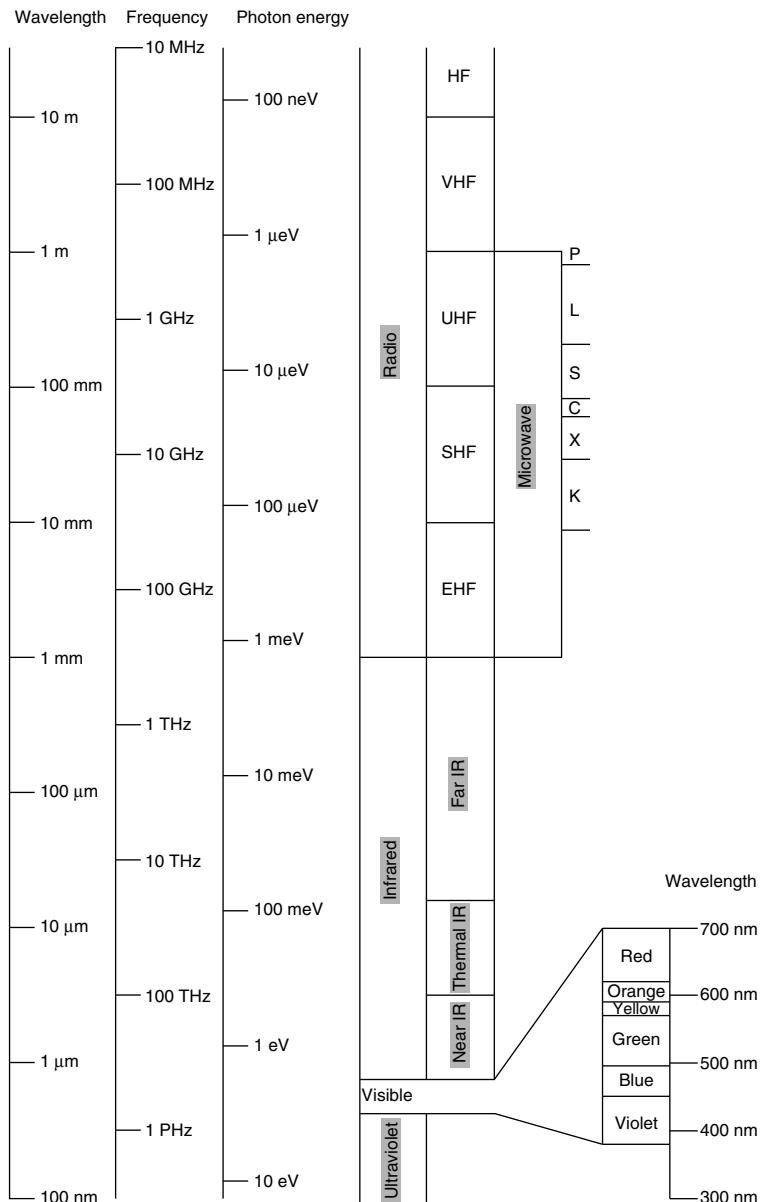


Figure 2.2. The electromagnetic spectrum. The diagram shows those parts of the electromagnetic spectrum that are important in remote sensing, together with the conventional names of the various regions of the spectrum. The letters (P, L, S etc.) used to denote parts of the microwave spectrum are in common use in remote sensing, being standard nomenclature amongst radar engineers in the USA. Note that this nomenclature varies somewhat in other countries, particularly in military usage. Note also that various terminologies are in use for the subdivisions of the infrared (IR) part of the spectrum. That adopted here defines the thermal IR band as lying between 3 and 15 μm , since this region contains most of the power emitted by black bodies at terrestrial temperatures.

Summary

Electromagnetic radiation in free space is a phenomenon intrinsic to the laws of electricity and magnetism. The simplest form of radiation is a plane wave, which can be characterised by its angular frequency ω and its wavenumber k , although these are related to one another through $\omega = ck$, where c is the speed of light, a fundamental physical constant. An alternative notation uses the cyclic frequency f and the wavelength λ , given respectively by $f = \omega/2\pi$ and $\lambda = 2\pi/k$. Any of these four equivalent variables can be used to characterise the type of radiation. The part of the electromagnetic spectrum that is important in remote sensing extends from a (cyclic) frequency of 10^8 Hz or less, to around 10^{15} Hz. It is conventionally divided into the microwave region (frequencies below 3×10^{10} Hz), infrared region (3×10^{10} to 4×10^{14} Hz), visible light (4 to 8×10^{14} Hz) and ultraviolet radiation (above 8×10^{14} Hz); all of these regions have useful subdivisions.

If the particle-like behaviour of the electromagnetic radiation is important, the radiation is conceived of as a stream of particles called photons. The energy of a photon is given by hf where h is the Planck constant. If the wave-like behaviour is important, the intensity of the radiation is proportional to the square of the electric field amplitude.

2.2

Polarisation

The wave specified by Equations (2.2) and (2.3) is not the most general electromagnetic wave propagating in the z -direction. We can find another such wave by simply rotating our coordinate system by 90° about the z -axis, to give

$$E_y = E_0 \cos(\omega t - kz), \quad (2.10)$$

$$B_x = -\frac{E_0}{c} \cos(\omega t - kz), \quad (2.11)$$

all other components being zero. If we now add the waves represented by (2.2) and (2.10), giving them different amplitudes and phases, we obtain an expression for a general wave propagating in the z -direction:

$$E_x = E_{0x} \cos(\omega t - kz - \phi_x), \quad (2.12)$$

$$E_y = E_{0y} \cos(\omega t - kz - \phi_y), \quad (2.13)$$

$$E_z = 0.$$

Note that we do not need to specify the components of the magnetic field \mathbf{B} , since they are defined uniquely by the components of the electric field \mathbf{E} . The two fields are always perpendicular to one another, and to the propagation direction, and the ratio of the amplitude of the electric field to that of the magnetic field is always equal to c . This isn't quite enough to define \mathbf{B} : we also need to specify that, at any instant, the vectors \mathbf{E} , \mathbf{B}

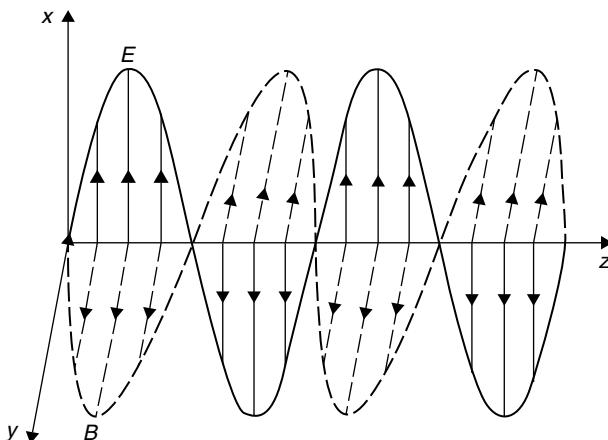


Figure 2.3. Plane-polarised radiation. The wave is propagating in the z -direction and is polarised with the electric field parallel to the x -axis and the magnetic field parallel to the y -axis. The arrows represent the instantaneous magnitudes and directions of the fields.

and the propagation direction (in this case the z -axis) form a right-handed set. (In other words, the vector $\mathbf{E} \times \mathbf{B}$ is in the direction of propagation.)

The values of E_{0x} , E_{0y} , ϕ_x and ϕ_y determine the way in which the direction of the electric field (and hence also of the magnetic field) varies with time. This is termed the *polarisation* of the radiation and, as we shall see later, it is often important to consider it in discussing the operation of a remote sensing system.

If the effect of the variables in Equations (2.12) and (2.13) is to cause the electric field vector \mathbf{E} to remain pointing in the same direction, the radiation is said to be *plane polarised*. This is illustrated in Figure 2.3. Clearly this requires that the phase difference $\phi_y - \phi_x = 0, \pi$ or $-\pi$. (We need only consider values of this phase difference in the range $-\pi$ to $+\pi$, since a value outside this range can be expressed as a value within it by adding or subtracting some integral multiple of 2π .)

Although in principle the direction of the polarisation could be specified using either the electric or the magnetic field, it is conventional to use the electric field, so the example in Figure 2.3 would be described as *x-polarised*.

If, instead of being confined to a fixed direction, the electric field vector rotates in the xy -plane with a constant amplitude, the radiation is said to be *circularly polarised* (Figure 2.4). If the sense of the rotation is clockwise when viewed along the propagation direction the polarisation is called *right-hand circular* (RHC), and if anticlockwise it is *left-hand circular* (LHC). Clearly, circular polarisation requires that

$$E_{0x} = E_{0y},$$

and right-hand polarisation requires that

$$\phi_y - \phi_x = \frac{\pi}{2}.$$

For left-hand polarisation,

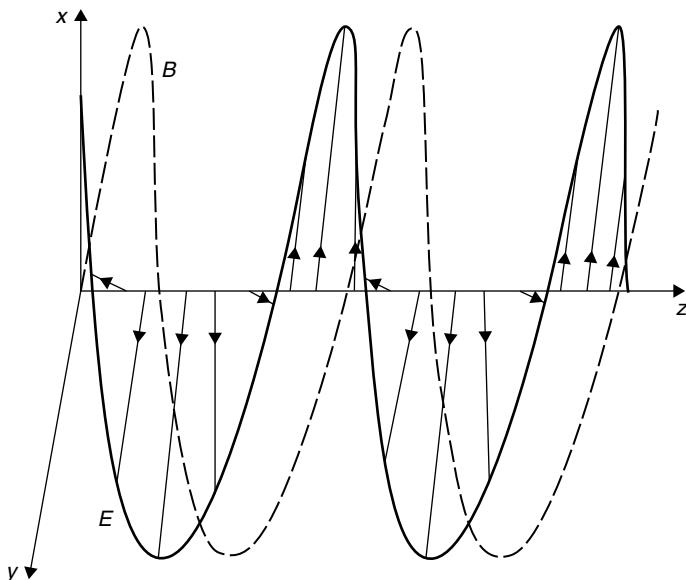


Figure 2.4. Right-hand circularly polarised radiation. The notation is the same as in Figure 2.3, although the magnetic field vectors have been omitted for clarity. They are, as always, oriented perpendicularly to the electric field vectors.

$$\phi_y - \phi_x = -\frac{\pi}{2}.$$

(The convention for defining left- and right-handed polarisations that is given here applies to radio-frequency radiation. For light, the opposite convention is used.)

The only other kind of ‘pure’ polarisation (i.e. completely polarised radiation) is *elliptically polarised* radiation, in which the path traced by the electric field vector in the xy -plane is an ellipse. This corresponds to a phase difference of $\pm\pi/2$ but different amplitudes for the x and y components of the field. In general, the polarisation of an electromagnetic wave will be a mixture of these various types (elliptical polarisation is itself a combination of linear and circular polarisation), and may also include a *randomly polarised* component in which the direction of the electric field vector changes randomly on a time-scale too short to measure. This kind of radiation is often called *unpolarised* radiation, although this is perhaps a somewhat misleading name since it suggests that the electric field vector does not point in any direction.

There are a number of notations for specifying the polarisation state of electromagnetic radiation. One of the most common is the *Stokes vector*, the four components of which can be defined in terms of (2.12) and (2.13) as follows:

$$S_0 = \langle E_{0x}^2 \rangle + \langle E_{0y}^2 \rangle, \quad (2.14.1)$$

$$S_1 = \langle E_{0x}^2 \rangle - \langle E_{0y}^2 \rangle, \quad (2.14.2)$$

Electromagnetic waves in free space

$$S_2 = \langle 2E_{0x}E_{0y} \cos(\phi_y - \phi_x) \rangle, \quad (2.14.3)$$

$$S_3 = \langle 2E_{0x}E_{0y} \sin(\phi_y - \phi_x) \rangle. \quad (2.14.4)$$

The angle brackets $\langle \rangle$ in these expressions denote time-averages.

Examples of some Stokes vectors are given below. In each case, the Stokes vector has been normalised so that $S_0 = 1$.

$[1 \ 0 \ 0 \ 0]$	random polarisation
$[1 \ 1 \ 0 \ 0]$	x -polarised linear
$[1 \ -1 \ 0 \ 0]$	y -polarised linear
$[1 \ 0 \ 1 \ 0]$	+45°linear
$[1 \ 0 \ -1 \ 0]$	-45°linear
$[1 \ 0 \ 0 \ 1]$	right-hand circular
$[1 \ 0 \ 0 \ -1]$	left-hand circular
$[1 \ 0.6 \ 0 \ 0.8]$	right-hand elliptical, $E_{0x}/E_{0y} = 2$

The *degree of polarisation* of an electromagnetic wave is defined as the fraction of the total power that is contained in polarised components. It is given in terms of the components of the Stokes vector by

$$\frac{\sqrt{S_1^2 + S_2^2 + S_3^2}}{S_0}.$$

It can be verified that in all of the examples above, with the exception of the first, the degree of polarisation is 1. The total flux density of the radiation is proportional to S_0 , and in fact is given by

$$F = \frac{S_0}{2Z_0}. \quad (2.15)$$

How strongly polarised is natural light?

Light from the Sun is randomly polarised, so its degree of polarisation is 0. However, light that has been scattered by molecules in the atmosphere (i.e. what we see as blue sky, as discussed in Chapter 4) is generally polarised. The degree of polarisation depends on the direction relative to the Sun, and is greatest when the line of sight to the sky is perpendicular to the line of sight to the Sun, in which case the skylight is approximately 75% linearly polarised if the air is clean. This can be verified, at least qualitatively, using a polarising filter such as the lens of a pair of polarising sunglasses.

The photographs show the sky viewed perpendicular to the direction of the Sun, close to sunset. The photographs were taken using a polarising filter: the photograph on the left shows the vertically polarised component and the photograph on the right the horizontally polarised component.

2.3 Spectra and the Fourier transform

The Stokes components of two electromagnetic waves of the same frequency, travelling in the same direction, can be added provided that the two waves are *incoherent* (i.e. that there is a randomly changing phase difference between them). This allows us to ‘decompose’ a Stokes vector into its polarised components, together with a randomly polarised component if necessary. If a remote sensing system responds only to one polarisation state (this is a common situation for microwave systems), we need to consider the component of the incident radiation that has that polarisation state. For example, randomly polarised radiation can be decomposed into incoherent x- and y-polarised components:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix},$$

so that an instrument capable of detecting only y-polarised radiation will collect half of the power available from a randomly polarised wave. Most natural sources of radiation are randomly polarised, although as we shall see, scattering and reflection may change the state of polarisation.

Summary

‘Polarisation’ of electromagnetic radiation refers to the orientation of the electric and magnetic field vectors. These must be oriented perpendicular to one another and to the direction of propagation, so it is sufficient to describe the orientation of the electric field vector. If it remains in a fixed plane the radiation is said to be linearly polarised or plane polarised, while if it rotates it is circularly polarised. Elliptically polarised radiation is a combination of linearly and circularly polarised radiation. Radiation is not necessarily polarised: in unpolarised, or randomly polarised, radiation the orientation of the electric field changes randomly on a short time-scale. Partially polarised radiation is a combination of unpolarised and polarised radiation. The intensity and polarisation state of electromagnetic radiation can be combined in the Stokes vector.

2.3

Spectra and the Fourier transform

Up to this point we have said nothing about the frequency (or wavelength) of the radiation, other than that electromagnetic radiation may, in principle, have any frequency we wish. It will often happen, however, that we wish to describe a particular radiation field in which a number (possibly a continuous distribution) of frequencies is present. This can be done by specifying the complete waveform, which obviously contains all the necessary information, or the *spectrum* of the radiation – the amplitudes of the various frequency components that are present in the waveform. These two methods are equivalent, and it is important to know how to convert from one description to another. The conversion is achieved using the *Fourier transform*, and since this is of great importance in many aspects of remote sensing it is worth deriving the theory.

It will be convenient to use the *complex exponential* notation to describe sinusoidal or cosinusoidal components, since it greatly simplifies the following analysis. Using this notation, we express a variation having angular frequency ω and amplitude A as

$$A \exp(i\omega t), \quad (2.16)$$

where $i^2 = -1$ and ‘exp’ is the exponential function, i.e. $\exp(x) \equiv e^x$. By allowing A to take complex values, and adopting the convention that it is the *real part* of (2.16) that corresponds to the variation of the physical quantity, we can represent both sinusoidal and cosinusoidal components. We can see this by writing A in terms of its real and imaginary parts, expanding $\exp(i\omega t)$ as $\cos(\omega t) + i \sin(\omega t)$, and taking the real part of (2.16):

$$\Re = [\left(\Re(A) + i\Im(A) \right) (\cos(\omega t) + i \sin(\omega t))] \Re(A)\cos(\omega t) - \Im(A)\sin(\omega t).$$

(Here we have used the symbols \Re and \Im to denote the real and imaginary parts respectively.) Let us suppose that some time-varying quantity (for example, the electric field amplitude at a give location as an electromagnetic wave passes through it) is written as a function of time $f(t)$, and that it is also possible to express it as the sum of components of various angular frequencies ω . If the distribution of frequencies is continuous, the amount of each frequency present can be expressed by a density function $a(\omega)$ such that the total amplitude of the components having frequencies in the range ω to $\omega + d\omega$ ($d\omega$ being very small) is $a(\omega) d\omega$. Thus the contribution from this range of frequencies is written as

$$a(\omega) \exp(i\omega t) d\omega$$

and the sum of the contributions from all frequencies can be obtained by integrating this expression:

$$f(t) = \int_{-\infty}^{\infty} a(\omega) \exp(i\omega t) d\omega. \quad (2.17)$$

So far this is merely an assertion. We have neither proved that the distribution $a(\omega)$ uniquely represents $f(t)$, nor shown how to find $a(\omega)$ given $f(t)$. It is beyond our scope to find a rigorous answer to the former problem, so we shall content ourselves with answering the latter. However, we don’t actually need to know how this is done in order to use the Fourier transform, so the derivation is given in the box. The result is shown there to be

$$a(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) \exp(-i\omega t) dt. \quad (2.18)$$

This is very similar to (2.17) and shows that, apart from a change of sign and scale, $a(\omega)$ is obtained from $f(t)$ in exactly the same way as $f(t)$ is obtained from $a(\omega)$.

Finding the inverse Fourier transform

If we multiply (2.17) by $\exp(i\omega' t)$, where ω' is an arbitrary angular frequency, we obtain

$$f(t)\exp(i\omega't) = \int_{-\infty}^{\infty} a(\omega)\exp(i(\omega + \omega')t)d\omega.$$

Next we integrate this with respect to t , giving

$$\begin{aligned} \int_{-\infty}^{\infty} f(t)\exp(i\omega't)dt &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(\omega)\exp(i(\omega + \omega')t)d\omega dt \\ &= \int_{-\infty}^{\infty} a(\omega) \int_{-\infty}^{\infty} \exp(i(\omega + \omega')t)d\omega dt. \end{aligned}$$

Now $\int_{-\infty}^{\infty} \exp(i\alpha t)dt$ is a function of α that is zero everywhere but at $\alpha = 0$, where it is infinite. The area underneath a graph of this function is, however, finite, and has a value of 2π . This can be written as

$$\int_{-\infty}^{\infty} \exp(i\alpha t)dt = 2\pi\delta(\alpha),$$

where $\delta(\alpha)$ is the *Dirac delta-function*. Thus we have

$$\int_{-\infty}^{\infty} f(t)\exp(i\omega't)dt = 2\pi a(-\omega'),$$

which can be rewritten, by changing the symbols and rearranging the expression, as

$$a(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\exp(-i\omega t)dt.$$

The integral transforms defined by (2.17) and (2.18) are called Fourier transforms, although we should note that some authors increase the symmetry between (2.17) and (2.18) still further by writing

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} a(\omega)\exp(i\omega t)d\omega \\ a(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\exp(-i\omega t)dt. \end{aligned}$$

Let us apply the Fourier transform to a practical example. Suppose we have a waveform $f(t)$, which consists of a single angular frequency ω_0 which is turned on for a finite time

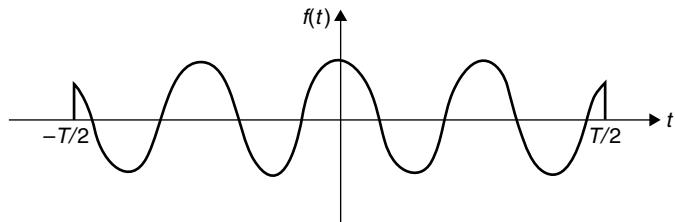


Figure 2.5. A truncated cosine wave. The Fourier transform of this function is shown in Figure 2.6.

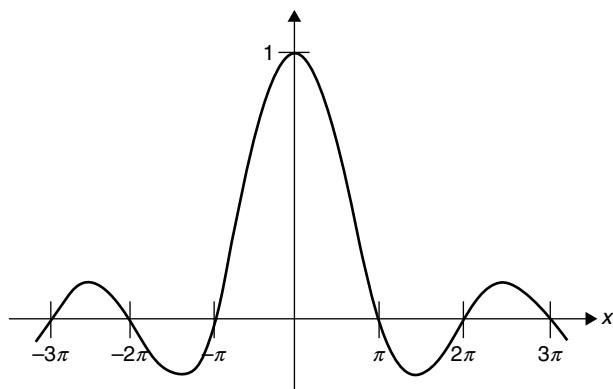


Figure 2.6. The function $\text{sinc}(x)$, defined as $(\sin x)/x$.

T (Figure 2.5). What is its spectrum $a(\omega)$? We might think that since we used only one frequency to construct $f(t)$, the spectrum would consist of a single spike or delta-function at that frequency. However, this cannot be correct since the spectrum $a(\omega)$ has to contain *all* the information contained by $f(t)$, including the fact that the waveform drops abruptly to zero for $|t| > T/2$. Using (2.18), then, we find that

$$\begin{aligned} a(\omega) &= \frac{1}{2\pi} \int_{-T/2}^{T/2} \cos(\omega_0 t) \exp(-i\omega t) dt \\ &= \frac{T}{4\pi} \left[\text{sinc}\left(\frac{(\omega_0 - \omega)T}{2}\right) + \text{sinc}\left(\frac{(\omega_0 + \omega)T}{2}\right) \right]. \end{aligned}$$

This is evidently the sum of two functions, each of the form $\text{sinc}(x)$, which we define as $(\sin x)/x$, centred at frequencies ω_0 and $-\omega_0$. The function $\text{sinc}(x)$ is shown in Figure 2.6. (Note that some authors define $\text{sinc}(x)$ to be $(\sin \pi x)/(\pi x)$).

Thus the complete spectrum of the waveform whose time dependence was shown in Figure 2.5 is shown by Figure 2.7. It can be seen that the delta-functions that we might have expected at $\omega = \pm \omega_0$ have been spread out over a range $2\delta\omega$ in angular frequency, where

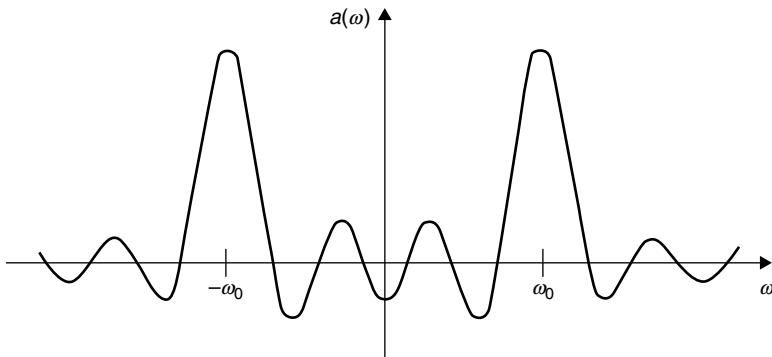


Figure 2.7. The Fourier transform of the function shown in Figure 2.5.

$$\delta\omega \approx \frac{2\pi}{T}, \quad (2.19)$$

or $\delta f \approx 1/T$. This is in fact a general result of fundamental importance: in order to represent a waveform of duration δt , we need a range of frequencies of at least $\pm 1/\delta t$. It is a form of ‘uncertainty principle’. Defining exactly what is usefully meant by ‘length’ and ‘range’ is not always obvious, so we leave (2.19) in its approximate form, although a more exact formulation of the result is possible.

A related result, which is however quite exact, is the *Nyquist sampling theorem*. This states that if a signal is to be sampled at regular intervals, the sampling frequency must exceed some minimum value if it is to be possible to reconstruct the original signal unambiguously from the samples. This frequency is the *Nyquist frequency*, and it is twice the bandwidth of the signal. The bandwidth is defined as the range of frequencies f over which the signal spectrum is non-zero. If the signal is undersampled, i.e. sampled at a rate below the Nyquist frequency, *aliases* are introduced which, amongst other undesirable effects, degrade the signal-to-noise ratio. The practical implications of the Nyquist theorem are many, but it clearly finds an important application in the design of electronic systems in which a signal is first filtered to define a bandwidth, and then sampled at regular intervals. An example of the aliasing phenomenon is given in Section 10.3.2.4.

The Fourier transform is a particular example of what are in general termed *integral transforms*. All of these have the same property that they convert a function from one ‘domain’ into another, either of the two representations of the function containing all the information about it. In the case of the Fourier transforms that we have been considering, the two domains are those of time and frequency. Analogous Fourier transforms can be defined between the spatial domain and the spatial frequency domain, and we shall make use of this idea in Section 3.3.4, where we consider the interaction of microwave radiation with a rough surface, and in Chapter 11, where we consider the spatial properties of images. However, there are many other integral transforms and several of them can usefully be used to describe the time-structure of a signal. Two particularly common ones are the Laplace and Hartley transforms.

Wavelet transforms are also useful for describing situations where the periodic content of a function changes over time. We shall consider the spatial equivalent of these wavelet transforms in Chapter 11.

Summary

Any time-varying quantity, such as a wave, can be represented explicitly as a function of time, or as a (possibly infinite) sum of components of different frequencies. The Fourier transform converts between these two equivalent representations. If a time-varying quantity is non-zero only for a finite time T , the range of frequencies (the bandwidth) needed to represent it is at least approximately $2/T$.

2.4

The Doppler effect

If a source of electromagnetic radiation of frequency f is in motion with respect to an observer (e.g. a sensor), the observer will in general detect the radiation at a different frequency f' . If the source is approaching the observer, or equivalently if the observer is approaching the source, f' will be greater than f , and conversely. This is known as the Doppler effect, and is analogous to the similar (and familiar) effect observed with sound waves. However, whereas the Doppler effect for sound is not the same for the source approaching the observer and for the observer approaching the source, the effect is symmetrical in this manner for electromagnetic radiation in free space. The result has to be derived using Einstein's Special Theory of Relativity, so it will merely be stated here.

If the source S approaches the observer O with a velocity v directed at an angle θ to the line of sight, as shown in Figure 2.8, the Doppler shift is given by

$$\frac{f'}{f} = \frac{\sqrt{1 - v^2/c^2}}{1 - v\cos\theta/c}, \quad (2.20)$$

where c is the speed of light. However, in all cases that will concern us, the relative speed v will be very much smaller than the speed of light, in which case a very good approximation to Equation (2.20) is given by

$$\frac{f'}{f} = 1 + \frac{v\cos\theta}{c}. \quad (2.21)$$

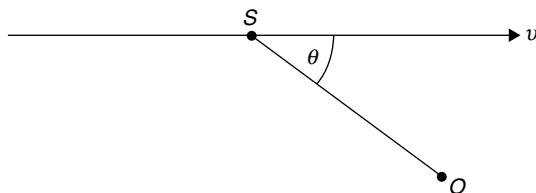


Figure 2.8. The Doppler effect. The source of electromagnetic radiation is located at S , travelling with velocity v . The observer is located at O .

2.5 Describing angular distributions of radiation

For example, if a satellite is travelling away from an observer on the Earth with a speed of 7 km s⁻¹ at an angle of 10° to the line of sight (thus $\theta = 170^\circ$), and it emits a signal with a frequency of exactly 5 GHz, the received frequency will be 4.999 885 GHz. In other words, the frequency has been shifted downwards by 115 kHz. The error in calculating this shift using the approximate equation (2.21) is only about 1 Hz and may be ignored.

Although it is small, a consideration of the Doppler effect is important for some radar systems, particularly the synthetic aperture radar systems discussed in Chapter 9.

Summary

If a source of electromagnetic radiation and a receiver are in relative motion, the frequency measured at the receiver will differ from the frequency transmitted from the source. If the distance between the source and the receiver is decreasing the frequency is increased, and conversely. Provided that the relative speed is very small compared with the speed of light, the fractional change in the frequency is simply the rate of decrease of the distance between source and receiver, expressed as a ratio of the speed of light.

2.5

Describing angular distributions of radiation

We have already seen how to describe electromagnetic radiation that contains a range of frequencies or a range of polarisations. Up to this point, however, we have considered only collimated radiation, that is, radiation travelling in a single direction. It is clear that we also need to be able to describe radiation distributed over a range of directions in space. The radiometric quantities introduced in this section are also discussed by Hapke (2005).

Let us begin by considering a plane surface that is illuminated by radiation from a variety of directions. To specify a particular direction of incident radiation we need two angles: θ , the angle between the propagation direction and the normal to the surface element; and ϕ , the azimuthal angle, measured around the normal in the plane of the surface (see Figure 2.9).

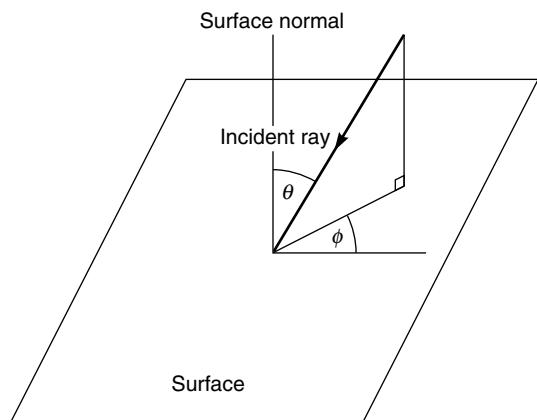


Figure 2.9. Definition of the angles θ and ϕ to describe the angular distribution of radiation.

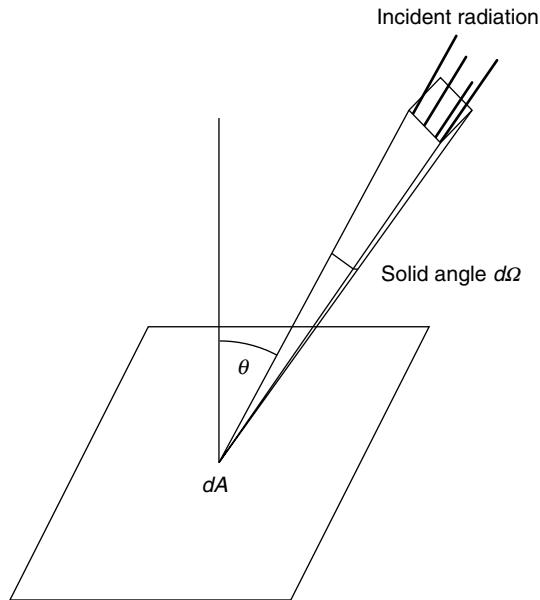


Figure 2.10. Geometrical construction to explain the concept of radiance.

Now we consider an element dA of this surface, and radiation incident from the range of directions between θ and $\theta + d\theta$ and between φ and $\varphi + d\varphi$ (Figure 2.10). The solid angle (unit: steradian; symbol: sr; see box) defined by this range of directions is

$$d\Omega = \sin\theta d\theta d\phi \quad (2.22)$$

and it is clear that the power incident on the element dA from this range of directions must be proportional to dA and $d\Omega$ as well as to a term that defines the strength of the radiation. This relationship can be expressed as

$$dP = L \cos\theta dAd\Omega \quad (2.23)$$

where dP is the contribution to the power incident on the area dA from solid angle $d\Omega$ in the direction (θ, φ) , and L is the *radiance* of the incident radiation in that direction. From this definition, it follows that the SI unit of radiance is $\text{W m}^{-2} \text{ sr}^{-1}$.

Solid angle

Solid angles can be defined analogously to planar angles. A planar angle can be defined by measuring the length s of the arc that it subtends on a circle of radius r : the angle θ (in radians) is given by s/r . Similarly, if a solid angle subtends an area A on the surface of a sphere of radius r , the solid angle (in steradians) is given by $\Omega = A/r^2$. Since the total area of the sphere is $4\pi r^2$, the solid angle represented by the full range of directions is 4π steradians.

2.5 Describing angular distributions of radiation

The inclusion of the factor $\cos \theta$ in Equation (2.23) seems perverse at first sight. However, it gives the radiance the valuable property that, if the medium through which the radiation propagates does not scatter or absorb and has a constant refractive index (and these conditions are obviously all met if we are still considering radiation in free space), then the radiance is constant along any ray. The concept of radiance is of prime importance in considering measurements made by optical and near-infrared remote sensing systems, discussed in Chapter 6.

The *irradiance* E at the surface is defined as the total incident power per unit area, and its SI unit is W m^{-2} . It is found by integrating Equation (2.23) over all the directions for which $\theta \leq \pi/2$, i.e. the hemisphere of directions from which the surface can be illuminated:

$$E = \int_{\theta=0}^{\pi/2} \int_{\phi=0}^{2\pi} L_{\text{incoming}} \cos \theta d\Omega. \quad (2.24)$$

Although the radiance may be a function of direction, the irradiance clearly cannot be.

We can use the same ideas to describe radiation emitted or reflected from a surface. Since the concept of radiance describes radiation in space, the same terminology will suffice for both incoming and outgoing radiation. All we need to do is to ‘label’ the radiation so that we know in which direction it is propagating, and terms such as ‘upwelling’ and ‘downwelling’ radiation are frequently used for this purpose. The outgoing analogue of irradiance is termed *radiant exitance*, and given the symbol M :

$$M = \int_{\theta=0}^{\pi/2} \int_{\phi=0}^{2\pi} L_{\text{outgoing}} \cos \theta d\Omega. \quad (2.25)$$

For *isotropic* radiation, the radiance is independent of direction. In this case, the relationship between the radiance and the exitance is given by

$$M = L \int_{\theta=0}^{\pi/2} \int_{\phi=0}^{2\pi} \cos \theta d\Omega = \pi L. \quad (2.26)$$

Summary

The fundamental concept in describing how much electromagnetic radiation is travelling in different directions is the radiance L , which can be defined in terms of the power reaching unit area of a surface and the solid angle from which the radiation originates. If electromagnetic radiation reaches a surface from a range of directions, the total power incident on unit area of the surface is called the irradiance E and is obtained by integrating the radiance over all directions. Similar concepts can be defined in the case of a surface emitting radiation: the term corresponding to the irradiance is the radiant exitance M . Isotropic radiation has equal radiance in all directions. This implies that the relationship between radiant exitance and emitted radiance for an isotropic emitter is $M = \pi L$.

2.6

Thermal radiation

Thermal radiation is emitted by all objects above absolute zero (-273.15°C – see box) and is, at first or second hand, the radiation that is detected by the great majority of passive remote sensing systems.

The absolute temperature scale

In describing thermal radiation, it is convenient to use the absolute scale in which temperature is measured in *kelvin* (K). The relationship between a temperature T in kelvin and a temperature t in degrees Celsius is

$$T(K) = t({}^{\circ}\text{C}) + 273.15$$

so that the absolute zero of the temperature scale is at $-273.15 {}^{\circ}\text{C}$.

In general, a hot object (by which, for the present, we mean one that is not at absolute zero) will distribute its emission over a range of wavelengths in a continuous spectrum. To describe this radiation we can use the same radiometric quantities that were defined in Section 2.5, but we need to modify the definitions to include the variation with wavelength or frequency. This is done by defining the *spectral radiance* L_{λ} such that the radiance ΔL contained in a small range of wavelengths $\Delta\lambda$ is given by

$$\Delta L = L_{\lambda} \Delta\lambda. \quad (2.27)$$

In other words, L_{λ} is just the differential of L with respect to λ , or more strictly the absolute value (modulus) of this differential:

$$L_{\lambda} = \left| \frac{\partial L}{\partial \lambda} \right|. \quad (2.28)$$

It is clear that the SI unit of spectral radiance is $\text{W m}^{-2} \text{ sr}^{-1} \text{ m}^{-1}$, although the units $\text{W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$ and $\text{W m}^{-2} \text{ sr}^{-1} \text{ nm}^{-1}$ are also commonly used.

The spectral radiance can also be defined in terms of the frequency f :

$$L_f = \left| \frac{\partial L}{\partial f} \right|, \quad (2.29)$$

so that its unit is $\text{W m}^{-2} \text{ sr}^{-1} \text{ Hz}^{-1}$, and the relationship between the definitions (2.28) and (2.29) is therefore given by

$$\frac{L_{\lambda}}{L_f} = \left| \frac{\partial f}{\partial \lambda} \right| = \frac{c}{\lambda^2} = \frac{f^2}{c}, \quad (2.30)$$

where c is the speed of light.

All of the radiometric quantities defined in Section 2.5, not just the radiance, can similarly be defined spectrally.

If we make a closed cavity with opaque walls, and hold the cavity at an absolute temperature T , the electromagnetic radiation inside it is known as *black-body radiation*.

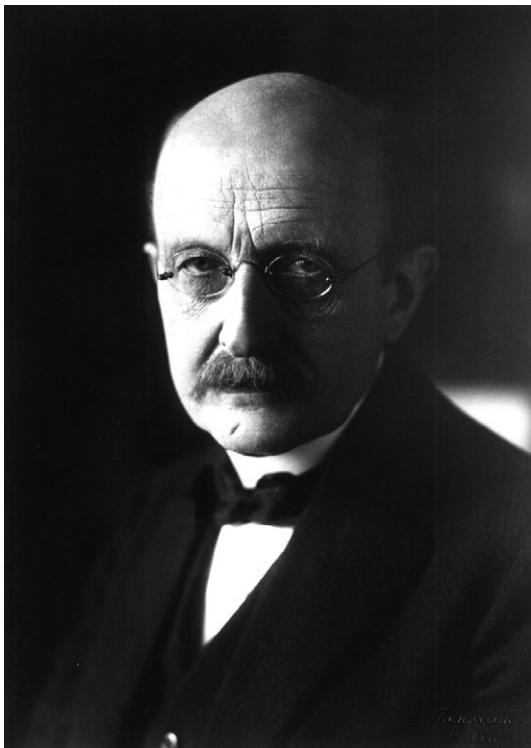


Figure 2.11. Max Planck (1858–1947), the founder of quantum mechanics. He was awarded the Nobel Prize in physics in 1918, the year of this photograph. (Source: Wikipedia. [http://en.wikipedia.org/wiki/File:Max_Planck_\(Nobel_1918\).jpg](http://en.wikipedia.org/wiki/File:Max_Planck_(Nobel_1918).jpg))

The spectral radiance of this radiation was calculated by Max Planck (Figure 2.11) during the early years of the twentieth century, using quantum mechanics (see e.g. Longair 2003)). It is given by

$$L_f = \frac{2hf^3}{c^2(\exp(hf/kT) - 1)}, \quad (2.31)$$

which may also be expressed, using Equation (2.30), as

$$L_\lambda = \frac{2hc^2}{\lambda^5(\exp(hc/\lambda kT) - 1)}. \quad (2.32)$$

In these equations, h is the Planck constant and k is the Boltzmann constant. Equation (2.32) is plotted in Figure 2.12 for a temperature of 300 K, and in Figure 2.13 on logarithmic axes for two temperatures, 300 K and 6000 K. Note the steep rise at short wavelengths, and the long tail at long wavelengths.

The radiation inside a closed cavity may not seem particularly interesting or relevant, but we may observe it by making a small hole in the cavity and letting some of it escape.

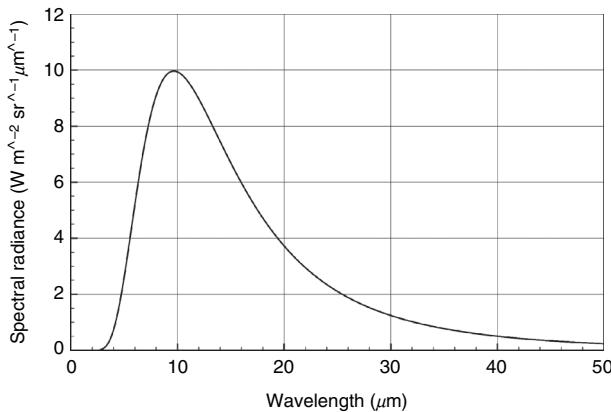


Figure 2.12. Spectral radiance of black-body radiation at a temperature of 300 K.

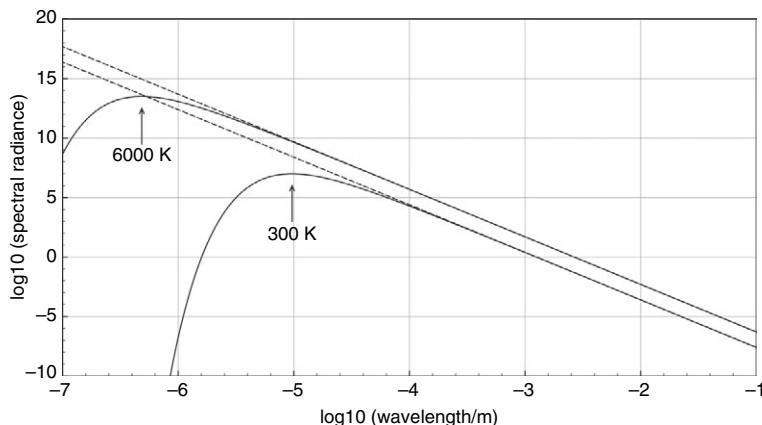


Figure 2.13. Spectral radiance of black-body radiation at 300 K and 6000 K, plotted on logarithmic axes. The spectral radiance is plotted in units of $\text{W m}^{-2} \text{sr}^{-1} \text{m}^{-1}$. The arrows show the position of the maxima and the dashed lines the Rayleigh-Jeans approximation.

In this case, Equation (2.31) or (2.32) describes the radiation emerging from the hole, and from any *black body* (perfect emitter of thermal radiation) at temperature T .

At sufficiently long wavelengths, Equation (2.31) can be approximated as

$$L_f \approx \frac{2kTf^2}{c^2} = \frac{2kT}{\lambda^2} \quad (2.33)$$

or equivalently as

$$L_\lambda \approx \frac{2kTc}{\lambda^4}. \quad (2.34)$$

This is called the *Rayleigh-Jeans approximation*, and corresponds to the right-hand part of Figure 2.13, where the graphs can be approximated as straight lines with slopes of -4 .

2.6 Thermal radiation

The condition for this approximation to be valid is $\frac{hc}{\lambda kT} = \frac{hf}{kT} \ll 1$. For $T = 280$ K, this gives $f = 6000$ GHz or $\lambda \approx 50$ μm, so the approximation is valid for microwave and radio frequencies for objects at typical terrestrial temperatures.

We can integrate the Planck formula (either (2.31) or (2.32), it does not matter which) to calculate the total radiance of black-body radiation over all wavelengths:

$$L = \int_0^\infty L_\lambda d\lambda = \frac{2\pi^4 k^4}{15c^2 h^3} T^4. \quad (2.35)$$

Since the radiation is isotropic, the total radiant exitance M is found, using Equation (2.26), to be

$$M = \pi L = \frac{2\pi^5 k^4}{15c^2 h^3} T^4.$$

This is normally written more compactly as

$$M = \sigma T^4, \quad (2.36)$$

where $\sigma = 2\pi^5 k^4 / 15c^2 h^3 \approx 5.67 \times 10^{-8}$ W m⁻² K⁻⁴ is called the Stefan–Boltzmann constant, and Equation (2.35) is called Stefan’s law. It shows how much power is emitted by a black body at temperature T , integrated over all wavelengths. If we want to know how this power is distributed in wavelength, we can of course use Equation (2.31) directly, but it may be sufficient merely to know the wavelength λ_{\max} at which L_λ reaches a maximum. This is found by differentiating Equation (2.31), which shows that

$$\lambda_{\max} = \frac{A}{T}, \quad (2.37)$$

where A is a constant whose value is about 2.898×10^{-3} K m. Equation (2.37) is called *Wien’s law*, or Wien’s displacement law. For example, the Sun is a fairly good approximation to a black body at a temperature of around 5800 K (see Figure 2.15), so the peak spectral radiance occurs at $\lambda_{\max} \approx 0.50$ μm, in the middle of the visible spectrum where we expect it to be. If, on the other hand, we consider a black body at a temperature of 280 K, which is fairly typical of temperatures on the Earth’s surface, we find $\lambda_{\max} \approx 10.3$ μm, in the thermal-infrared region of the electromagnetic spectrum.

Wien’s law and Stefan’s law in the kitchen

With an electric hot plate, a digital camera, and great care, we can demonstrate qualitatively the behaviour predicted by Wien’s displacement law and by Stefan’s law. Switch on the hotplate and wait until it is glowing red. Then get an assistant to switch out the light (this experiment is best performed when it is dark outside) and take a series of digital photographs of the hotplate as it cools down. Be very careful! Set the camera for maximum, and fixed, sensitivity. By analysing the images quantitatively, you should be able to show that as the hotplate cools down the radiation emitted by it becomes less intense, and more purely red, over time.

We may also sometimes need to calculate the radiance or radiant exitance of a black body over a finite range of wavelengths. This can be simplified a little by integrating with respect to a dimensionless variable. Specifically, we can put

$$\int_{\lambda_1}^{\lambda_2} M_\lambda d\lambda = \sigma T^4 (f(x_1) - f(x_2)), \quad (2.38)$$

where the dimensionless variables x_1 and x_2 are defined by $x_i = \frac{hc}{\lambda_i kT}$ and the function $f(x)$ is defined by

$$f(x) = \frac{15}{\pi^4} \int_0^x \frac{z^3 dz}{e^z - 1}. \quad (2.39)$$

This integral cannot be evaluated analytically, although numerical integration using computer programs is straightforward. The function is plotted in Figure 2.14.

We remarked earlier that a small hole in the wall of a cavity behaves as a black body. This is not a particularly plausible model for real materials, so we introduce the idea of the *emissivity* ε to relate the actual radiance of a body at temperature T to the black-body value. (Note that emissivity and dielectric constant – defined in Section 3.1 – are both conventionally given the symbol epsilon, which has potential for confusion. The usage is too well established, however, for us to introduce a different notation, and we rely on the context, or an explicit statement, to differentiate between them.) The emissivity is often dependent on wavelength, so in general we should write it as $\varepsilon(\lambda)$, and we can define it through

$$L_\lambda = \varepsilon(\lambda) L_{\lambda,P}, \quad (2.40)$$

where we have now written $L_{\lambda,P}$ for the black-body radiance defined by Equation (2.32) (the ‘P’ stands for ‘Planck’). A simple thermodynamic argument shows that a body which

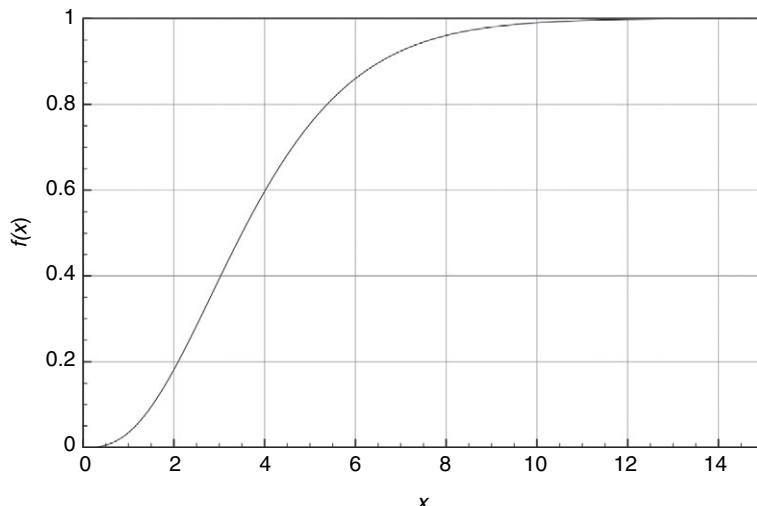


Figure 2.14. The function $f(x)$ defined in Equation (2.39). It shows the fraction of the total black-body radiation emitted up to $x = hc/\lambda kT$.

2.6 Thermal radiation

is a good emitter (high ε) must also be a good absorber of radiation – in fact the two factors must be equal (this is Kirchhoff's law of radiation). We can see this quite easily by realising that any body at temperature T must be in equilibrium with black-body radiation whose spectrum corresponds to the same temperature. If the body absorbs better than it emits, say, it will heat up, and thus cannot in fact be at equilibrium. Thus the reflectivity is given by $1 - \varepsilon$. It also follows from this argument that the emissivity (strictly, the angular average of the emissivity) must lie between 0 and 1. The factors that determine emissivity are discussed in more detail in Sections 3.6.2 and 3.6.3.

It is often convenient, especially when discussing passive microwave systems (Chapter 7), to define the *brightness temperature* of a body that is emitting thermal radiation. This is the temperature of the equivalent black body that would give the same radiance at the wavelength under consideration. By combining Equations (2.32) and (2.40), we can see that at wavelength λ , a body with temperature T and emissivity ε has a brightness temperature T_b that is given by

$$\varepsilon \frac{2hc^2}{\lambda^5 (\exp(hc/\lambda kT) - 1)} = \frac{2hc^2}{\lambda^5 (\exp(hc/\lambda kT_b) - 1)}.$$

The exact solution of this equation for T is

$$T = \frac{hc}{k\lambda \ln \left(1 + \frac{\exp(hc/\lambda kT_b) - 1}{\varepsilon} \right)} \quad (2.41)$$

but at sufficiently long wavelengths (low frequencies) this can be approximated very simply, using the Rayleigh–Jeans approximation, as

$$T_b = \varepsilon T. \quad (2.42)$$

If the emissivity ε is close to 1, (2.41) can be approximated as

$$T \approx \frac{T_b}{1 + \frac{\lambda k T_b \ln \varepsilon}{hc}} \quad (2.43)$$

which can be useful for calibration of satellite imagery. This is discussed in Section 11.2.1.1.

We saw earlier that a black body at a typical terrestrial temperature of 280 K will radiate maximally at a wavelength of 10.3 μm. How well does it radiate at other wavelengths? Specifically, let us calculate the fraction of the total radiant exitance that is emitted in four wavelength ranges: 0.5–0.6 μm, 1.55–1.75 μm, 10.5–12.5 μm and 1.52–1.56 cm. These have been chosen to be typical of remote sensing measurements in the optical, near-infrared, thermal infrared and passive microwave regions respectively. Using the methods described from Equation (2.38) onwards, we find that these fractions are approximately 6×10^{-33} , 7×10^{-10} , 0.12 and 1×10^{-10} respectively. This illustrates the very rapid fall in the Planck function at shorter wavelengths and the much slower decline at longer wavelengths. It also shows that, while objects at normal terrestrial temperatures do not emit thermal radiation in the form of visible light (which is a fact of everyday experience), small but potentially measurable quantities of radiation are emitted in the microwave region. In fact, it is possible to build receivers sensitive enough

to detect this microwave radiation, and this forms the basis of the passive microwave radiometry techniques that will be discussed in Chapter 7.

2.6.1 Characteristics of solar radiation

By way of illustration, we now apply some of the results of Section 2.5 and Section 2.6 to characterise radiation from the Sun. To a fairly good approximation the Sun can be taken to be a black body with an effective temperature T of about 5800 K (Figure 2.15). It can be assumed to be a sphere of radius $r = 6.96 \times 10^8$ m located a distance $D = 1.50 \times 10^{11}$ m (this is called the *astronomical unit*) from the Earth.

From Equations (2.36) and (2.40), we can write the Sun's radiant exitance, integrated over all wavelengths, as $M = \sigma T^4 = 6.35 \times 10^7 \text{ W m}^{-2}$. The total power radiated by the Sun (this is usually called its *luminosity*) is obtained by multiplying this by the Sun's surface area: $P = 4\pi r^2 \sigma T^4 = 3.87 \times 10^{26} \text{ W}$. By considering a sphere of radius D centred on the Sun, we can see that the irradiance at the Earth (but above the Earth's atmosphere so we that we do not need to consider atmospheric absorption) is given by

$$E = \frac{P}{4\pi D^2} = 1.37 \times 10^3 \text{ W m}^{-2}.$$

This value is often called the *mean exoatmospheric irradiance*. We can calculate the corresponding exoatmospheric radiance L by considering the range of directions over which this radiation is distributed. Seen from a distance D , the Sun subtends a solid angle $\Delta\Omega = \frac{\pi r^2}{D^2}$ which is much less than 1 sr, so a sufficiently accurate estimate is given by

$$L = \frac{E}{\Delta\Omega} = \frac{\sigma T^4}{\pi} = 2.02 \times 10^7 \text{ W m}^{-2} \text{ sr}^{-1}.$$

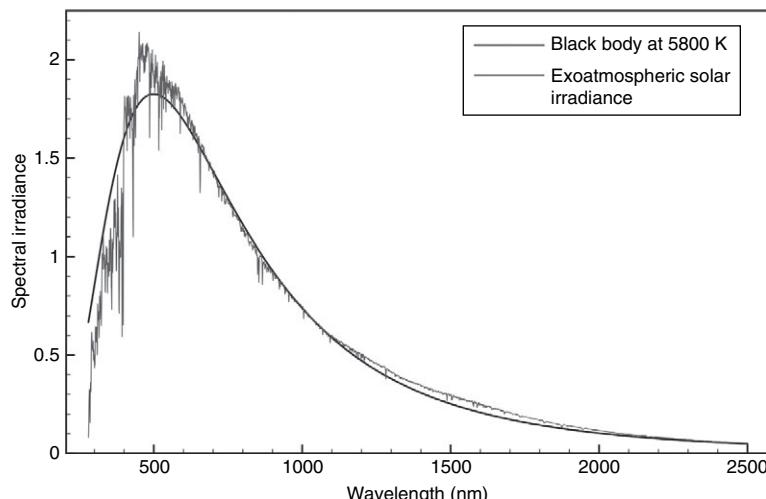


Figure 2.15. Spectral irradiance from the Sun at a distance of 1 astronomical unit (grey curve) (Gueymard 2004) and spectral irradiance from a black body at 5800 K subtending the same solid angle as the Sun (black curve). The units of spectral irradiance are $\text{W m}^{-2} \text{ nm}^{-1}$.

This radiance is confined to the range $\Delta\Omega$ of solid angle. Outside this range, the radiance is of course zero.

We can also calculate the exoatmospheric radiance spectrally, taking Equation (2.32) as our starting point and following the same procedure. We find that

$$L_\lambda = \frac{2\epsilon hc^2}{\lambda^5 (\exp(hc/\lambda kT) - 1)},$$

which of course is just the Planck formula for the radiance, modified by Equation (2.40) to take account of the emissivity. For example, at a wavelength of $0.5 \mu\text{m}$ this gives $L_\lambda = 2.65 \times 10^{13} \text{ W m}^{-2} \text{ sr}^{-1} \text{ m}^{-1}$. The corresponding spectral irradiance is obtained by multiplying this by $\Delta\Omega$ to give $E_\lambda = 1.79 \times 10^9 \text{ W m}^{-2} \text{ m}^{-1}$, which can also be expressed in less standard but more common units as $1.79 \text{ W m}^{-2} \text{ nm}^{-1}$ (confirmed by Figure 2.15) or as $179 \text{ mW cm}^{-2} \mu\text{m}^{-1}$.

Summary

All objects emit electromagnetic radiation as a result of thermal motion. An ideal emitter, which is one that absorbs all radiation that is incident on it, is termed a *black body* and the amount of radiation emitted by a black body depends on its absolute temperature T . The spectral radiance (i.e. the radiance per unit wavelength interval) of black-body radiation is given by the Planck formula: $L_\lambda = \frac{2hc^2}{\lambda^5 (\exp(hc/\lambda kT) - 1)}$. The physical constants in this

expression are the Planck constant h , the speed of light c and the Boltzmann constant k . An equivalent formulation can be given for the radiance per unit frequency interval. At sufficiently long wavelengths the Planck formula can be approximated as $L_\lambda = \frac{2kTc}{\lambda^4}$; this is the Rayleigh–Jeans approximation and it is generally valid in the microwave region. At shorter and shorter wavelengths, the spectral radiance does not continue to increase as would be implied by the Rayleigh–Jeans approximation. Instead, it reaches a maximum value at a wavelength that is inversely proportional to the temperature (this is Wien’s law) and then falls increasingly rapidly at shorter wavelengths. The wavelength of the maximum spectral radiance is around $10 \mu\text{m}$ for objects at a typical terrestrial temperature of 280 K and is around $0.5 \mu\text{m}$ for the Sun, which behaves more or less as a black body at 5800 K . The total radiant exitance, integrated over all wavelengths, of a black body at temperature T is σT^4 (Stefan’s law) where σ is the Stefan–Boltzmann constant.

For bodies that are not perfect absorbers and emitters of radiation we introduce the concept of emissivity ϵ . This is the ratio of the spectral radiance actually emitted by the body to the spectral radiance that would be emitted by a black body at the same temperature. It is equal to 1 for a black body and cannot exceed 1. It can vary with wavelength. The brightness temperature T_b of an object that is emitting thermal radiation is the temperature of a black body that would give rise to the same spectral radiance. It is equal to the temperature T of the object if the emissivity is 1, otherwise it is less than the temperature of the object. If the Rayleigh–Jeans approximation is valid, the relationship between the actual temperature and the brightness temperature is simply $T_b = \epsilon T$. In other cases the relationship is more complicated.

2.7

Diffraction

We conclude this review of the propagation of electromagnetic radiation in free space by discussing diffraction. Diffraction can be roughly defined as the changes that occur to the direction of electromagnetic radiation when it encounters an obstacle of some kind. It could therefore be argued that, since the radiation is interacting with matter (the obstacle), the phenomenon should be discussed in Chapter 3. However, it is more convenient to treat it here since (1) we shall assume that, until the radiation encounters the obstacle and after it has left it, it is propagating in free space, and (2) our approach will build upon the discussion of Fourier transforms developed in Section 2.3. The treatment presented here, which will lead to results that are of fundamental importance in understanding the *spatial resolution* of remote sensing systems, will be very brief. Much fuller treatments can be found in any textbook on optics, for example Hecht (2003).

We begin by considering plane parallel radiation (i.e. radiation travelling in a single direction) incident on a very long slit, of width w , in an infinite opaque screen. The slit has its long axis parallel to the x -axis of a Cartesian coordinate system, and the centre of the slit is located at the origin of this coordinate system. We wish to determine the amplitude of the electric field at the point P shown in Figure 2.16.

If the distance z is sufficiently large (we shall discuss later how large it needs to be), the rays OP and AP may be regarded as parallel, and AP is shorter than OP by $y \sin \theta$. The phase difference between the two rays is thus $ky \sin \theta$, where k is the wavenumber of the radiation. If this condition, that the phase difference in a given direction varies linearly with the position in the slit, is met, what we are describing is termed *Fraunhofer diffraction*. The complex amplitude at P contributed by an element of the slit of width dy , located at A , is thus proportional to

$$\exp(iky \sin \theta) dy.$$

(We are ignoring the reduction of amplitude with distance due to geometrical spreading, as well as one or two other effects.) The total amplitude at P is found by integrating this expression over the entire slit:

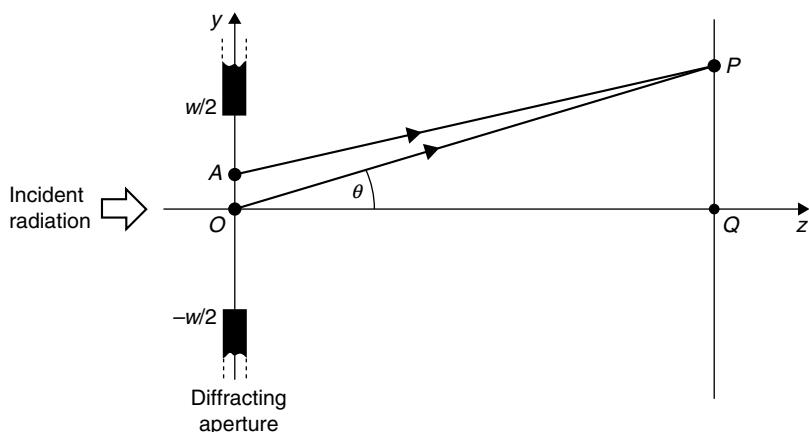


Figure 2.16. Geometry of Fraunhofer diffraction at a slit.

2.7 Diffraction

$$a(\theta) = \int_{-w/2}^{w/2} \exp(iky \sin \theta) dy.$$

This expression looks very similar to the Fourier transform defined in Equation (2.18). We can make the correspondence exact by introducing the idea of an *amplitude transmittance function* $f(y)$ for the plane of the screen, which defines the fraction of the incident amplitude that is transmitted. For the slit we have been discussing, $f(y) = 1$ for $-w/2 < y < w/2$, and 0 everywhere else. Using $f(y)$ to characterise any general one-dimensional aperture distribution, the expression for the complex amplitude in the direction θ becomes

$$a(\theta) = \int_{-\infty}^{\infty} f(y) \exp(iky \sin \theta) dy \quad (2.44)$$

which is clearly a Fourier transform, though often called the *Fraunhofer diffraction integral*.

In Section 2.3 we identified time t and angular frequency ω as a pair of conjugate variables related by the Fourier transform; here, the corresponding variables are y and $(k \sin \theta)$. Again, a form of uncertainty principle applies. Evaluating the integral (2.44) for our slit of width w , we find that

$$a(\theta) \propto \sin c\left(\frac{kw \sin \theta}{2}\right).$$

This is a function that has the same shape as Figure 2.6, and it first falls to zero when $\sin \theta = \pm 2\pi/kw = \pm \lambda/w$. If $w \gg \lambda$, $\sin \theta$ will be much less than 1 so that we can put $\sin \theta \approx \theta$, and hence

$$\delta\theta \approx \frac{\lambda}{w}. \quad (2.45)$$

This is the result that corresponds to Equation (2.19), and it shows that, if a beam of plane parallel radiation of wavelength λ passes through an aperture of width w , it will spread into a diverging beam whose angular width will be of the order of λ/w radians.

Equation (2.44) applies to one-dimensional diffraction, i.e. to the case in which the amplitude transmission function depends only on y . For the two-dimensional case in which the amplitude transmission function must be written as $f(x, y)$, the diffraction integral becomes

$$a(\theta_x, \theta_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \exp(ikx \sin \theta_x) \exp(iky \sin \theta_y) dx dy. \quad (2.46)$$

This double integral is rather hard to solve in general, although there are two special cases that should be mentioned. The first is when $f(x, y)$ can be factorised into two independent parts: $f(x, y) = g(x) h(y)$. The double integral can then be factorised into the product of two single integrals of the form of Equation (2.44). This approach allows us to calculate, for example, the diffraction pattern of rectangular apertures. The second special case is when the amplitude transmission function has circular symmetry. In this case it is simpler to use

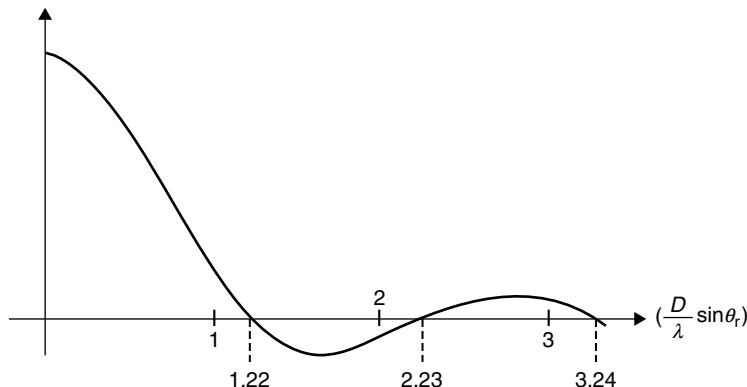


Figure 2.17. The Fraunhofer diffraction pattern of a circular aperture of diameter D , plotted as a function of $(D \sin \theta_r)/\lambda$, where θ_r is the radial angle, i.e. the angle from the normal to the plane of the aperture.

polar coordinates. We shall need only one result for general reference, and that is the diffraction pattern of a uniform circular aperture of diameter D . The amplitude of the diffracted wave in this case is given by

$$a(\theta_r) \propto \frac{J_1(\xi)}{\xi}, \quad (2.47)$$

where $J_1(x)$ is the first-order Bessel function, $\xi = \frac{kD \sin \theta_r}{2}$ and θ_r is the radial angle. The function in Equation (2.47) is sketched in Figure 2.17. $J_1(x)$ first falls to zero when $x = 3.832$, so the first zero occurs when $\sin(\theta_r) = 7.66/kD = 1.22\lambda/D$.

Now we return to the comment we made regarding Figure 2.16, that the distance z must be large enough for the two rays OP and AP to be regarded as parallel. How large is this? We assume conventionally that the Fraunhofer description is valid if the phase differences computed using it are accurate to within $\pi/2$ radians. Inspection of Figure 2.16 shows that this is equivalent to putting

$$AQ - OQ < \frac{\lambda}{4}$$

and since OA takes a maximum value of $w/2$ we may use Pythagoras' theorem to derive

$$\sqrt{\left(\frac{w}{2}\right)^2 + z^2} - z < \frac{\lambda}{4}.$$

Now if $w/2 \ll z$, we can use the binomial approximation to simplify this condition to

$$\frac{w^2}{8z} < \frac{\lambda}{4}$$

or

$$z > \frac{w^2}{2\lambda} \equiv z_F. \quad (2.48)$$

2.7 Diffraction

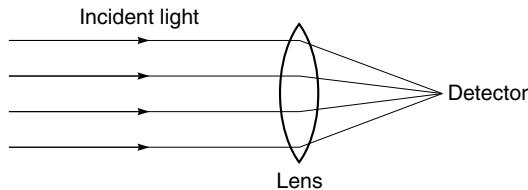


Figure 2.18. Ray diagram for a simple model of a remote sensing system.

The distance z_F is often called the *Fresnel distance*, after A. Fresnel who made many important discoveries in physical optics in the early nineteenth century, and if the condition (2.48) is not satisfied, a more rigorous form of diffraction theory, known as *Fresnel diffraction*, must be used. The region in which $z < z_F$ is often called the *near field*, and $z > z_F$ is the *far field*.

As was mentioned earlier, one important practical implication of diffraction is that it limits the spatial resolution of a remote sensing system. Without developing a rigorous theory for this phenomenon, we can see the principles involved by considering a very simple system (Figure 2.18) consisting of a lens arranged so that it focusses plane parallel light onto an extremely small detector.

Figure 2.18 is a ray diagram, so it includes the phenomena of geometrical optics but not the effects of diffraction. As it stands, the diagram implies that all of the radiation that reaches the detector was originally travelling in the same direction, i.e. that it subtended an angular width of zero. In fact, a range of incident directions will contribute to the signal that reaches the detector. By imagining that the light is propagating in the opposite direction to that shown in the figure, from the detector to the lens, we can see that the effect of the finite aperture represented by the lens will be to spread the outgoing light into a cone with an angular width of the order of λ/D , where D is the diameter of the lens. Thus, in general, we expect that diffraction will limit the angular resolution of any remote sensing system to $\sim \lambda/D$, where D is the width of the lens, antenna, mirror, or whatever is at the ‘front end’ of the system to define the spatial extent of the wavefront captured by the system. Other parts of the system may further degrade the resolution, of course.

We can illustrate this calculation with two examples. The first is a spaceborne optical sensor operating at a wavelength of $0.5 \text{ }\mu\text{m}$ with a lens diameter of 5 cm . Using the formula λ/D we find that the angular resolution is limited to about 10^{-5} radians (about 2 seconds of arc), which corresponds to a spatial resolution of about 10 m at a distance of 1000 km . (The second of arc, sometimes called an arcsecond or an arcsec, is a common unit of angular resolution. It is equal to $1/3600$ of a degree.) The Fresnel distance (Equation 2.45) is 2.5 km , so the simple λ/D calculation is valid. This is in fact typical of the spatial resolution of many spaceborne optical remote sensing systems. The second example is a passive microwave radiometer operating at a wavelength of 3 cm with an antenna diameter of 1 m . In this case, the angular resolution is about 0.03 radians (1.7 degrees), corresponding to a spatial resolution of 30 km at a distance of 1000 km (the Fresnel distance is 17 m). Thus we can see why the passive microwave systems to be discussed in Chapter 7 have very much poorer angular resolution than the optical and infrared systems described in Chapter 6.

Finally, it should be noted that some microwave remote sensing systems have been designed to circumvent the diffraction limit. The methods by which this is possible are discussed in Chapters 8 and 9.

Summary

When electromagnetic radiation encounters an obstruction or aperture, its angular distribution is changed. This is especially relevant in determining the angular resolution of imaging systems, since the radiation is collected by an antenna, lens or other structure which limits the spatial extent over which it can be received. If the wavelength of the radiation is λ and the width of the radiation-collecting structure is w , the angular resolution set by the diffraction limit is of order λ/w (radians) provided that the distance from the structure to the object being investigated is greater than about w^2/λ (the Fresnel distance). However, not all imaging instruments are diffraction-limited.

Review questions

Describe the electromagnetic spectrum as it is relevant to remote sensing.

Explain what is meant by *polarisation* of electromagnetic radiation.

Explain qualitatively how the Fourier transform can be used to analyse the frequency components present in a time-varying signal. If a signal has a finite duration ΔT , what implication does this have for the range of frequencies present in the signal?

Describe the Doppler effect.

Explain the concepts of radiance, irradiance and radiant exitance.

Describe the phenomenon of black-body radiation.

Discuss Stefan's law and Wien's law as they relate to black-body radiation.

Explain what is meant by brightness temperature.

Explain what is meant by diffraction. Why is it relevant to the angular resolving power of a remote sensing instrument?

Problems

1. The electric field of an electromagnetic wave in free space is given by

$$E_x = 0$$

$$E_y = E \cos(\omega t - kx)$$

$$E_z = 2E \sin(\omega t - kx)$$

where $E = 1 \text{ kV m}^{-1}$. Find (i) the direction of propagation, (ii) the polarisation, (iii) the magnetic field and (iv) the flux density of the radiation.

2. When radiation having a Stokes vector $\mathbf{S} = [S_0, S_1, S_2, S_3]$ is incident on an antenna that receives only linearly x -polarised radiation, the detected power is proportional to $\mathbf{S}\cdot\mathbf{P}$, where $\mathbf{P} = [1, 1, 0, 0]$. Show how the detected power varies with the polarisation state for radiation of a given flux density.
3. Show that Equation (2.31) can be approximated by Equation (2.33) at sufficiently low frequencies. Hint: $\exp(x) \approx 1 + x$ when $|x| \ll 1$.
4. Calculate the ratio of the spectral radiances of black bodies at 300 K and 6000 K at (i) a wavelength of 0.1 μm , (ii) a wavelength of 1 μm , (iii) a frequency of 1000 GHz, (iv) a frequency of 1 GHz.
5. Show that the Fourier transform of the Gaussian function

$$f(t) = \exp - \left(\frac{(t - t_0)^2}{2\sigma^2} \right)$$

is proportional to

$$\exp \left(-\frac{i\omega t_0}{2} - \frac{\omega^2 \sigma^2}{2} \right)$$

and interpret the result.

6. When is Equation (2.46) valid?
7. The Planck formula can be inverted to obtain the brightness temperature from the spectral radiance as follows:

$$T_b = \frac{K_2}{\ln \left(\frac{K_1}{L_\lambda} + 1 \right)}$$

Show that, for observations at a wavelength of 10.7 μm , K_1 and K_2 have values of 851 $\text{W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$ and 1347 K respectively. Hence calculate the brightness temperature of a body whose spectral radiance is 5.84 $\text{W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$ at a wavelength of 10.7 μm .

3

Interaction of electromagnetic radiation with matter

The interaction of electromagnetic radiation with matter is evidently fundamental to remote sensing. The subject is a vast one, embracing many areas of physics, and a fully systematic treatment of it would require at least a book in itself. In this chapter, therefore, we attempt to provide an overview that will be sufficient to gain an understanding of the operation of remote sensing systems. In order to keep the chapter to a manageable length, we reserve a discussion of the interaction of electromagnetic radiation with the Earth's atmosphere to Chapter 4. Nevertheless, this is still a long chapter, and it is also the most technical in the book. It is not necessary to understand all of the material in this chapter in order to follow the subsequent material but, as usual, a reading of the summaries at the end of each section should give a general understanding of the material.

The chapter first extends some of the concepts of Chapter 2 into a consideration of how electromagnetic radiation propagates in homogeneous dielectric media. The key concepts here are the dielectric constant (also known as the relative electric permittivity) and the refractive index. By allowing these parameters to take complex, rather than purely real, values, the concept of absorption of radiation is included, and by allowing them to vary with frequency (or equivalently with wavelength) we are able to include the idea of dispersion, which will be important in Chapter 8 where we consider ranging systems.

Sections 3.2 and 3.3 discuss how radiation interacts with boundaries between two media, particularly the important case where one of the media is free space. Section 3.2 examines the behaviour of radiation at planar interfaces while Section 3.3 deals with rough surfaces. Optical and microwave cases are discussed separately.

In S3.4 we consider the way electromagnetic radiation interacts with particles, which will in general scatter some of the radiation into different directions but may also absorb some of the radiation. This is true of particles as small as atoms and molecules as well as of larger particles such as water droplets. In Section 3.5 these concepts are related to the properties of the medium treated as a continuum, leading to a discussion of the radiative transfer equation. Finally, Section 3.6 draws on the ideas discussed in the earlier parts of the chapter to describe the interaction of electromagnetic radiation with a number of real materials.

3.1

Propagation through homogeneous materials

To describe propagation in a uniform homogeneous material we need to introduce two properties of the medium: its *relative electric permittivity* ϵ_r and its *relative magnetic permeability* μ_r . The relative electric permittivity, which is also known as the *dielectric constant*, is the ratio of the electric permittivity ϵ of the material to ϵ_0 , the permittivity of free space, and the relative magnetic permeability is the ratio of the magnetic permeability μ to μ_0 , the permeability of free space. Thus we can put

$$\mu = \mu_r \mu_0 \quad (3.1)$$

and

$$\epsilon = \epsilon_r \epsilon_0. \quad (3.2)$$

Both ϵ_r and μ_r are pure numbers, i.e. they are dimensionless. Note that some authors use the symbols ϵ and μ to represent the relative, rather than the absolute, values of the permittivity and the permeability.

An electromagnetic wave can propagate in such a medium. For consistency with Chapter 2, we write the equation for the electric field in exactly the same form as for the free-space wave defined in Equation (2.1):

$$E_x = E_0 \cos(\omega t - kz). \quad (3.3)$$

However, the equation for the magnetic field now differs from equation (2.2), becoming

$$B_y = \frac{E_0 \sqrt{\epsilon_r \mu_r}}{c} \cos(\omega t - kz). \quad (3.4)$$

The ratio of the amplitudes of the electric and magnetic fields has become

$$\frac{c}{\sqrt{\epsilon_r \mu_r}}$$

instead of c as it is for an electromagnetic wave propagating in free space.

In order for Equations (3.3) and (3.4) to represent a valid solution of Maxwell's equations, the angular frequency ω and the wavenumber k must be related by

$$\frac{\omega}{k} = \frac{c}{\sqrt{\epsilon_r \mu_r}} \quad (3.5)$$

This is the wave speed, or more precisely the *phase velocity* of the wave (see Section 3.1.3), which we give the symbol v . The *refractive index* n of the medium is defined as c/v , thus

$$n = \sqrt{\epsilon_r \mu_r} \quad (3.6)$$

Clearly, free-space propagation is the special case $n = 1$.

Most of the discussion of polarisation presented in Chapter 2 can be applied equally well to the case of a medium in which the refractive index n is not equal to 1. The electric and magnetic fields are still perpendicular to one another and to the direction of propagation, though of course the ratio of the amplitudes is now v rather than c . The Stokes

parameters are still defined by Equations (2.14). However, the flux density of the radiation (the power transmitted per unit area normal to the propagation direction) is

$$F = \frac{S_0}{2Z}, \quad (3.7)$$

where Z , the impedance of the medium, is given by

$$Z = Z_0 \sqrt{\frac{\mu_r}{\epsilon_r}}. \quad (3.8)$$

3.1.1 Complex dielectric constants: absorption

We have seen how the behaviour of an electromagnetic wave in a uniform homogeneous medium is controlled by the refractive index n , defined by Equation (3.6). For the media we shall need to consider, the relative magnetic permeability μ_r can be taken as 1 (these are so-called ‘non-magnetic materials’), and we can therefore focus our attention on the dielectric constant ϵ_r .

For the case of a uniform homogeneous medium that does not absorb any energy from an electromagnetic wave propagating through it, the dielectric constant must be a real number. However, if the medium does absorb energy from the wave, we must use a complex number to represent the dielectric constant. The conventional way of doing this is to put

$$\epsilon_r = \epsilon' - i\epsilon'' \quad (3.9)$$

where $i^2 = -1$ and ϵ' and ϵ'' are referred to as the real and imaginary parts of the dielectric constant although, strictly speaking, ϵ'' is the negative of the imaginary part. Another commonly encountered way of writing the complex dielectric constant is to put

$$\epsilon_r = \epsilon'(1 - i\tan \delta) \quad (3.10)$$

where $\tan \delta$ is called the *loss tangent*. This notation is clearly equivalent to Equation (3.9), and is also equivalent to making the refractive index complex (we are still assuming that $\mu_r = 1$):

$$n = m - ik, \quad (3.11)$$

which we can see as follows.

From Equation (3.6), and taking $\mu_r = 1$, we know that $n^2 = \epsilon_r$. Squaring Equation (3.11), and equating it to Equation (3.9), gives $n^2 = m^2 - \kappa^2 - 2im\kappa = \epsilon' - i\epsilon''$ and so the two descriptions are equivalent provided that

$$\epsilon' = m^2 - \kappa^2 \quad (3.12.1)$$

$$\epsilon'' = 2m\kappa. \quad (3.12.2)$$

(It should be noted here that although $n = m - ik$ is a common notation for the complex refractive index, others are in use. Some authors reverse the roles of n and m , and in another convention the imaginary parts of both n and ϵ_r are positive, not negative.)

3.1 Propagation through homogeneous materials

We can also see how a complex dielectric constant represents propagation with absorption by considering an x -polarised wave propagating in the z -direction. The electric field can be written using the complex exponential notation

$$E_x = E_0 \exp(i(\omega t - kz)), \quad (3.13)$$

where E_0 is a constant. This is equivalent to Equation (3.3). From Equations (3.5) and (3.6) we know that $k = \omega n/c$, which can be rewritten using Equation (3.11) as $k = \frac{\omega}{c}(m - ik)$.

Substituting this expression for k into Equation (3.13), and rearranging, we obtain

$$E_x = E_0 \exp\left(-\frac{\omega kz}{c}\right) \exp\left(i\omega(t - mz/c)\right).$$

This represents a simple harmonic wave whose amplitude decreases exponentially with the distance z . Since we are usually more interested in the flux density F of the wave, and this is proportional to the square of the amplitude, we can put

$$F = F_0 \exp\left(-\frac{2\omega kz}{c}\right) \quad (3.14)$$

where F_0 is a constant. This is a version of the *Beer–Lambert law*, which is known by various names that permute some or all of the names Beer, Lambert and Bouguer. The Beer–Lambert law is discussed in more detail in Section 3.5.1.

Equation (3.14) shows that the flux density of the electromagnetic wave is reduced by a factor of e (≈ 2.718) as the wave advances a distance l_a through the medium, where l_a , the *absorption length*, is given by

$$l_a = \frac{c}{2\omega k}. \quad (3.15)$$

If there are no other factors influencing the intensity of the radiation (i.e. we can ignore scattering and emission for the present – although we shall discuss these phenomena in Section 3.4), the absorption length provides an order-of-magnitude estimate for the distance radiation will propagate through the material before its intensity is significantly reduced. For example, after travelling two absorption lengths, the intensity is reduced by a factor of e^2 , which means that it has been reduced to about 14% of its original value. After five absorption lengths, the intensity is only 0.7% of its original value, and so on. Figure 3.1 shows the variation with frequency of the absorption length in various materials. (The data for water are from Segelstein (1981) and for ice from data compiled by Rott *et al.* (1988) and by Rees (2006).)

3.1.2 Dielectric constants and refractive indices of real materials

The dielectric constant of a real material is normally dependent on the frequency, and the variation can be quite complicated (for example, see Figure 3.2, and Feynman, Leighton and Sands (2005) for a very clear discussion of the physical principles that determine the dielectric constants of real materials). However, over a limited frequency range a simple physical model can often give a satisfactory description. Although the foregoing theory has been developed in terms of the angular frequency ω , it is more usual to specify the frequency f , or, especially for optical and infrared radiation, the *free-space wavelength* λ_0 . This is the

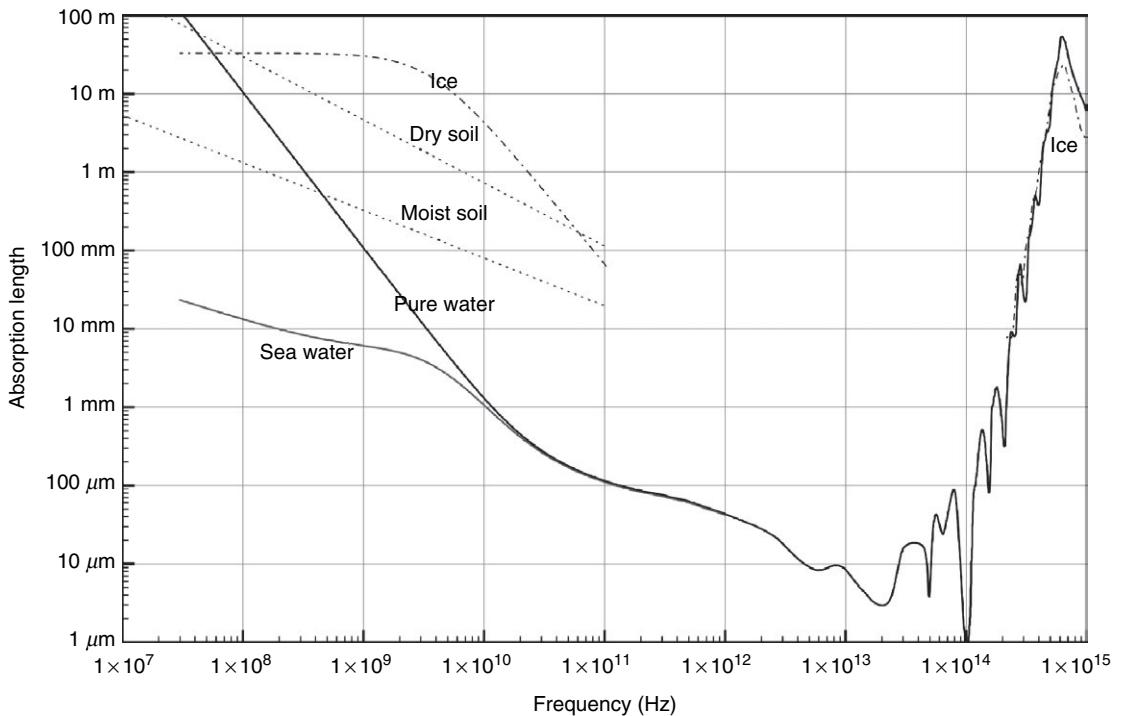


Figure 3.1. Absorption lengths (schematic) of various materials. Note that the absorption lengths are strongly influenced by such factors as temperature and the content of trace impurities, especially at low frequencies.

wavelength that electromagnetic radiation of the same frequency would have if it were propagating in free space, and is given by Equations (2.4) and (2.6) as

$$\lambda_0 = \frac{2\pi c}{\omega} = \frac{c}{f}. \quad (3.16)$$

The wavelength of the radiation in the medium is given by

$$\lambda = \frac{\lambda_0}{m}. \quad (3.17)$$

3.1.2.1 Gases

The dielectric constant of a gas is given to reasonable accuracy, provided that the radiation is not strongly absorbed by the gas, by

$$\epsilon_r = 1 + \frac{N\alpha}{\epsilon_0}, \quad (3.18)$$

where N is the number density of the gas molecules (i.e. the number of molecules per unit volume) and α is the *polarisability* of the gas molecule. Since the term $N\alpha/\epsilon_0$ is much smaller than 1 for a gas, the refractive index is given, to a good approximation, by

3.1 Propagation through homogeneous materials

Table 3.1. Values of α/ϵ_0 (where α is the molecular polarisability) of various gases at optical and radio frequencies. The values are given in units of 10^{-30} m^3

Gas	Optical	Radio
Air	21.7	21.4
Carbon dioxide	33.6	36.8
Hydrogen	9.8	10.1
Oxygen	20.2	19.8
Water vapour	18.9	368

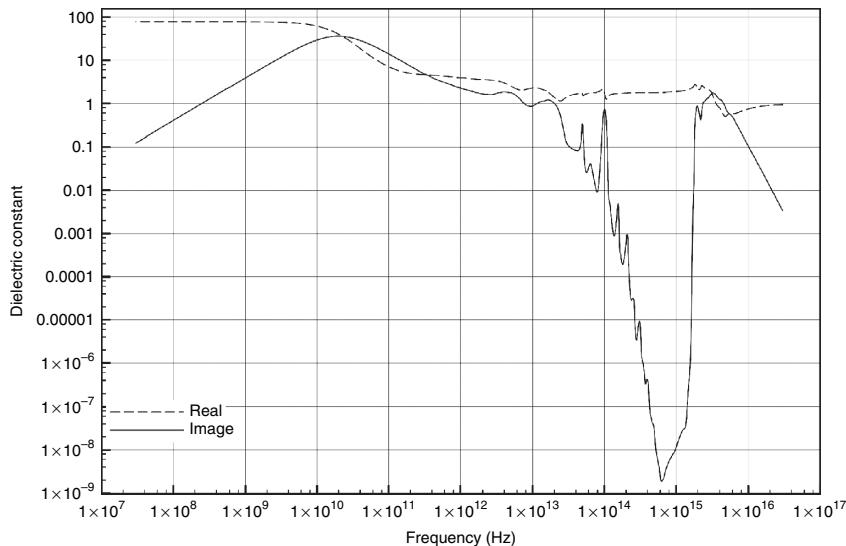


Figure 3.2. Real and imaginary parts of the dielectric constant of pure water, following Segelstein (1981).

$$n = 1 + \frac{N\alpha}{2\epsilon_0}. \quad (3.19)$$

The quantity α/ϵ_0 has the dimensions of a volume, and is normally similar to the actual physical volume of the molecule. Table 3.1 shows typical values of this quantity for various gases at optical ($\lambda_0 = 589 \text{ nm}$) and radio (f typically 1 MHz) frequencies.

3.2.1.2 Solids and liquids that are electrical insulators

Simple *non-polar* materials, i.e. those that are composed of molecules that do not have a permanent dipole moment, are characterised by a constant (possibly complex) value of ϵ_r . Simple *polar* materials can be described by the *Debye equation* (3.20), which represents a resonant phenomenon with a time-constant (relaxation time) τ :

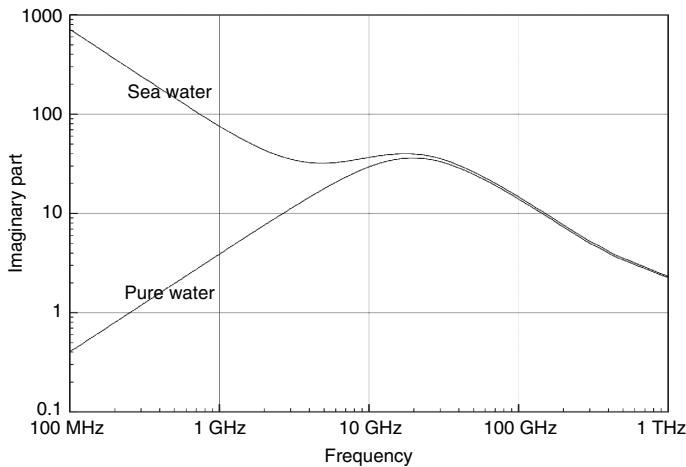


Figure 3.3. Imaginary parts of the dielectric constants of pure water and sea water below 1 THz.

$$\epsilon' = \epsilon_\infty + \frac{\epsilon_p}{1 + \omega^2\tau^2}, \quad (3.20.1)$$

$$\epsilon'' = \frac{\omega\tau\epsilon_p}{1 + \omega^2\tau^2}. \quad (3.20.2)$$

In these equations, ϵ_∞ is the dielectric constant at ‘infinite’ frequency (in practice, at frequencies much greater than $1/\tau$), and ϵ_p is the polar contribution to the dielectric constant. For example, pure water follows the Debye equation fairly closely between 1 MHz and 1000 GHz, with values (at 20 °C) of $\epsilon_\infty = 5.0$, $\epsilon_p = 75.4$, $\tau = 9.2 \times 10^{-12}$ s. The corresponding variation of ϵ' and ϵ'' is shown in Figure 3.2, while Figure 3.3 shows how the imaginary part of the dielectric constant of sea water (which is an electrical conductor) diverges from that of pure water at frequencies below about 20 GHz.

3.2.1.3 Metals

The electrical properties of metals are dominated by the very high densities of delocalised electrons. In general, the dielectric constant of a metal can be written as

$$\epsilon' = 1 - \frac{\sigma\tau}{\epsilon_0(1 + \omega^2\tau^2)}, \quad (3.21.1)$$

$$\epsilon'' = \frac{\sigma}{\epsilon_0\omega(1 + \omega^2\tau^2)}. \quad (3.21.2)$$

In these expressions, σ is the electrical *conductivity* of the metal, and

$$\tau = \frac{m_e\sigma}{Ne^2} \quad (3.22)$$

where m_e is the mass of the electron, e is the charge on the electron, and N is the number density of delocalised electrons in the metal. Equations (3.21) can be simplified for the cases where $\omega \gg 1/\tau$ and where $\omega \gg 1/\tau$. For metals, τ has a value of typically 10^{-15} to 10^{-14} s, so these cases correspond to radio frequencies and optical or ultraviolet frequencies respectively. At low (radio) frequencies, we obtain

3.1 Propagation through homogeneous materials

$$\epsilon_r \approx -\frac{i\sigma}{\epsilon_0 \omega}. \quad (3.23)$$

From Equations (3.12) we see that this corresponds to real and imaginary components of the refractive index of

$$m = \kappa = \sqrt{\frac{\sigma}{2\epsilon_0\omega}},$$

and hence from Equation (3.15) the absorption length is given by

$$l_a = c \sqrt{\frac{\epsilon_0}{2\sigma\omega}}.$$

For example, let us consider electromagnetic radiation at a frequency of 5 GHz ($\omega = 3.14 \times 10^{10} \text{ s}^{-1}$) propagating in stainless steel ($\sigma = 1.0 \times 10^6 \Omega^{-1} \text{ m}^{-1}$). We find that the absorption length is 3.6 μm, which shows that the material is opaque to radio frequency radiation unless it is extremely thin.

At high (optical or ultraviolet) frequencies, the dielectric constant of a metal can be approximated as

$$\epsilon_r \approx 1 - \frac{Ne^2}{\epsilon_0 m_e \omega^2}, \quad (3.24)$$

which is real and very slightly less than 1. Thus, at sufficiently high frequencies, metals become transparent.

Equation (3.24) also describes the dielectric constant of a *plasma*, a state of matter in which all the atoms have been ionised. Because the mass of the electron is so much smaller than that of any other charged particle, the latter may effectively be ignored in considering the response of the material to an electromagnetic wave. It is clear from Equation (3.24) that a plasma is transparent (ϵ_r is real) for angular frequencies higher than

$$\omega_p = \sqrt{\frac{Ne^2}{\epsilon_0 m_e}} \quad (3.25)$$

and absorbs radiation at angular frequencies below this value. ω_p is called the *plasma frequency*, and considerations of the properties of plasmas will be important when we discuss the ionosphere in Chapters 4 and 8.

3.1.3 Dispersion

We noted earlier that, in a number of cases of practical importance, the dielectric properties (and hence refractive index) of a medium vary with frequency. Such media are said to be *dispersive*, and a wave propagating in such a medium is called a *dispersive wave*. It is usual to characterise this behaviour by expressing the angular frequency ω as a function of the wavenumber k , and this relationship is called the *dispersion relation*.

We saw in Equation (3.5) that the wave velocity v is given by

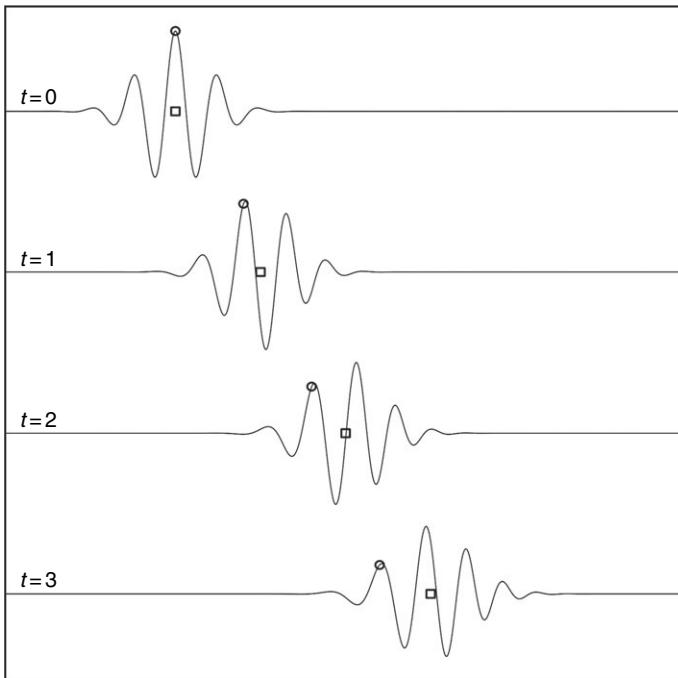


Figure 3.4. Dispersion of a modulated wave. The figure shows a sinusoidal wave that has been modulated by a Gaussian envelope, at four successive instants of time. After time t , a particular wave crest (shown by the circle) has moved a horizontal distance vt , whereas the centre of the modulating function (shown by the square) has moved a distance v_gt . In this case v is the phase velocity and v_g is the group velocity.

$$v = \frac{\omega}{k}. \quad (3.26)$$

This is true even if ω varies with k , and v (the wave velocity or *phase velocity*) is the speed at which the crests and troughs of the wave move in the propagation direction. However, if we modulate the wave in some way, for example by breaking it up into pulses, it is this modulation which carries information, and we therefore need to know the speed at which the modulating function travels. This is called the *group velocity*, and it is given by

$$v_g = \frac{d\omega}{dk}. \quad (3.27)$$

Only in the case of a non-dispersive wave, for which ω is proportional to k , will (3.26) and (3.27) be in general equal to one another.



Figure 3.4 illustrates the idea of a dispersive wave. It shows a sinusoidal wave that has been modulated by a Gaussian envelope to give a pulse. The pulse is travelling to the right, and is shown at four equally spaced intervals of time. The little rings drawn on each position of the pulse show the progress of a particular crest of the wave, and it can be seen that this crest is travelling more slowly than the envelope itself. Thus, in this particular case, the phase velocity is less than the group velocity. The phase velocity in the example of Figure 3.4 increases with increasing frequency. This can be seen by the fact that the higher frequencies (shorter

3.1 Propagation through homogeneous materials

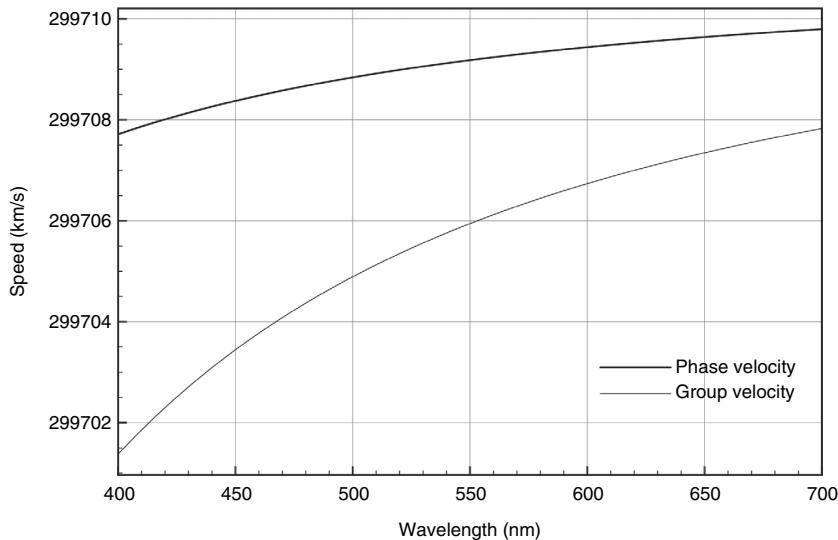


Figure 3.5. Phase and group velocities for light propagating in dry air at standard atmospheric pressure and 15 °C.

wavelengths) move towards the front of the pulse, and the lower frequencies towards the back, as the pulse travels. Figure 3.4 also illustrates another consequence of a wave travelling in a dispersive medium, which is the spreading (elongation) of the envelope as time progresses. Whether this phenomenon occurs depends on the precise form of the dispersion relation.

It will sometimes happen in practice that information about the dispersion relation for a particular medium will be given not as $\omega(k)$, but as $n(\lambda_0)$, the dependence of the refractive index on the free-space wavelength. In this case, (3.27) may conveniently be expressed as

$$\frac{c}{v_g} = n - \lambda_0 \frac{dn}{d\lambda_0}. \quad (3.28)$$

As an example, we can consider the dispersion of visible light in air. In the optical region, the refractive index of dry air at atmospheric pressure and 15 °C can be approximated (e.g. see http://www.kayelaby.npl.co.uk/general_physics/2_5/2_5_7.html) as

$$n = 1 + \frac{1}{a - b/\lambda_0^2},$$

where $a = 3669$ and $b = 2.1173 \times 10^{-11} \text{ m}^2$. Applying Equation (3.28) gives

$$\frac{c}{v_g} = 1 + \frac{(a + b/\lambda_0^2)}{(a - b/\lambda_0^2)^2}.$$

Figure 3.5 shows the phase velocity v and the group velocity v_g calculated from these formulae, for free-space wavelengths between 0.4 and 0.7 μm.

Returning to Equation (3.24), which describes the dielectric constant of a metal (at sufficiently high frequencies) or a plasma, we note that above the plasma frequency the dielectric constant is real and less than 1. This means that the phase velocity v is greater than c , the speed of light, and at first sight this seems to contradict Einstein's principle that

nothing can travel faster than c . However, as we remarked earlier, *information* is carried at the group velocity, not the phase velocity. It is easy to show, using Equation (3.27), that if the dielectric constant is given by Equation (3.24) the relationship between v and v_g is

$$vv_g = c^2, \quad (3.29)$$

so the group velocity is indeed less than c .

Summary

Propagation of electromagnetic radiation through a homogeneous non-magnetic material is characterised by the medium's dielectric constant ϵ_r , or equivalently by its refractive index n , related through $\epsilon_r = n^2$. If these are purely real the medium does not absorb radiation. Absorption is represented by making the refractive index, and hence the dielectric constant, complex. The notations for these complex quantities are $\epsilon_r = \epsilon' - i\epsilon''$ and $n = m - ik$ respectively. If n varies with frequency the medium is said to be dispersive, and the phase velocity, at which waves propagate, and the group velocity, at which modulations of the wave propagate (and hence pulses, and information in general), are not necessarily the same. The group velocity cannot exceed the speed of light.

3.2

Plane boundaries

In this section we review the phenomena of reflection and transmission when electromagnetic radiation encounters a plane boundary between two uniform homogeneous media (Figure 3.6). We call these media 1 and 2. The radiation is travelling in medium 1 towards the boundary with medium 2, and makes an angle θ_1 with the normal to the boundary. In general, some of the radiation will be reflected back into medium 1, again at an angle θ_1 but on the opposite side of the normal, and some will be refracted across the boundary so that it makes an angle θ_2 in medium 2.

Snell's law relates the angles θ_1 and θ_2 through



$$n_1 \sin \theta_1 = n_2 \sin \theta_2, \quad (3.30)$$

where n_1 and n_2 are the refractive indices of the two media, and also states that the incident, reflected and refracted rays, and the normal to the boundary, all lie in the same plane.

We also need to know the reflection and transmission coefficients, r and t . The reflection coefficient r is defined as the electric field amplitude of the reflected radiation, expressed as a fraction of the electric field amplitude of the incident radiation, and similarly for the transmission coefficient t . Since the values of these coefficients depend on the polarisation of the incident radiation, we need to specify each coefficient for two orthogonal polarisations, giving a total of four coefficients. (The coefficients for any other polarisation state can be calculated by resolving the state into the components for the two states we have chosen, as shown in Section 2.2.) The two polarisations that are usually chosen are called perpendicular and parallel, denoted by the symbols \perp and \parallel (Figure 3.7).

3.2 Plane boundaries

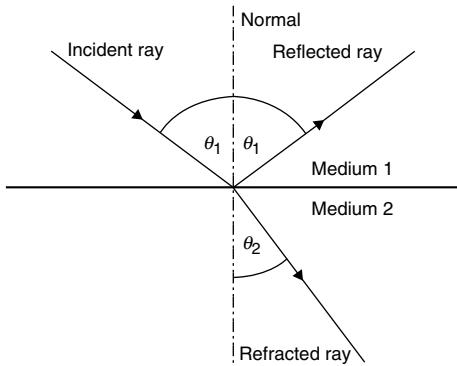


Figure 3.6. Reflection and refraction at a plane boundary between two media.

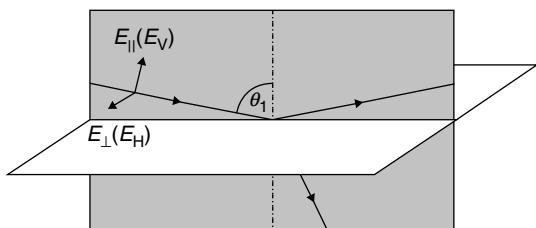


Figure 3.7. Parallel and perpendicular (vertical and horizontal) polarisations of radiation incident at and reflected from a plane boundary between two media. The boundary between the media is shown in white, and the perpendicular plane containing the incident, reflected and refracted rays is shown in grey.

The term ‘parallel polarisation’ means that the electric field vector of the radiation is parallel to the plane containing the incident, reflected and refracted rays (and the normal to the boundary), and ‘perpendicular radiation’ means that the electric field vector is perpendicular to this plane. Sometimes, especially in describing microwave systems, the terms ‘horizontal polarisation’ and ‘vertical polarisation’ are used instead. To understand this notation it is necessary to think of the boundary as being horizontal, and to realise that ‘vertically’ polarised radiation merely has a vertical component. Provided that the two media are homogeneous, parallel-polarised incident radiation will give rise to parallel-polarised reflected and refracted radiation, and no perpendicular-polarised radiation. The converse of this statement is also true, so that perpendicular-polarised incident radiation does not produce any parallel-polarised components.

Now that we have defined our terms, we can proceed to state the formulae for the reflection and transmission coefficients. These are calculated, in terms of the impedances Z_1 and Z_2 of the media, by solving Maxwell’s equations at the boundary:

$$r_{\text{perp}} = \frac{Z_2 \cos \theta_1 - Z_1 \cos \theta_2}{Z_2 \cos \theta_1 + Z_1 \cos \theta_2}, \quad (3.31.1)$$

$$t_{\text{perp}} = \frac{2Z_2 \cos \theta_1}{Z_2 \cos \theta_1 + Z_1 \cos \theta_2}, \quad (3.31.2)$$

$$r_{\text{par}} = \frac{Z_2 \cos \theta_2 - Z_1 \cos \theta_1}{Z_2 \cos \theta_2 + Z_1 \cos \theta_1}, \quad (3.31.3)$$

$$t_{\text{par}} = \frac{2Z_2 \cos \theta_1}{Z_2 \cos \theta_2 + Z_1 \cos \theta_1}, \quad (3.31.4)$$

The full expressions of Equations (3.31), which are called *Fresnel coefficients*, become rather complicated in the case where both media have significant absorption coefficients (i.e. the complex form of their refractive indices has to be taken into account). However, in many cases of practical importance we may assume that medium 1 has a refractive index of 1 (a vacuum, or air, to a good approximation). If medium 2 is absorbing, the Fresnel reflection coefficients for non-magnetic media are given by the following formulae:

$$r_{\text{perp}} = \frac{\cos \theta_1 - \sqrt{\epsilon_r 2 - \sin^2 \theta_1}}{\cos \theta_1 + \sqrt{\epsilon_r 2 - \sin^2 \theta_1}}, \quad (3.32.1)$$

$$t_{\text{perp}} = \frac{2 \cos \theta_1}{\cos \theta_1 + \sqrt{\epsilon_r 2 - \sin^2 \theta_1}}, \quad (3.32.2)$$

$$r_{\text{par}} = \frac{\sqrt{\epsilon_r 2 - \sin^2 \theta_1} - \epsilon_r 2 \cos \theta_1}{\sqrt{\epsilon_r 2 - \sin^2 \theta_1} + \epsilon_r 2 \cos \theta_1}, \quad (3.32.3)$$

$$t_{\text{par}} = \frac{2 \sqrt{\epsilon_r 2 - \cos \theta_1}}{\sqrt{\epsilon_r 2 - \sin^2 \theta_1} + \epsilon_r 2 \cos \theta_1}. \quad (3.32.4)$$

Note that these expressions must in general be evaluated using complex arithmetic. If medium 1 has $n = 1$ and medium 2 is non-absorbing, the reflection coefficients become

$$r_{\text{perp}} = \frac{\cos \theta_1 - \sqrt{n_2^2 - \sin^2 \theta_1}}{\cos \theta_1 + \sqrt{n_2^2 - \sin^2 \theta_1}}, \quad (3.33.1)$$

$$r_{\text{par}} = \frac{\sqrt{n_2^2 - \sin^2 \theta_1} - n_2^2 \cos \theta_1}{\sqrt{n_2^2 - \sin^2 \theta_1} + n_2^2 \cos \theta_1}. \quad (3.33.2)$$

We can see from Equation (3.33.2) that $r_{\text{par}} = 0$ when θ_1 takes the value θ_B , given by

$$\tan \theta_B = n_2. \quad (3.34)$$

This is called the *Brewster angle*. Parallel (vertically) polarised radiation incident on a surface at the Brewster angle cannot be reflected, and so must all be transmitted into the medium. Consequently, we can note that randomly polarised radiation incident from an arbitrary direction on a boundary between two media will in general, on reflection, be partially polarised, and if it is incident at the Brewster angle it will be completely plane polarised. This is the simplest justification for the remark we made in Section 2.2 that the degree of polarisation is changed on reflection.

To illustrate Equations (3.33) and the phenomenon of the Brewster angle, we calculate the power reflection coefficients for light meeting an air–water interface. The refractive index of

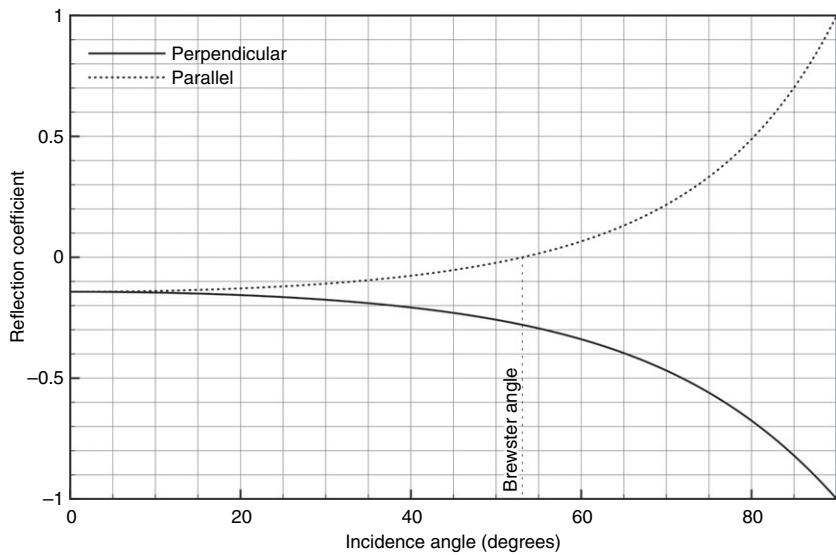


Figure 3.8. Amplitude reflection coefficients for light incident on a water surface.

air can be taken as 1 and that of pure water as 1.33 with no imaginary part. Figure 3.8 shows the reflection coefficients as a function of the incidence angle θ_1 .

The Brewster angle by the river

As we noted in Chapter 2, direct sunlight is not polarised. However, if the sunlight is reflected from a horizontal air–water interface at an angle near the Brewster angle, the reflected light will be almost totally polarised with its electric field vector horizontal. A suitably oriented polarising filter, such as a pair of polarising sunglasses, will thus eliminate almost all of this reflected component.

The photograph on the left shows sunlight reflected from the surface of a river, viewed through a polarising filter that blocks vertically polarised radiation. In the photograph on the right the filter has been rotated through 90° to block horizontally polarised radiation.



The figure shows that the value of r_{par} falls to zero near 53° , which is confirmed by calculating the Brewster angle from Equation (3.34) as 53.1° . The figure also shows that when $\theta = 90^\circ$ (grazing incidence) the intensity reflection coefficients are both 1 (i.e. all the radiation is reflected). At normal incidence, i.e. when $\theta = 0$, the figure shows that the two reflection coefficients have the same value, and this must be correct since for normally incident radiation there can be no distinction between parallel and perpendicular polarisations. In this case, Equation (3.33) shows that the amplitude reflection coefficient is given by

$$r_{\text{normal}} = \frac{1 - n_2}{1 + n_2}. \quad (3.35)$$

More generally, the intensity reflection coefficient at normal incidence to the boundary between two non-absorbing media is given by

$$r^2_{\text{normal}} = \left[\frac{n_1 - n_2}{n_1 + n_2} \right]^2 \quad (3.36)$$

and thus depends only on the *ratio* of the two refractive indices. If this ratio has a value of one, i.e. the two refractive indices are identical, nothing is reflected at the interface. On the other hand, if the ratio is very large or very small compared with one, i.e. there is a large *refractive contrast* (or dielectric contrast) between the two media, the reflection coefficient will be large. In the case of visible light incident on an air–water interface, the ratio of refractive indices is about 1.33 as we have already noted, and the intensity reflection coefficient at normal incidence is about 2.0%.

Summary

When electromagnetic radiation is incident on a planar interface between two media, it can be refracted into the second medium, reflected from the interface or, in general, both things can happen. The directions of the reflected and refracted rays are given by Snell's laws. The amplitudes of the rays are given by the Fresnel coefficients, and depend on the polarisation. For parallel-polarised (vertically polarised) radiation, the reflection coefficient is zero at the Brewster angle. The intensity reflection coefficient is 100% at glancing incidence, while at normal incidence it is $((n_1 - n_2)/(n_1 + n_2))^2$ for both polarisations, so that it depends on the contrast in refractive index between the two media.

3.3

Scattering from rough surfaces

Scattering (reflection) of radiation from the Earth's surface is a fundamental process in most remote sensing situations. The exceptions to this principle are atmospheric sounding observations, and those passive observations (of thermal infrared or microwave emission) that do not respond to reflected sunlight. Thus, a consideration of the reflectance properties of real surfaces will be of considerable importance. In fact, as we saw in Chapter 2,

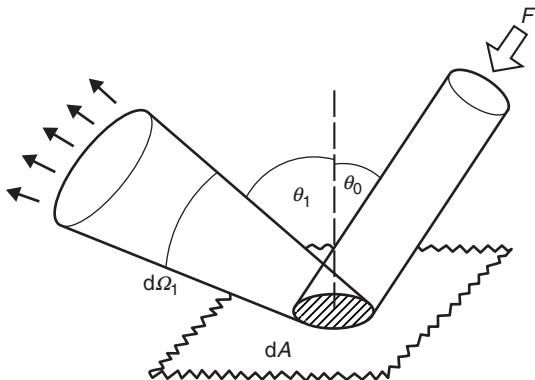


Figure 3.9. Radiation, initially of flux density F , is incident at angle θ_0 on an area dA and is then scattered into solid angle $d\Omega_1$ in the direction θ_1 . The azimuthal angles ϕ_0 and ϕ_1 are omitted for clarity.

the thermal emissivity of a surface is directly related to its reflectance, so these properties will be important even in the case of passive microwave and thermal infrared remote sensing.

In Section 3.2 we reviewed the behaviour of electromagnetic radiation when it is incident on a planar (i.e. perfectly smooth) boundary between two homogeneous media. In this section we consider radiation incident, from within a vacuum (to which air is a reasonably good approximation), on a rough surface. The material below this surface is assumed to be homogeneous, but we consider what happens when it is *not* homogeneous in Section 3.4.

3.3.1 Description of surface scattering

The first thing we need to do is to develop some of the terminology needed to describe rough surface scattering. The treatment presented in this section is amplified by Hapke (2005).

Figure 3.9 shows a well-collimated beam of radiation of flux density F , measured in a plane perpendicular to the direction of propagation, incident on a surface at an angle θ_0 . This angle is often called the *local incidence angle*, and its complement $\pi/2 - \theta_0$ the *depression angle*. A proportion of the incident radiation will be scattered into the solid angle $d\Omega_1$, in a direction specified by the angle θ_1 . For simplicity the azimuthal angles have been omitted from Figure 3.9. These will be denoted by φ_0 and φ_1 respectively.

The irradiance E at the surface is given by $F \cos \theta_0$. If we write $L_1(\theta_1, \varphi_1)$ for the radiance of the scattered radiation in the direction (θ_1, φ_1) , we can define the *bidirectional reflectance distribution function* (BRDF) R as

$$R = \frac{L_1}{E}. \quad (3.37)$$

R has no dimensions, and its unit is sr^{-1} . It is also commonly represented by the symbols f or r . In considering radar systems (Chapter 9), the BRDF is usually replaced by the equivalent *bistatic scattering coefficient* γ , which also has no units and is related to R by

$$\gamma = 4\pi R \cos\theta_1. \quad (3.38)$$

In fact most radar systems, and all those we shall consider in Chapter 9, detect only the *backscattered* component of the radiation, which retraces the path of the incident radiation. In this case $\theta_1 = \theta_0$ and $\phi_1 = \phi_0$, and the usual way of specifying the proportion of scattered radiation is through the (dimensionless) *backscattering coefficient* σ^0 , defined by

$$\sigma^0 = \gamma \cos\theta_0 = 4\pi R \cos^2\theta_0. \quad (3.39)$$

The BRDF is a function of the incidence and scattered directions (σ^0 is a function of the incidence direction only, since the scattered direction is the same), so in principle it should be written as a function of its arguments: $R(\theta_0, \phi_0, \theta_1, \phi_1)$. This notation is useful since it allows us to state the reciprocity relation obeyed by the BRDF:

$$R(\theta_0, \phi_0, \theta_1, \phi_1) = R(\theta_1, \phi_1, \theta_0, \phi_0), \quad (3.40)$$

but for compactness we just write R , and take the arguments to be implied. In the majority of cases the surface will lack azimuthally dependent features so that the dependence on ϕ_0 and ϕ_1 simplifies to a dependence on $(\phi_0 - \phi_1)$, and often the azimuthal dependence can be neglected altogether.

The *reflectivity* r of the surface is a function only of the incidence direction. It defines the ratio of the total power scattered to the total incident power. It is thus given by

$$r(\theta_0, \phi_0) = \frac{M}{E},$$

where M is the radiant exitance of the surface, and on substituting for M from Equation (2.25) we find that

$$r(\theta_0, \phi_0) = \int_{\theta=0}^{\pi/2} \int_{\phi=0}^{2\pi} R \cos\theta_1 d\theta_1 d\phi_1. \quad (3.41)$$

The reflectivity is also commonly called the *albedo* (from the Latin for ‘whiteness’) of the surface, and it is related to the emissivity ε in the direction (θ_0, ϕ_0) through

$$r = 1 - \varepsilon. \quad (3.42)$$

We can define the *diffuse albedo* r_d , also called the *hemispherical albedo*, as the average value of r over the hemisphere of possible incidence directions. In this case it represents the ratio of the total scattered power to the total incident power when the latter is distributed isotropically. Since the incident radiance is therefore constant we may write it as L_0 , so that the contribution dE to the irradiance from the direction (θ_0, ϕ_0) is $L_0 \cos\theta_0 \sin\theta_0 d\theta_0 d\phi_0$. The contribution dM that this makes to the radiant exitance in the direction (θ_1, ϕ_1) must therefore be given by $RL_0 \cos\theta_0 \sin\theta_0 d\theta_0 d\phi_0 \cos\theta_1 \sin\theta_1 d\theta_1 d\phi_1$. The radiant exitance is thus

$$M = L_0 \int_{\theta_0=0}^{\pi/2} \int_{\phi_0=0}^{2\pi} \int_{\theta_1=0}^{\pi/2} \int_{\phi_1=0}^{2\pi} R \cos\theta_0 \sin\theta_0 \cos\theta_1 \sin\theta_1 d\theta_0 d\phi_0 d\theta_1 d\phi_1$$

and the irradiance is

3.3 Scattering from rough surfaces

$$M = L_0 \int_{\theta_0=0}^{\pi/2} \int_{\phi_0=0}^{2\pi} \cos\theta_0 \sin\theta_0 d\theta_0 d\phi_0 = \pi L_0.$$

Since the diffuse albedo is given by M/E in this case, we can write it (making use of Equation 3.40 to simplify the formula a little) as

$$r_d = \frac{1}{\pi} \int_{\theta_0=0}^{\pi/2} \int_{\phi_0=0}^{2\pi} r(\theta_0, \phi_0) \cos\theta_0 \sin\theta_0 d\theta_0 d\phi_0. \quad (3.43)$$

The albedo, in either its directional or diffuse form, plays a significant role in our understanding of the Earth's climate system since it controls the Earth's energy budget by determining the proportion of incident solar radiation that is reflected back into space. Measurement of the albedo, usually integrated over wavelength, is thus of considerable importance. Some satellite-borne sensors are designed specifically to measure this quantity.

3.3.2 Simple models of surface scattering

In this section, we discuss a few of the important models of the BRDF of real surfaces. Further information can be found in, for example, Hapke (2005).

If the scattering surface is sufficiently smooth, it will behave like a mirror. This is called *specular scattering* or specular reflection (the word comes from the Latin *speculum*, meaning a mirror). Radiation incident from the direction (θ_0, ϕ_0) will be scattered only into the direction $\theta_1 = \theta_0, \phi_1 = \phi_0 - \pi$, as illustrated schematically in Figure 3.10a. The BRDF must therefore be a delta-function, and we can write it as

$$R = \frac{|r(\theta_0)|^2}{\cos\theta_0 \sin\theta_0} \delta(\theta_1 - \theta_0) \delta(\phi_1 - \phi_0 + \pi), \quad (3.44)$$

where $r(\theta_0)$ is the appropriate Fresnel amplitude reflection coefficient for radiation with incidence angle θ_0 . Inserting this expression into Equation (3.41), we find that the reflectivity r for radiation with incidence angle θ_0 is just

$$r = |r(\theta_0)|^2,$$

as of course it must be, and from Equation (3.43) the diffuse albedo is

$$r_d = 2 \int_0^{\pi/2} |r(\theta_0)|^2 \cos\theta_0 \sin\theta_0 d\theta_0.$$

Specular scattering is one limiting case of surface scattering, and arises when the surface is very smooth (later we shall consider just how smooth it needs to be). The other important limiting case is that of an ideally rough surface, giving *Lambertian scattering* (named after Johann Heinrich Lambert, Figure 3.11). This has the property that, for any illumination which is uniform across the surface, the scattered radiation is distributed isotropically, and so the BRDF has a constant value. This is illustrated schematically in Figure 3.10b. From Equations (3.41) and (3.43) it can easily be seen that, for such a surface,

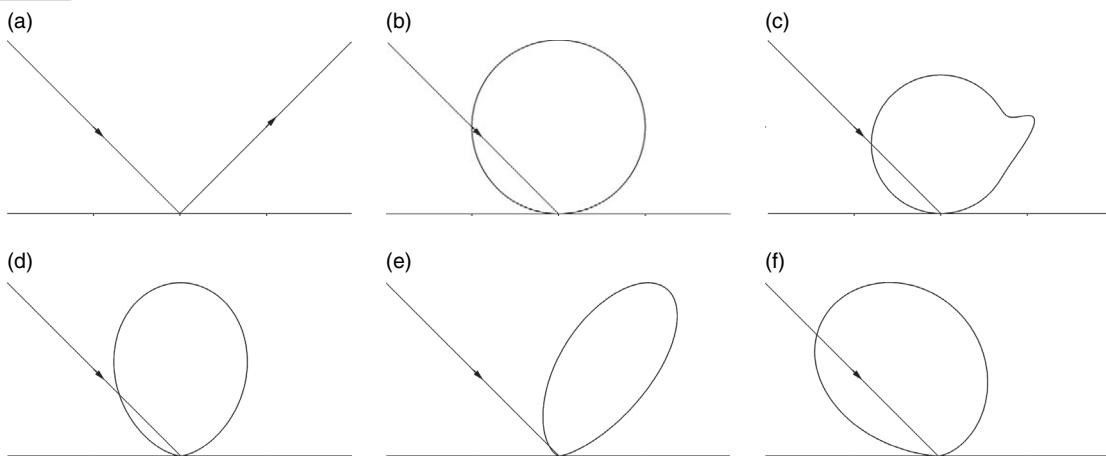


Figure 3.10. Schematic illustration of different types of surface scattering. The lobes are polar diagrams of the scattered radiation. The length of a line joining the point where the radiation is incident on the surface to the lobe is proportional to the radiance scattered in the direction of the line. (a) Specular reflection; (b) Lambertian scattering; (c) quasi-Lambertian scattering with a pseudospecular component; (d) Minnaert model ($\kappa = 2$); (e) Henyey–Greenstein model of forward scattering ($g = +0.5$); (f) Henyey–Greenstein model of backscattering ($g = -0.7$).



Figure 3.11. Johann Heinrich Lambert (1728–77) was a Swiss mathematician who carried out important work in optics in the mid eighteenth century. He also developed several of the map projections, including the Transverse Mercator projection, that are in widespread use today. (Source: Wikipedia. <http://en.wikipedia.org/wiki/File:JHLambert.jpg>).

3.3 Scattering from rough surfaces

$$R(\theta_0, \phi_0, \theta_1, \phi_1) = r(\theta_0, \phi_0) = \frac{r_d}{\pi}. \quad (3.45)$$

Thus, for example, a Lambertian surface that scatters all of the radiation incident upon it has $R = 1/\pi$.

Demonstrating Lambertian scattering

Plain white paper, for example the kind used in photocopiers, scatters light in an almost Lambertian manner. This can be demonstrated by illuminating a sheet of paper in a strongly non-uniform manner (e.g. by allowing direct sunlight to fall on it) and then exploring the angular distribution of the scattered light using a light meter. Cameras usually have built-in light meters so can be used to estimate the amount of light.

The scattering behaviour of real surfaces is often specified, not by using the BRDF, but instead by measuring the *bidirectional reflectance factor* (BRF). This is defined as the ratio of the flux scattered into a given direction by a surface under given conditions of illumination to the flux that would be scattered in the same direction by a perfect Lambertian scatterer under the same conditions. The usefulness of this approach is that surfaces can be manufactured to have a BRF very close to unity for a fairly wide range of wavelengths and of incidence and scattering angles. The most common materials are barium sulphate, which, as a pressed powder and for $\theta < 45^\circ$, has a BRF greater than 0.99 for wavelengths between 0.37 and 1.15 μm , and magnesium oxide, which has a BRF greater than 0.98 over roughly the same range of conditions. ‘Spectralon’, a fluoropolymer material, has now largely replaced these materials.

Although the Lambertian model is simple and idealised, the scattering from many natural surfaces can often, to a first approximation at least, be described using it. A simple modification is provided by the *Minnaert model*, in which the BRDF is given by

$$R \propto (\cos\theta_0 \cos\theta_1)^{\kappa-1}, \quad (3.46)$$

where the parameter κ has the effect of increasing or decreasing the radiance scattered in the direction of the surface normal (Figure 3.10d). (M. G. J. Minnaert was a Belgian astronomer who spent most of his working life in the Netherlands.) Lambertian scattering is the special case of the Minnaert model with $\kappa = 1$.

The scattering from real rough surfaces can often, as we have just remarked, be described by the Lambert or Minnaert models. However, neither of these models accounts for the fact that real surfaces may also show additional backscattering (where the radiation is scattered back into the incidence direction) or specular scattering. These can of course be incorporated by devising an empirical model that combines a Lambertian or Minnaert component with, for example, a ‘quasi-specular’ component. One common modification is to multiply the Lambert or Minnaert BRDF by the *Henyey–Greenstein* term (Henyey and Greenstein 1941)

$$\frac{1 - g^2}{(1 - 2g \cos\Theta + g^2)^{3/2}}, \quad (3.47)$$

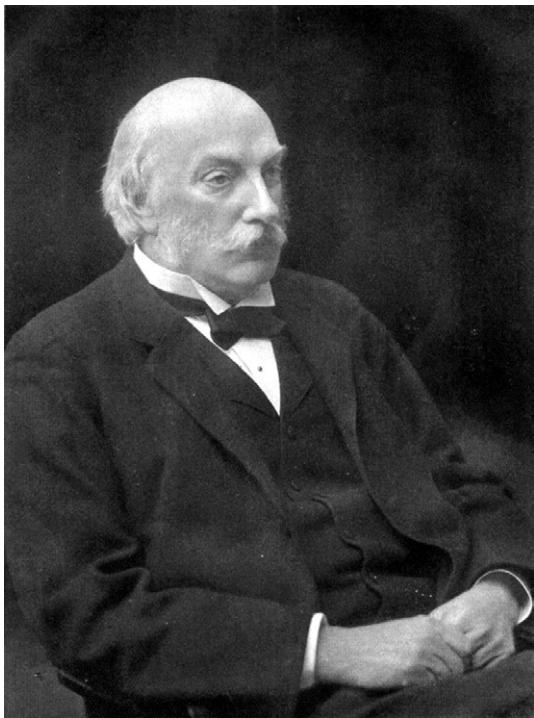


Figure 3.12. John William Strutt (1842–1919), third Lord Rayleigh, was a British physicist who worked extensively on wave theory. He was awarded the Nobel Prize in physics in 1904. His contributions to physics appear several times in this chapter. (Source: Royal Society, via Wikipedia. http://en.wikipedia.org/wiki/File:John_William_Strutt.jpg).

where the parameter g represents the anisotropy of the scattering, with $0 < g \leq 1$ corresponding to forward scattering and $-1 \leq g < 0$ corresponding to backscattering. Θ is the scattering phase angle, given by

$$\cos\Theta = \cos\theta_0 \cos\theta_1 + \sin\theta_0 \sin\theta_1 \cos(\phi_1 - \phi_0). \quad (3.48)$$

Figures 3.10e and f illustrate typical BRDFs that incorporates the Henyey–Greenstein term.

3.3.3

The Rayleigh roughness criterion

We have distinguished between the behaviour of a perfectly smooth surface and a Lambertian surface, which is in some sense perfectly rough. It is clear that in order to understand which of these simple models is likely to give a better model of scattering from a real surface, some measure of surface roughness must be developed. The usual approach is via the Rayleigh criterion, which we develop in this section. (This is named after the third Lord Rayleigh, Figure 3.12)

Figure 3.13 shows schematically the detailed behaviour when radiation is incident on an irregular surface at angle θ_0 , and scattered specularly from it at the same angle. We consider two rays: one is scattered from a reference plane, and the other from a plane at a

3.3 Scattering from rough surfaces

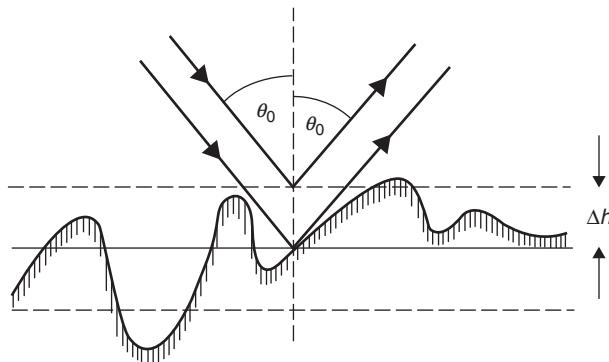


Figure 3.13. The Rayleigh criterion. Radiation is specularly reflected at an angle θ_0 from a surface whose r.m.s. height deviation is Δh . The difference in the lengths of the two rays is $2\Delta h \cos \theta_0$.

height Δh above this reference plane. After scattering, the path difference between these two rays is $2\Delta h \cos \theta_0$, so the phase difference between them is

$$\Delta\phi = \frac{4\pi\Delta h \cos\theta_0}{\lambda},$$

where λ is the wavelength of the radiation. If we now let Δh stand for the root mean square (r.m.s.) variation in the surface height, $\Delta\phi$ becomes the r.m.s. variation in the phase of the scattered rays. A surface can be defined as smooth enough for scattering to be specular if $\Delta\phi$ is less than some arbitrarily defined value of the order of 1 radian. The conventional value is $\pi/2$, and this is called the Rayleigh criterion. Thus, for a surface to be smooth according to this criterion,

$$\Delta h < \frac{\lambda}{8\cos\theta_0}. \quad (3.49)$$

Note that other criteria have also been adopted for the value of $\Delta\phi$ at which the surface becomes effectively smooth. A common definition that provides for the possibility of some intermediate cases between rough and smooth is that if $\Delta\phi$ is greater than $\pi/2$ the surface is rough, and if $\Delta\phi$ is less than $4\pi/25$ (so that the numerical part of the denominator in Equation (3.49) becomes 25 instead of 8) it is smooth.

Equation (3.49) evidently dictates that for a surface to be effectively smooth at normal incidence irregularities must be less than about $\lambda/8$ (or perhaps $\lambda/25$) in height. Thus, for a surface to give specular reflection at optical wavelengths (say $\lambda = 0.5 \mu\text{m}$), Δh must be less than about 60 nanometres. This is a condition of smoothness likely to be met only in certain man-made surfaces such as sheets of glass or metal. On the other hand, if the surface is to be examined using VHF (very high frequency) radio waves (say $\lambda = 3 \text{ m}$), Δh need only be less than about 40 cm, a condition that could be met by a number of naturally occurring surfaces. A further aspect of Equation (3.49) is the dependence on θ_0 . The smoothness criterion is more easily satisfied at large values of θ_0 than at normal incidence, so that a moderately rough surface may be effectively smooth at glancing incidence. This fact is well known to anyone who has endured the glare of reflected

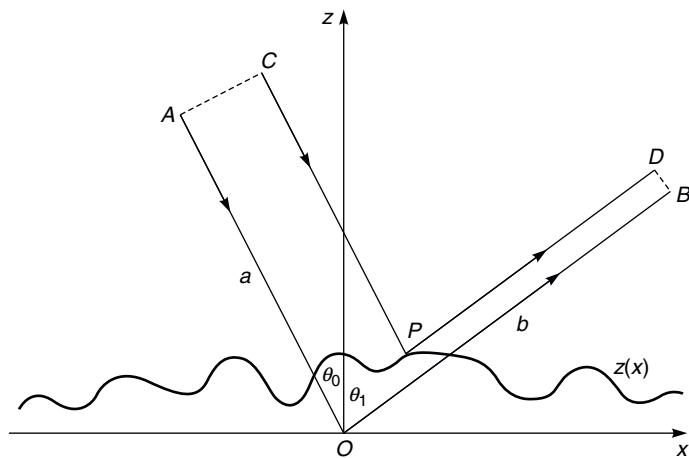


Figure 3.14. Model for developing the small perturbation model of rough surface scattering.

sunlight from a low sun over an ordinary road surface. Although the scattering cannot really be described as specular in this case, the component of the BRDF in the specular direction is greatly enhanced.

3.3.4

Models for microwave backscatter

In this section we shall discuss some of the commonest physical methods used to model the microwave backscatter from rough surfaces. This is a large and important area in which considerable research is still taking place, and the reader who wishes to pursue it in greater depth is recommended to study, for example, the books by Bass and Fuks (1979), Ulaby, Moore and Fung (1981, 1982), and Beckman and Spizzichino (1987), and the more recent research literature. The mathematical development of these models is generally rather advanced, and in this section we can do little more than sketch their principles.

3.3.4.1

The small perturbation method

The most helpful way to begin to look at the problem of microwave scattering from a rough surface is through the small perturbation method. This is essentially a Fraunhofer diffraction approach to rough surface scattering, in which the interaction of the incident radiation with the surface is used to calculate the outgoing radiation field in the vicinity of the surface. This field can then be regarded as having been produced from a uniform incident radiation field by a fictitious screen that changes both the amplitude and the phase, and the far-field radiation pattern is obtained by calculating the Fraunhofer diffraction pattern of this screen.

In order to see how this is applied, but without becoming too deeply immersed in mathematical detail, we consider a surface $z(x, y)$ in which the height z depends only on x , and which scatters all radiation incident upon it (i.e. the diffuse albedo is 1).

Figure 3.14 illustrates radiation incident on the surface at angle θ_0 and scattered from it at angle θ_1 . The rays AO and OB , of length a and b respectively, provide a reference from

3.3 Scattering from rough surfaces

which phase differences can be measured. P is a point on the surface, having coordinates $x, z(x)$. Simple trigonometry shows that the length of the ray CP is

$$a + x\sin\theta_0 - z(x)\cos\theta_0$$

and the length of the ray PD is

$$b - x\sin\theta_1 - z(x)\cos\theta_1,$$

so the phase of the wave at D relative to the reference ray can be written as

$$\phi(x) = k\alpha x - k\beta z(x),$$

where $\alpha = (\sin \theta_0 - \sin \theta_1)$, $\beta = (\cos \theta_0 + \cos \theta_1)$, and k is the wavenumber of the radiation. The total amplitude E scattered from the direction θ_0 into the direction θ_0 can therefore be written as

$$E = \int_{-\infty}^{\infty} \exp(-i\phi(x))dx = \int_{-\infty}^{\infty} \exp(-ik\alpha x)\exp(ik\beta z(x))dx.$$

This expression is clearly the Fourier transform of the function $\exp(ik\beta z(x))$, which we can expand as a power series:

$$\exp(ik\beta z(x)) = 1 + ik\beta z(x) - \frac{(k\beta z(x))^2}{2} \dots, \quad (3.50)$$

so that our expression for the scattered field amplitude becomes

$$E = \int_{-\infty}^{\infty} \left[1 + ik\beta z(x) - \frac{(k\beta z(x))^2}{2} \dots \right] \exp(-ik\alpha x)dx. \quad (3.51)$$

The first term in this expression is a delta-function centred at $\alpha=0$. Using the fact that $\alpha = (\sin \theta_0 - \sin \theta_1)$, we can see that this is just the specularly scattered component $\theta_1 = \theta_0$. We can write the second term as

$$ik\beta \int_{-\infty}^{\infty} z(x)\exp(-ik\alpha x)dx,$$

which is proportional to the Fourier transform of the surface height function $z(x)$. This suggests that it will be helpful to write the height function in terms of its Fourier transform $a(q)$, where q is the spatial frequency:

$$z(x) = \int_{-\infty}^{\infty} a(q)\exp(-iqx)dq.$$

Thus, the second term in Equation (3.49) becomes

$$ik\beta \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(q)\exp(i(q - k\alpha))dq dx.$$

Using the definition of the Dirac delta-function that we met in Section 2.3, we see that this can be written as

$$2\pi ik\beta \int_{-\infty}^{\infty} a(q)\delta(q - k\alpha)dq. \quad (3.52)$$

This is the result that we need. It shows that the amplitude of the radiation scattered into the direction specified by α is proportional to the component of the surface height function with spatial frequency $q = k\alpha$.

There is another way of thinking about this result. The wave vector of the incident radiation has a horizontal component $k \sin \theta_0$ and the wave vector of the scattered radiation has a horizontal component $k \sin \theta_1$, so the term $k\alpha$ is just the change in this horizontal component. We can therefore say that the component of the scattered radiation amplitude is proportional to the spatial frequency component of the surface profile corresponding to the change in the horizontal component of the radiation's wave vector.

Up to this point, we have assumed that the third and subsequent terms in the power-series expansion of Equation (3.50) are negligible in comparison with the first two. If this is true, the phenomenon is known as *Bragg scattering*. Clearly, the condition that needs to be satisfied is

$$k\beta\Delta h \ll 1,$$

which is equivalent to

$$\Delta h \ll \frac{\lambda}{2\pi(\cos\theta_0 + \cos\theta_1)}.$$

Thus the surface must be smooth according to the Rayleigh roughness criterion (3.49). The Bragg scattering mechanism is thought to be largely responsible for the reflection of microwave radiation from small-scale (of the order of a centimetre) roughness on water surfaces, especially where the structure of this roughness contains a dominant spatial frequency, in which case the Bragg scattering is said to be *resonant*.

We have also assumed that the surface height z depends only on the x -coordinate. A more general derivation for two-dimensional isotropic surfaces, which more or less follows the argument we have just presented, leads to the following expression for the backscattering coefficient σ^0 :

$$\sigma_{pp}^0 = 4k^4 L^2 (\Delta h)^2 \cos^4 \theta |f_{pp}(\theta)|^2 \exp(-k^2 L^2 \sin^2 \theta). \quad (3.53)$$

In this expression, σ_{pp}^0 is the backscattering coefficient for pp -polarisation (so that, for example, $p = H$ means HH-polarisation, or radiation both incident and scattered in the horizontal polarisation state), θ is the incidence angle of the radiation, L is the correlation length of the surface (i.e. the ‘width’ of the irregularities), which contains information about the shape of the spatial frequency spectrum of the surface. In fact, Equation (3.53) is based on the assumption that the surface has a Gaussian autocorrelation function. L is the distance over which the autocorrelation coefficient falls to a value of $1/e$ (see box for more information), and $f_{pp}(\theta)$ is a measure of the surface reflectivity for radiation with incidence angle θ . For HH-polarised radiation we have

3.3 Scattering from rough surfaces

$$f_{HH}(\theta) = \frac{\cos\theta - \sqrt{\varepsilon_r - \sin^2\theta}}{\cos\theta + \sqrt{\varepsilon_r - \sin^2\theta}}, \quad (3.54.1)$$

which is just the Fresnel reflection coefficient for radiation incident at angle θ from a vacuum onto the surface of a medium with a (complex) dielectric constant ε_r . For VV-polarised radiation, the corresponding formula is

$$f_{VV}(\theta) = (\varepsilon_r - 1) \frac{\sin^2\theta - \varepsilon_r(1 + \sin^2\theta)}{(\varepsilon_r \cos\theta + \sqrt{\varepsilon_r - \sin^2\theta})^2}. \quad (3.54.2)$$

The conditions for the validity of Equation (3.53) are usually given as

$$k\Delta h < 0.3 \quad (3.55.1)$$

and

$$kL < 3, \quad (3.55.2)$$

although these are somewhat approximate.

The autocorrelation function

The r.m.s. height variation Δh is the simplest measure of the roughness of a surface, but it tells us nothing about the *scale* of the irregularities. The autocorrelation function is one way of providing this information.

Suppose, for simplicity, we consider a one-dimensional surface $z(x)$, where z is the height at position x . The mean height is $\langle z \rangle$, where the angle brackets denote averaging over all values of x , and the r.m.s. height deviation is defined by

$$\Delta h = \langle (z(x) - \langle z \rangle)^2 \rangle^{1/2}.$$

The autocorrelation function is defined by

$$\rho(\xi) = \frac{\left\langle (z(x + \xi) - \langle z \rangle)(z(x) - \langle z \rangle) \right\rangle}{(\Delta h)^2}$$

and is a measure of the similarity of the heights at two points separated by distance ξ . By definition, $r(0) = 1$, and for most surfaces, $r(\infty) = 0$.

Common models for the autocorrelation function are the Gaussian

$$\rho(\xi) = \exp\left(-\frac{\xi^2}{L^2}\right)$$

and the negative exponential

$$\rho(\xi) = \exp\left(-\frac{|\xi|}{L^2}\right).$$

In each case, L (the correlation length) is a measure of the width of the irregularities of the surface.

The extension to two dimensions is straightforward.

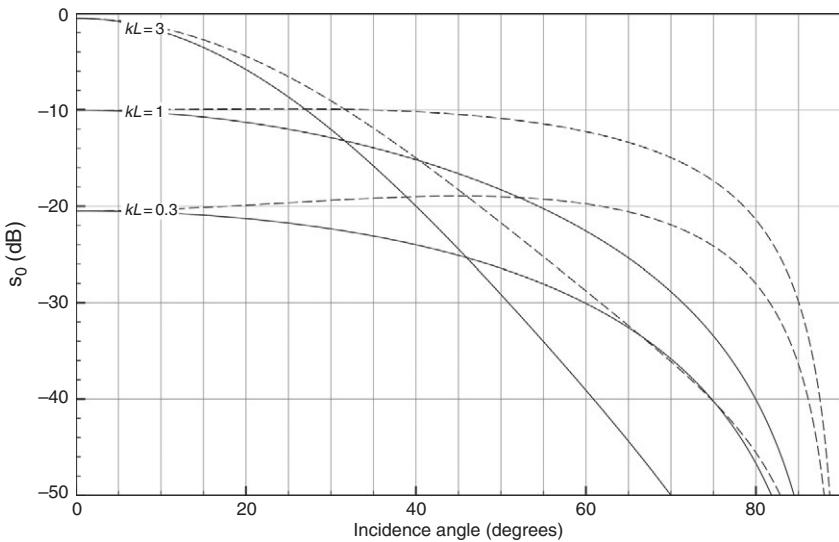


Figure 3.15. Backscatter calculated according to the small perturbation model for a surface having dielectric constant $\varepsilon_r = 10 - 2i$. In each case $k \Delta h = 0.3$, and the curves are labelled with the values of kL . The solid curves are for HH-polarisation and the dashed curves are for VV-polarisation.

Figure 3.15 illustrates the backscatter predicted by the small perturbation model for a surface with $\varepsilon_r = 10 - 2i$. In each case, the r.m.s. height variation Δh is the same. It can be seen that the effect of increasing kL , the spatial scale of the surface roughness features, is to increase the specular scattering ($\theta = 0$) at the expense of the scattering at larger angles. Since the r.m.s. surface slope is of the order of $\Delta h/L$, increasing the value of kL while keeping the value of $k \Delta h$ constant corresponds to decreasing the r.m.s. surface slope, so it is not surprising that this increases the specular component of the scattering. The effect of varying $k \Delta h$, not shown in the figure, is very straightforward since Equation (3.53) shows that the backscatter coefficient is just proportional to $(\Delta h)^2$. Thus decreasing $k \Delta h$ by a factor of 10, say, would have the effect of shifting all the curves in the figure down by 20 dB without changing their shapes.

3.3.4.2 The Kirchhoff model

The approach that is adopted by the Kirchhoff model (named after Gustav Kirchhoff, Figure 3.16) for scattering for randomly rough surfaces is to model the surface as a collection of variously oriented planes, each of which is locally tangent to the surface. This is called the *tangent plane approximation*. The scattered radiation field can then be calculated using the results for radiation incident on a plane interface. Two variants of the Kirchhoff model are in common use: the *stationary phase* (or *geometrical optics*) model, which is valid for rougher surfaces, and the *scalar approximation*.

The backscatter coefficient is given by the stationary phase model as

$$\sigma_{HH}^0(\theta) = \sigma_{VV}^0(\theta) = \frac{|r(0)|^2 \exp(-(\tan^2 \theta)/2m^2)}{2m^2 \cos^4 \theta} \quad (3.56)$$

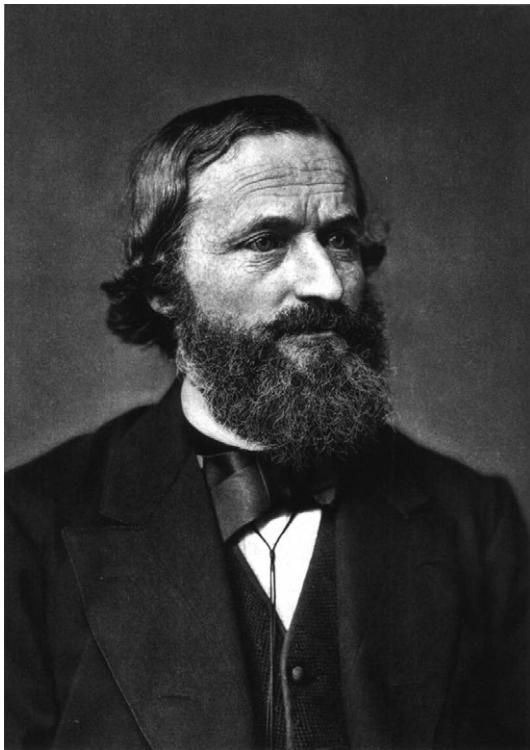


Figure 3.16. Gustav Kirchhoff (1824–87) was a German physicist who worked particularly in the fields of spectroscopy and thermal radiation. He conceived the term ‘black-body radiation’. With Robert Bunsen, of burner fame, he discovered the elements caesium and rubidium. (Source: Wikipedia. http://en.wikipedia.org/wiki/File:Gustav_Robert_Kirchhoff.jpg)

where $r(0)$ is the Fresnel amplitude reflection coefficient for normally incident radiation, and m is the r.m.s. surface slope. For a surface having a Gaussian autocorrelation function with correlation length L and r.m.s. height variation Δh ,

$$m = \sqrt{2} \frac{\Delta h}{L}. \quad (3.57)$$

The conditions for the validity of this model are

$$k\Delta h \cos\theta > 1.58, \quad (3.58.1)$$

$$kL > 6 \quad (3.58.2)$$

and

$$kL^2 > 17.3\Delta h \quad (3.58.3)$$

The backscatter coefficient is given by the scalar approximation as

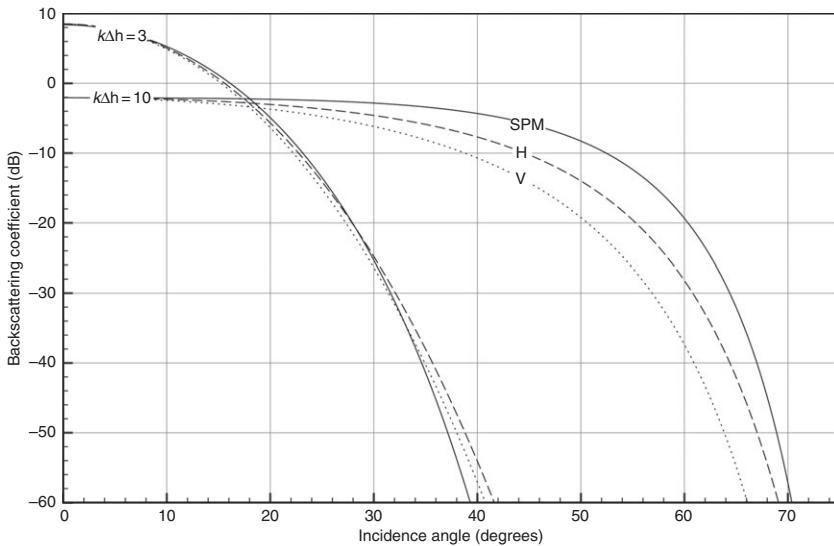


Figure 3.17. Backscatter calculated according to the Kirchhoff model for a surface having dielectric constant $\epsilon_r = 10 - 2i$. In each case $kL = 30$, and the curves are labelled with the values of $k \Delta h$. Solid lines: stationary phase model (SPM); dashed lines: scalar approximation, HH-polarisation; dotted lines: scalar approximation, VV-polarisation.

$$\sigma_{pp}^0(\theta) = k^2 L^2 \cos^2 \theta |r_p(\theta)|^2 \exp(-4k^2 \Delta h^2 \cos^2 \theta) \times \sum_{n=1}^{\infty} \frac{(2k\Delta h)^{2n}}{n! n} \exp\left(-(k^2 L^2 \sin^2 \theta)/n\right), \quad (3.59)$$

where we have again assumed that the autocorrelation function of the surface is Gaussian. $r_p(\theta)$ is the Fresnel coefficient for p -polarised radiation incident at angle θ . The conditions for the validity of this model are

$$\Delta h < 0.18L, \quad (3.60.1)$$

$$kL > 6 \quad (3.60.2)$$

and

$$kL^2 > 17.3\Delta h. \quad (3.60.3)$$

We can note that the second and third of these conditions are the same as for the stationary phase model.

Figure 3.17 illustrates the backscatter predicted by both variants of the Kirchhoff model for a surface with $\epsilon_r = 10 - 2i$. In each case, the value of kL has been kept constant. Two sets of curves show the predicted backscatter from a smooth ($k \Delta h = 3$) and rough ($k \Delta h = 10$) surface. Again, we see that smoother surfaces give enhanced scattering near the specular direction.

There are in fact further restrictions on the validity of the Kirchhoff model that we should note. As was mentioned above, the first step in constructing the model is to replace

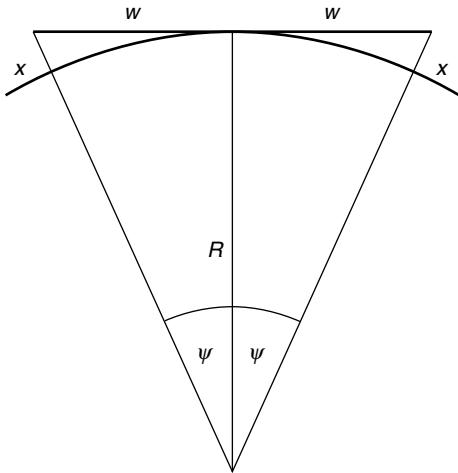


Figure 3.18. A one-dimensional facet of length $2w$ is tangent to a surface with radius of curvature R . The facet subtends an angle 2ψ at the centre of curvature, and its maximum deviation from the surface is x .

the surface by a set of planes or facets, each of which is locally tangent to the surface. It is clear that, for this programme to succeed, we must be able to define facets whose spatial extent is much greater than the wavelength λ (so that diffraction effects do not dominate), but whose deviation from the real surface is much less than λ (so that we do not incur large phase errors in modelling the surface). This is in fact a restriction on the local curvature of the surface, as we can see by the following simple one-dimensional argument.

We assume that, locally, the surface has a constant radius of curvature R , so that it forms part of a sphere or, in one dimension, a circle. Figure 3.18 shows a facet of length $2w$ tangent to this circle.

The facet subtends an angle 2ψ at the centre of curvature, where $\psi = \tan^{-1}(w/R)$, and the maximum deviation of the facet, x , from the surface is $R(\sec \psi - 1)$. If $w = R$, this can be approximated as $x \approx w^2/2R$. Now we assume that $w > \lambda$ (so that the facet is large enough for diffraction effects not to dominate) and that $x < \lambda/2$ (so that we do not incur large phase errors in approximating the surface by the facet). Thus we obtain the condition that $R > \lambda$. In other words, in order for the surface to be adequately represented by facets, its radius of curvature must exceed a few wavelengths.

Another condition that must be satisfied is that the incidence or scattering angles should not be so large that one part of the surface can obscure another. If this occurs in practice, it can usually be dealt with by an appropriate modification of the model, or by specifying that the model is valid only up to some maximum angle.

3.3.4.3 Other models

It is clear that the models of backscatter from randomly rough surfaces that we have considered in the two preceding sections do not provide a complete description of the possible phenomena. First, we can note that the ranges of validity of the three models that we have discussed, defined in Equations (3.55), (3.58) and (3.60), do not cover all the

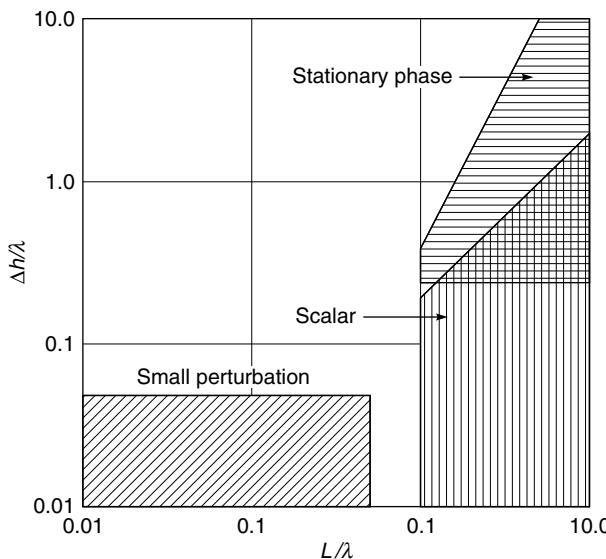


Figure 3.19. Range of validity of the small perturbation, stationary phase and scalar approximation models of rough surface scattering. For the stationary phase model, the incidence angle has been taken as zero.

possibilities. This fact is illustrated in Figure 3.19, which shows the valid range of each model in terms of the dimensionless parameters ($\Delta h/\lambda$) and (L/λ).

Second, we can note from Figure 3.17 that, even where two models are both apparently valid, they can give different predictions for the backscatter. This is to some extent a consequence of the approximations inherent in the models. Third, we can observe that none of the three models that have been discussed has an explicit dependence on the imaginary part of the dielectric constant, which is in contradiction to the experimental evidence.

Many of these difficulties can be circumvented by the use of the *integral equation model*. This is still an approximation, but its grounding in the physics of the interactions between the radiation and the surface is more nearly fundamental. The integral equation model has a larger region of validity and is sensitive to both the real and the imaginary parts of the dielectric constant. It is thus useful for estimating these parameters from backscattering measurements. Unfortunately, its mathematical complexity is such that it is beyond our scope to discuss it any further here.

Summary

When electromagnetic radiation is incident on a rough surface, its angular distribution is in general changed. The most detailed information about how the radiation is scattered is provided by the bidirectional reflectance distribution function (BRDF). The reflectivity r is an average of the BRDF over all outgoing directions, and is therefore a function of the incidence direction. It is related to the emissivity ε of the surface through $\varepsilon = 1 - r$. An even more generalised measure of the scattering is the diffuse albedo, which is an average

3.4 Absorption and scattering by particles

of the reflectivity over all incidence directions. If the incident radiation is isotropically distributed, the diffuse albedo is the ratio of the scattered power to the incident power. The effect of a real surface on electromagnetic radiation is governed primarily by the roughness of the surface. The Rayleigh roughness criterion relates the roughness (variation in surface height) to the wavelength of the radiation and also the direction of the incident radiation. Very smooth surfaces show specular (mirror-like) reflection, while very rough surfaces show Lambertian scattering, in which the scattered radiance is the same in all directions. Although this is often a reasonable approximation to the behaviour of natural surfaces illuminated by visible light, several models exist to deal with intermediate cases. A separate class of models has been developed to describe backscattering of microwave radiation from rough surfaces. Apart from the dielectric constant of the surface material, these are mainly characterised by the surface roughness and the horizontal scale of the roughness features.

3.4

Absorption and scattering by particles

In Section 3.1 we met the idea of electromagnetic radiation being absorbed by matter, which we quantified using the complex refractive index or the absorption length. These concepts are appropriate if we think of matter as a homogeneous continuum, but at the microscopic level it is clear that the radiation must be absorbed by individual particles – atoms, molecules or larger particles. In this section we develop some of the theory needed to understand the interactions between electromagnetic radiation and small particles. The classic work on this topic is van de Hulst's book originally published in 1957 (van de Hulst 1981), and a more modern work is Bohren and Huffman (1983).

The effectiveness of a single particle at absorbing radiation can be quantified through its *absorption cross-section* σ_a . This is defined as the ratio of the power P_a absorbed by the particle to the flux density F incident on it:

$$P_a = \sigma_a F. \quad (3.61)$$

Since we saw in Section 3.1 that the flux density is the power per unit area carried by the radiation, it is clear that the absorption cross-section has the dimensions of an area and it can be thought of as an ‘effective area’ over which the particle collects and absorbs radiation. If we are considering a macroscopic particle (one larger than an atom or a small molecule), we can also define its geometrical cross-sectional area, and hence a dimensionless ratio Q_a called the *absorption efficiency*. For example, if the particle is a sphere of radius r , we define Q_a through

$$\sigma_a = \pi r^2 Q_a. \quad (3.62)$$

The case of atoms and molecules is discussed separately in Section 3.4.1.

In addition to absorbing energy from the incident electromagnetic radiation, a particle can also *scatter* it. Scattering is changing the direction of the radiation without losing any of the energy. However, if we consider the situation in which the particle is illuminated by a collimated beam of radiation of flux density F travelling in a particular direction, scattering will reduce the power travelling in that direction. It is thus meaningful to think

of the power P_s scattered away from the original direction, and we can define a *scattering cross-section* σ_s similarly to (3.59):

$$P_s = \sigma_s F. \quad (3.63)$$

For macroscopic particles with a well-defined geometrical area we can therefore also define a scattering efficiency similarly to (3.62). For a spherical particle of radius r we have $\sigma_s = \pi r^2 Q_s$. (3.64)

The combined effect of absorption and scattering is referred to as *attenuation* or *extinction*, and a cross-section and an efficiency can be defined for this term too (they are simply the sums of the corresponding terms for absorption and scattering).

The values of Q_a and Q_s are determined by the size of the particle and its dielectric constant. They cannot be very much larger than 1 but they can be much smaller. The effect of size is relative to the wavelength of the radiation, and it is helpful to introduce another dimensionless parameter to describe this effect. For a spherical particle of radius r , the dimensionless size parameter is given by

$$\omega = ck \left(1 - \frac{Ne^2}{2\epsilon_0 m_e \omega^2} \right)^{-1}, \quad (3.65)$$

where λ is the wavelength of the radiation and $k = 2\pi/\lambda$ is the wavenumber. We distinguish three cases:

3.4.1

Very small particles ($\chi \ll 1$)

If the particle is very small compared with the wavelength, the radiation is in phase across the whole particle. This produces a situation that is generally called *Rayleigh scattering*. A useful concept in understanding absorption and scattering in this case is the polarisability, which we have already met in Section 3.1.1 and which will again give the symbol α . This is defined such that the dipole moment p induced in the particle when it is placed in an electric field E is given by αE . The direction of the dipole moment is the same as that of the electric field vector, but in an oscillating electric field, such as we have in the case of electromagnetic radiation, the dipole moment might not be in phase with the electric field. For this reason, the polarisability α will in general be complex.

First we consider the absorption of electromagnetic energy by the particle. If the induced dipole moment p is varying with time, the instantaneous rate of power dissipation is given by $E dp/dt$. Putting $E = E_0 \exp(i\omega t)$ for the electric field in the vicinity of the particle, we see that the dipole moment must be given by $p = \alpha E_0 \exp(i\omega t)$. However, we recall that the physical meaning of these complex exponential expressions is contained in their real parts. We can assume that E_0 is real without any loss of generality, so we can rewrite our expression for E as $E = E_0 \cos(\omega t)$ and obtain the following expression for dp/dt :

$$\frac{dp}{dt} = -\Re(\alpha)\omega E_0 \sin(\omega t) - \Im(\alpha)\omega E_0 \cos(\omega t).$$

The time-average of the dissipated power can then be obtained by multiplying these two expressions together, and averaging the result over time. This gives the mean dissipated power as

3.4 Absorption and scattering by particles

$$\langle P \rangle = \frac{-\omega \Im(\alpha)}{2} E_0^2.$$

Using Equations (2.4), (2.8), (2.9) and (3.61), we find that the absorption cross-section is given by

$$\sigma_a = \frac{-k}{\epsilon_0} \Im(\alpha), \quad (3.66)$$

where k is the wavenumber of the radiation. We can note from Equation (3.66) that the imaginary part of the polarisability cannot be positive.

For a small spherical particle of radius r and refractive index n (which may be complex), the polarisability is given by

$$\alpha = 4\pi\epsilon_0 r^3 \left(\frac{n^2 - 1}{n^2 + 2} \right), \quad (3.67)$$

so we obtain

$$Q_a = -4\chi \Im \left(\frac{n^2 - 1}{n^2 + 2} \right)$$

and hence, writing $n = m - ix$,

$$Q_a = \frac{24m\kappa}{(m^2 - \kappa^2 + 2)^2 + 4m^2\kappa^2} \chi \quad (3.68)$$

Because of our assumption that the radius r of the particle is much smaller than the wavelength λ , this expression will almost always be much smaller than 1. Thus, in the small-particle limit, the particle will absorb much *less* power than the power carried by a cross-section of the radiation equal to the geometrical area of the particle.

Next, we consider scattering by our small ($= \lambda$) particle. Classical electrodynamics shows that an oscillating electric dipole whose strength is described by $p = p_0 \sin(\omega t)$ radiates a mean power of

$$\langle P \rangle = \frac{\mu_0 \omega^4 p_0^2}{12\pi c},$$

so the time-average of the power reradiated by a particle of polarisability α in an electromagnetic wave whose electric field vector is given by $E = E_0 \exp(i\omega t)$ must be

$$\langle P \rangle = \frac{\mu_0 \omega^4 E_0^2}{12\pi c} |\alpha|^2$$

and so

$$\sigma_s = \frac{k^4}{6\pi\epsilon_0^2} |\alpha|^2. \quad (3.69)$$

From (3.65), this gives

$$Q_s = \frac{8(m^2 - 2m + \kappa^2 + 1)(m^2 + 2m + \kappa^2 + 1)}{3(m^2 - \kappa^2 + 2)^2 + 4m^2\kappa^2} \chi^4. \quad (3.70)$$

Equation (3.68) shows that if $\kappa = 0$, the absorption efficiency is zero as expected, and from (3.70) the scattering efficiency is

$$\frac{8}{3} \left(\frac{m^2 - 1}{m^2 + 2} \right)^2 \chi^4.$$

Thus, for a non-absorbing spherical particle of radius r , the scattering cross-section is

$$\sigma_s = \frac{128\pi^5}{3} \left(\frac{m^2 - 1}{m^2 + 2} \right)^2 \frac{r^6}{\chi^4}. \quad (3.71)$$

We can note a number of important features in this result. The scattering efficiency is much less than 1, and the amount of scattering is very strongly dependent on the wavelength. This characteristic is in general referred to as *selective scattering*; in the case of Rayleigh scattering we see that the scattering is inversely proportional to the fourth power of the wavelength. Equation (3.71) also emphasises the importance of dielectric contrast, which we noted in Section 3.2: if $m = 1$ there is no dielectric contrast between the particle and its surroundings and there is no scattering from it.

Although the scattering cross-section tells us how much of the incident radiation is scattered by a particle, it does not tell us how this scattered power is distributed over different directions. This is specified using the *phase function* $p(\cos \Theta)$ of the scattering, which describes the angular distribution of the scattered radiation in terms of the angle Θ through which the radiation has been deflected. For Rayleigh scattering, the phase function is given by

$$p(\cos \Theta) = \frac{3}{4}(1 + \cos^2 \Theta). \quad (3.72)$$

Equation (3.72) shows that the scattering is maximum in the forward ($\Theta = 0$) and backward ($\Theta = \pi$) directions, and minimum for scattering through $\pi/2$. The factor of $3/4$ ensures that the expression is correctly normalised, so that the integral over all directions is equal to 4π . Equation (3.72) is illustrated in Figure 3.20. For comparison, the phase function for an isotropic scatterer would be 1 in all directions. The phase function will be important when we consider the radiative transfer equation in Section 3.5.

3.4.2 Larger particles

When particles do not satisfy the criterion that the size parameter χ is much less than 1, the calculations of scattering and absorption become much more complicated. The term *Mie scattering* (after Gustav Mie, Figure 3.21) is used to describe the situation for larger values of χ , although when $\chi \gg 1$ the term *geometrical scattering* is also used.

The extinction (attenuation) and scattering efficiencies can be expressed as

$$Q_e = \frac{2}{\chi^2} \sum_{l=1}^{\infty} (2l+1)(|a_1^2| + |b_1^2|), \quad (3.73)$$

$$Q_s = \frac{2}{\chi^2} \sum_{l=1}^{\infty} (2l+1)\Re(a_l + b_l), \quad (3.74)$$

3.4 Absorption and scattering by particles

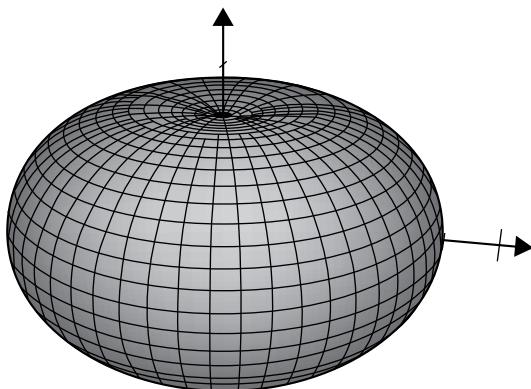


Figure 3.20. Polar diagram representing the phase function of Rayleigh scattering. Radiation is incident, in the direction of the horizontal axis, on a scatterer located at the origin of the coordinate system. The distance to the surface in any particular direction is proportional to the radiance scattered in that direction.

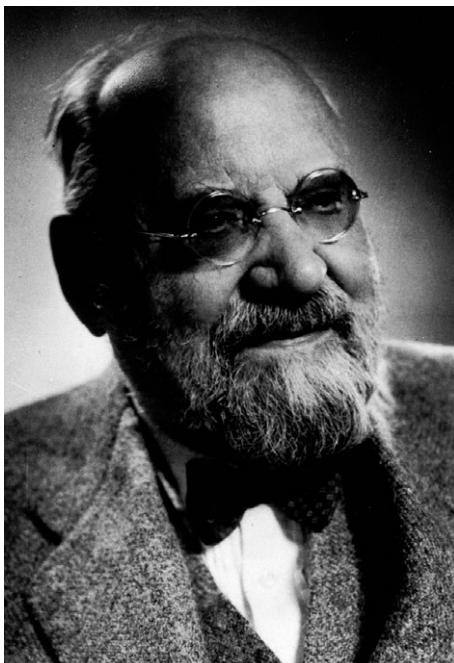


Figure 3.21. Gustav Mie (1869–1957) was a German physicist. His greatest contribution was to the understanding of scattering of electromagnetic radiation by small dielectric spheres. He was a friend of Wien's and they used to go skiing together. (Source: ThomasHB4 at Wikipedia. <http://en.wikipedia.org/wiki/File:GustavMie.gif>)

where a_l and b_l , which are the *Mie coefficients*, depend on the complex refractive index and on the value of χ . Equations (3.73) and (3.74) can be written as power series in χ , in which case the first few terms are

$$Q_e = -\Im \left(4\chi \frac{n^2 - 1}{n^2 + 2} + 4 \frac{\chi^3}{15} \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \frac{n^4 + 27n^2 + 38}{2n^2 + 3} \right) + \chi^4 \Re \left(\frac{8}{3} \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \right) + \dots \quad (3.75)$$

and

$$Q_s = \frac{8\chi^4}{3} \left| \frac{n^2 - 1}{n^2 + 2} \right|^2 + \frac{16\chi^6}{45} \left| \frac{(n^2 - 1)^2(n^2 - 2)}{(n^2 + 2)^3} \right| + \frac{32\chi^7}{27} \left| \frac{n^2 - 1}{n^2 + 2} \right|^3 + \dots \quad (3.76)$$

We can recognise the first term in Equation (3.75) as the result we have already derived for the absorption cross-section of a very small sphere, and the first term in Equation (3.76) as the Rayleigh scattering cross-section. When χ is large it is necessary to compute the Mie coefficients. A particularly effective technique was described by Bohren and Huffman (1983) and Matlab codes that use the Bohren and Huffman method have been developed by Mätzler (2002).

As an example, Figure 3.22 shows the dependence of Q_s and Q_a on χ for a refractive index $n = 5.67 - 2.88i$. For small values of χ (the Rayleigh scattering region), the scattering cross-section is proportional to χ^4 , and the absorption cross-section is proportional to χ . The figure also shows that, when χ increases beyond about 1, the cross-sections pass through a maximum value, which is of the order of the geometrical cross-section, and then tend towards a constant value, usually with some oscillations. This illustrates the fact that, for very large particles, the scattering (and the absorption) are independent of the wavelength of the radiation (unless, of course, the refractive index has a significant wavelength dependence). This phenomenon is termed *non-selective scattering*, in distinction to the selective scattering that we noted earlier.

The angular distribution of scattered radiation becomes more complicated as the size parameter χ increases. The number of wiggles in a graph of the scattering efficiency Q_s plotted against scattering angle Θ is of the order of χ . For example, if we consider visible light scattered from a spherical water drop of radius 0.1 mm, the size parameter χ is around 10^3 and a graph of the phase function has roughly a thousand wiggles in it. Figure 3.23 shows the phase function for this situation. It shows a very detailed variation with angle, not all of which can be resolved in the figure. It also shows narrow features where a lot of radiation is scattered in particular directions. This is something we expect, since we are familiar with the complicated angular distribution of light that results when it is incident on a raindrop and produces a rainbow (Figure 3.24).

3.4.3

Absorption and scattering by atoms and molecules

Atoms and molecules can scatter and absorb radiation. Absorption is primarily a quantum-mechanical phenomenon, and it is beyond the scope of this book to cover this in great detail. We may state, however, that the energy of an individual atom or molecule cannot be varied continuously, but must be one of a number, in principle infinite, of discrete values called *energy levels*. If a molecule absorbs electromagnetic radiation, it must be

3.4 Absorption and scattering by particles

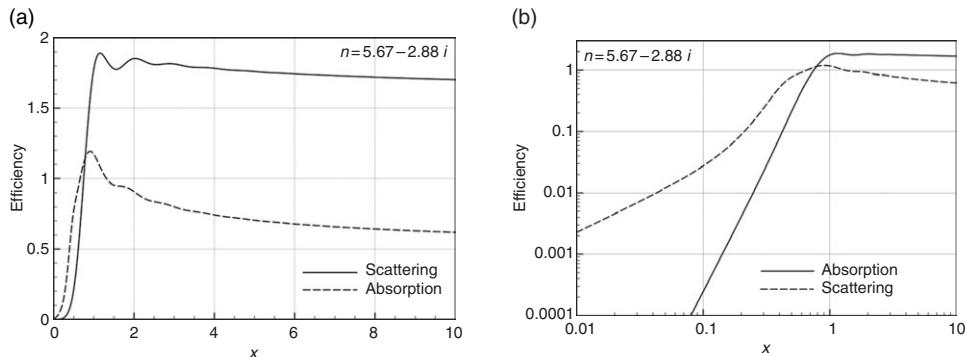


Figure 3.22. Dependence of the attenuation (extinction) and scattering cross-sections on the parameter χ for a spherical particle of refractive index $n = 5.67 - 2.88i$. (a) Linear scales; (b) logarithmic scales.

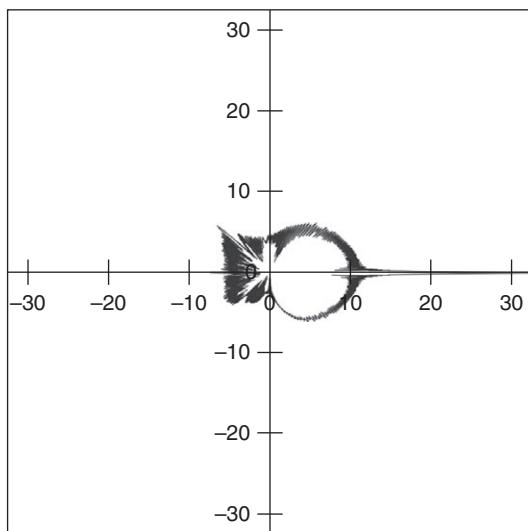


Figure 3.23. Phase function for light scattered from a spherical water droplet with a radius of 0.1 mm. The upper and lower halves of the diagram represent different polarisation states. The phase function is plotted on a logarithmic scale, and is somewhat schematic since not all details can be resolved. Note the strong forward scattering and the peak at about 140° corresponding to the primary rainbow.

promoted from one energy level to another, and hence only certain values of the energy increase ΔE are allowed. Planck's law states that the frequency f of the electromagnetic radiation is given by

$$\Delta E = hf, \quad (3.77.1)$$



Figure 3.24. The angular distribution of light in a rainbow (here a double rainbow can be seen) indicates the complexity of scattering when the size parameter is large. See also colour plates section.

where h is the Planck constant (this is in fact the same result that was stated in Equation 2.7), although it is often more convenient to write this equation in terms of the angular frequency ω :

$$\Delta E = \frac{h\omega}{2\pi} = h\omega. \quad (3.77.2)$$

Thus we expect that molecules will absorb ‘selectively’, at particular frequencies, which are usually called *absorption lines*.

There are three main mechanisms by which molecules can absorb electromagnetic radiation. The first of these, requiring the largest amounts of energy, involves the promotion of electrons to higher energy levels. These are termed *electronic transitions*. Calculation of the energy levels for any but the simplest of molecules is an extremely difficult task, so we illustrate the idea with reference to the hydrogen *atom*. In this case, the electronic energy levels are given by

$$E_n = -\frac{me^4}{32\pi^2\varepsilon_0^2\hbar^2n^2}, \quad (3.78)$$

where m is the electron mass (strictly, the electron’s reduced mass, which, in the hydrogen atom, is 99.95% of the electron mass) and n is a *quantum number* that can take only positive integer values. Substituting the values of the constants into the formula, we find that

$$E_n = -\frac{2.177 \times 10^{-18}}{n^2} \text{ J},$$

although it is often more convenient to use the electronvolt, so that the formula becomes

$$E_n = -\frac{13.59}{n^2} \text{ eV}.$$

In its *ground state* (the configuration with lowest energy), hydrogen has $n = 1$. The smallest increase in energy therefore corresponds to the transition from $n = 1$ to 2, which

3.4 Absorption and scattering by particles

requires an increase in energy of 10.2 eV. This is typical of the energies required for electronic transitions, and from Equation (3.77) we see that the frequency of the electromagnetic radiation needed to cause such transitions will therefore be of the order of 10^{15} Hz. The corresponding wavelength is thus a few tenths of a micrometre, so that we expect to find the absorption lines due to electronic transitions in the ultraviolet and visible regions of the electromagnetic spectrum.

Frequencies and wavenumbers

We have already met the concept of *wavenumber*, which is usually defined in physics as $k = 2\pi/\lambda$ where λ is the wavelength. In molecular spectroscopy, wavenumber is often defined as $1/\lambda$. For example, the spectroscopic wavenumber of the $n = 1 \rightarrow 2$ transition for atomic hydrogen is about $82\,000\text{ cm}^{-1}$ or $8.2 \times 10^6\text{ m}^{-1}$.

The second mechanism of molecular absorption that we shall consider is vibration. The molecular bond between atoms behaves more or less like a spring. To model this, we consider a diatomic molecule consisting of two atoms, with masses m_1 and m_2 , connected by a spring with force constant (defined as dF/dx , where F is the tension and x is the extension) k , as shown in Figure 3.25. Classical mechanics gives the natural angular frequency of this system as

$$\omega_0 = \sqrt{\frac{k(m_1 + m_2)}{m_1 m_2}} \quad (3.79)$$

and quantum mechanics gives the energy levels as

$$E_v = \left(v + \frac{1}{2}\right)\hbar\omega_0, \quad (3.80)$$

where v is a quantum number that can take any non-negative integer value. This quantum number can only change by ± 1 , so in fact the only possible amount of energy that can be absorbed is $\Delta E = \hbar\omega_0$, giving an absorption line at the resonant frequency $f = \omega_0/2\pi$. Because the force constant k is of the order of 1000 N m^{-1} , the resonant frequency is typically between 10^{13} and 10^{14} Hz, corresponding to wavelengths generally in the thermal infrared region.

The last absorption mechanism we discuss is rotation. We consider a simple diatomic molecule consisting of two atoms, with masses m_1 and m_2 , separated by a fixed distance d (Figure 3.26). Classically, this system can rotate about the centre of mass of the two atoms. The moment of inertia of the system is given by

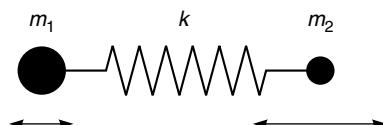


Figure 3.25. Classical model of molecular vibration. Atoms of masses m_1 and m_2 are connected by a spring of force constant k .

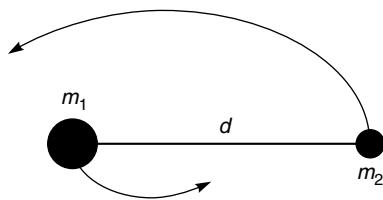


Figure 3.26. Classical model of molecular rotation. Atoms of masses m_1 and m_2 , separated by distance d , rotate about their common centre of mass.

$$I = \frac{m_1 m_2}{m_1 + m_2} d^2 \quad (3.81)$$

and, according to quantum mechanics, the energy of such a state is given by

$$E_J = \frac{J(J+1)\hbar^2}{2I}, \quad (3.82)$$

where J is a quantum number that can take any non-negative integer value. When electromagnetic radiation is absorbed, J must increase by 1. Calculating ΔE from Equation (3.82), for the transition from quantum number J to $J+1$, and substituting into Equation (3.77), we find that the frequency of a rotation absorption line is given by

$$f = \frac{(J+1)\hbar}{4\pi^2 I}.$$

As an example, we can consider the carbon monoxide, CO, molecule. This has $m_1 = 2.66 \times 10^{-26}$ kg, $m_2 = 1.99 \times 10^{-26}$ kg and $d = 1.13 \times 10^{-10}$ m, giving it a moment of inertia $I = 1.45 \times 10^{-46}$ kg m². We thus calculate that the $J = 0 \rightarrow 1$ transition will occur at a frequency of 116 GHz, which is in the microwave region. In general, we expect to find the rotational absorption lines in the microwave or far infrared regions of the electromagnetic spectrum, with frequencies typically between 10^{10} and 10^{12} Hz.

Although we have now outlined the most important mechanisms governing molecular absorption lines, there are further complications to be considered. *Combinations* of mechanisms can operate at the same time. For example, the energy level of a molecule can be described by both a rotational quantum number J and a vibrational quantum number v , and both of these may change in a transition. This gives rise to a more complicated vibrational–rotational spectrum in which a vibrational absorption line has fine structure superimposed on it as a result of different rotational transitions. We should also note that not all possible transitions can in fact be excited by electromagnetic radiation. For example, the hydrogen molecule, H₂, has a symmetric distribution of electric charge, which means that, classically speaking, an electric field cannot exert a force on it. We should therefore not expect molecular hydrogen to absorb electromagnetic radiation by vibrational or rotational transitions.

We have implicitly suggested that a molecular transition occurs at a single frequency, so that the absorption line in the spectrum has a width of zero. In fact, all lines are broadened to some extent. The Heisenberg uncertainty principle imposes a minimum line width, although this is negligible compared with other sources of line broadening. In proportion to the frequency of the line, the effect is largest for electronic transitions, and

3.4 Absorption and scattering by particles

even for these it is only of the order of 1 part in 10^8 . Much more significant is the effect of thermal motion of the gas. The line width Δf due to this effect, which is usually called *Doppler broadening*, is given by

$$\frac{\Delta f}{f} = \sqrt{\frac{RT}{m_m c^2}}, \quad (3.83)$$

where f is the frequency of the line, R is the gas constant, T is the absolute temperature, m_m is the mass of one mole of the gas, and c is the speed of light. Equation (3.83) shows that increasing the temperature will broaden the line, and that heavier molecules will exhibit narrower lines than lighter molecules. The fractional broadening due to this effect is typically 1 part in 10^6 .

Another important mechanism is *pressure broadening*, also called *collision broadening*. The molecules of the gas collide with one another and with other molecules in the atmosphere, and these collisions disturb the state of the molecule. The pressure broadening can be written approximately as

$$\Delta f \approx \frac{\sigma N_A p}{\sqrt{m_m RT}}, \quad (3.84)$$

where p is the gas pressure, σ is related to the collision cross-section for the molecules, and is of the order of 10^{-19} m^2 , and N_A is the Avogadro constant.

We can illustrate these formulae with two examples. First, we consider the $0.76 \mu\text{m}$ absorption line of the oxygen molecule, O_2 , at an altitude of 50 km, near the top of the stratosphere. The temperature at this altitude can be taken as 271 K, so Equation (3.83) gives the Doppler broadening $\Delta f/f$ as 9×10^{-7} . The pressure is about 80 Pa, so Equation (3.84) gives the pressure broadening as $\Delta f \approx 6 \times 10^5 \text{ Hz}$, and hence $\Delta f/f \approx 1.5 \times 10^{-9}$. In this case Doppler broadening dominates, and the line width can be specified as $3.5 \times 10^8 \text{ Hz}$ or about 0.01 cm^{-1} . Our second example is the 22.2 GHz absorption line of water, H_2O , at sea level. Taking the temperature as 288 K, we find from Equation (3.83) that the Doppler broadening is $\Delta f/f = 1.2 \times 10^{-6}$, and taking the pressure as 10^5 Pa , Equation (3.84) gives the pressure broadening as $\Delta f/f = 4.1 \times 10^{-2}$. In this case the pressure broadening dominates, and the line width is approximately 0.9 GHz (0.03 cm^{-1}).

We can also consider *scattering* by individual atoms and molecules. Since they are very much smaller than the wavelength of any radiation that we shall be considering, the Rayleigh scattering model is relevant. We can model a molecule rather crudely as a sphere of radius a , and if we assume that is electrically conducting, Equation (3.69) shows that its scattering cross-section will be given by

$$\sigma_s = \frac{128\pi^5 a^6}{3\lambda^4}. \quad (3.85)$$

The implications of this result for the propagation of light through the Earth's atmosphere are considered in Chapter 4. Other molecular scattering mechanisms also occur. In particular, scattering cross-sections are very much larger than implied by Equation (3.85) at frequencies close to absorption lines. This phenomenon is known as *resonant absorption*. Molecular fluorescence effects can also cause large scattering cross-sections.

Summary

When electromagnetic radiation encounters a particle, it can be absorbed (energy is transferred from the radiation to the particle) or scattered (radiation is redirected but no energy is transferred to the particle). The combined effect of absorption and scattering is called attenuation or extinction. The extent to which a particle absorbs or scatters radiation is specified by the appropriate cross-section σ , which has the dimensions of area, and if the particle is larger than an atom or a small molecule, a dimensionless absorption or scattering efficiency Q can also be defined as the ratio of the absorption or scattering cross-section to the geometrical cross-sectional area of the particle. As well as by the dielectric constant of the particle, the behaviour is controlled by the dimensionless size parameter χ , which expresses the size of the particle relative to the wavelength of the radiation. When the particle is very small compared with the wavelength, the scattering efficiency is proportional to χ^4 , i.e. the scattering cross-section is proportional to λ^{-4} . This is called Rayleigh scattering. Because of the dependence of the scattering cross-section on the wavelength of the radiation, this is also called selective scattering. If the dielectric constant of the particle has a non-zero imaginary part, absorption also occurs, and the absorption efficiency is proportional to χ . For particles much larger than the wavelength the absorption and scattering efficiencies reach values that are about 1, although the angular distribution of the scattered radiation becomes increasingly complicated as χ increases. The term Mie scattering is often used for the intermediate cases when the size parameter is about 1, and this case can be quite difficult to calculate.

Atoms and molecules also absorb and scatter radiation. Absorption is a quantum-mechanical process in which the atom or molecule absorbs just enough energy to promote it from one to another discrete energy level. Electronic transitions require the greatest amount of energy, typically needing radiation in the ultraviolet or visible parts of the spectrum. Vibrational transitions are the next most energetic, typically occurring in the infrared region, while the least energetic are rotational transitions, typically occurring in the microwave region of the spectrum. Absorption lines are broadened by the thermal motion of molecules. Atoms and molecules can scatter radiation by a number of processes including Rayleigh scattering.

3.5

The radiative transfer equation

The radiative transfer equation describes the behaviour of electromagnetic radiation in a medium in which both absorption (and hence emission) and scattering can occur. Because its form is rather daunting, we shall approach it through a number of special cases and then show how the general result is a natural extension from them. Much of the ground-work for this section has been laid in the previous sections of this chapter.

3.5.1

Propagation through an absorbing medium

We start by considering a medium in which only absorption occurs. If a beam of collimated radiation (i.e. radiation that is travelling a single direction) passes through an absorbing medium, the flux density F of the radiation varies with the propagation distance z according to the differential equation

$$\frac{dF}{dz} = -\gamma_a F. \quad (3.86)$$

This can be taken as a definition of the *absorption coefficient* γ_a , which clearly has the dimensions of 1/length. The negative sign in Equation (3.86) shows that the flux density of the radiation is reduced by absorption. We can see how the absorption coefficient, which is a macroscopic concept, is related to the absorption cross-section, which is a microscopic concept, by the following argument.

We suppose that the medium consists of absorbing particles, each of which has an absorption cross-section σ_a , and that the number of these particles per unit volume (the number density of the particles) is n . The radiation initially has a flux density F but then travels through a parallel-sided slab which has a cross-sectional area A and a thickness Δz . The number of particles contained within the slab is $nA\Delta z$ (Figure 3.27), and each one absorbs a power $\sigma_a F$ from the radiation (this follows from Equation 3.61) so the total

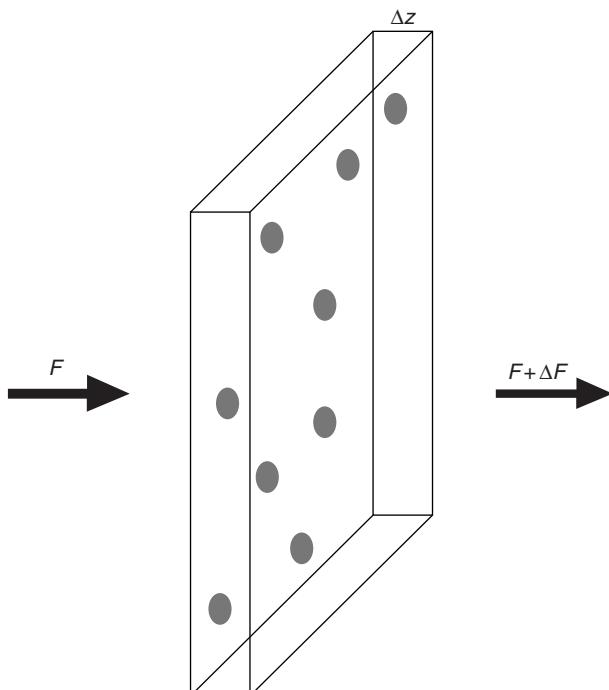


Figure 3.27. A beam of radiation is incident on a slab of thickness Δz containing absorbing particles. The emerging flux density has been shown as $F + \Delta F$ because ΔF denotes the change in the flux density. The effect of absorption is to cause ΔF to be negative.

power absorbed is $nA\Delta z\sigma_a F$. Dividing this by the cross-sectional area A , we find that the change in the flux density is given by

$$\Delta F = -n\sigma_a F \Delta z.$$

Taking the limit as Δz becomes very small, this expression can be rearranged to be the same as Equation (3.86) and we see that the absorption coefficient is related to the absorption cross-section through

$$\gamma_a = n\sigma_a. \quad (3.87)$$

If the absorption coefficient is constant, Equation (3.86) can be integrated to obtain the solution

$$F = F_0 \exp(-\gamma_a z). \quad (3.88)$$

This has the same form as Equation (3.14), and shows us that the absorption coefficient is just the reciprocal of the absorption length. F_0 is a constant in Equation (3.88), equal to the value of the flux density at $z = 0$.

If the absorption coefficient is not constant, Equation (3.86) can still be integrated although the solution has to be written in a different way:

$$F = F_0 \exp(-\tau). \quad (3.89)$$

In this expression, τ is a function of z given by

$$\tau = \int_0^z \gamma_a(z') dz'. \quad (3.90)$$

It measures the total amount of absorption between 0 and z , and is called the *optical thickness* between 0 and z . Optical thickness is an important concept in the radiative transfer equation.

3.5.2 Propagation through an absorbing and emitting medium

As we saw in Chapter 2, the phenomena of absorption and emission of radiation are closely connected. If an absorbing medium is in thermal equilibrium, i.e. it has a definite temperature T , it will also emit radiation. Although the radiation is emitted in all directions, none of the emitted radiation can find its way back into the direction of the original beam of radiation (we are not yet including any scattering in our model) so we can continue to represent the radiation by the amount travelling in a specific direction. However, it will now be more convenient to describe the amount of radiation in terms of the spectral radiance. The differential equation becomes

$$\frac{dL_f}{dz} = \gamma_a (B_f - L_f), \quad (3.91)$$

where

$$B_f = \frac{2hf^3}{c^2(e^{hf/kT} - 1)} \quad (3.92)$$

3.5 The radiative transfer equation

is just the spectral radiance of a black body at temperature T (Equation 2.31). Although Equation (3.91) can be solved as it stands, it is perhaps more instructive to assume that the frequency is low enough for the Rayleigh–Jeans approximation (Section 2.6) to be valid. In this case, Equation (3.91) can be rewritten in terms of the brightness temperature T_b of the radiation:

$$\frac{dT_b}{dz} = \gamma_a(T - T_b). \quad (3.93)$$

Suppose radiation is propagating in the $+z$ direction through an absorbing and emitting medium that has a uniform temperature T , and that the brightness temperature of the radiation is $T_b(0)$ at $z = 0$. Solving Equation (3.93) shows that the brightness temperature at some value of $z > 0$ is given by

$$T_b = T_b(0)\exp(-\tau) + T(1 - \exp(-\tau)), \quad (3.94)$$

where τ is the optical thickness of the medium between 0 and z , as defined in Section 3.5.1. This result has an intuitive interpretation. The brightness temperature at z is a weighted average of the input brightness temperature $T_b(0)$ and the temperature T of the medium itself. The coefficients of these two temperatures are $\exp(-\tau)$ and $(1 - \exp(-\tau))$ respectively. If the optical thickness τ is zero, the coefficients are 1 and 0, meaning that the output brightness temperature is the same as the input brightness temperature and the medium has contributed nothing. We say that the medium is *transparent*. If the optical thickness is infinite, the coefficients are 0 and 1, so that the output brightness temperature is the same as the physical temperature of the medium. The input brightness temperature has contributed nothing to the output because none of this radiation can penetrate through the medium, and we say that the medium is *opaque*. Figure 3.28 shows how these coefficients vary with the optical thickness of the medium.

We can also consider the situation when the physical temperature of the medium is not uniform. This could be, for example, a model of atmospheric temperature sounding, where the brightness temperature of upward-travelling radiation is measured at a point above the bulk of the Earth's atmosphere and used to make deductions about the temperature distribution within the atmosphere (Sections 6.7 and 7.5). It is much simpler in this case to work with the optical thickness τ rather than the position z , and we can note from Equation (3.90) that there is a monotonic relationship between τ and z . On this basis, the solution of Equation (3.91) can be written as

$$L_f(\tau) = L_f(0)\exp(-\tau) + \int_0^{\tau} B_f(\tau')\exp(\tau' - \tau)d\tau' \quad (3.95)$$

or, if the Rayleigh–Jeans approximation is valid, as

$$T_b(\tau) = T_b(0)\exp(-\tau) + \int_0^{\tau} T(\tau')\exp(\tau' - \tau)d\tau' \quad (3.96)$$

As before, an intuitive understanding of these equations is possible. For simplicity, we consider Equation (3.96), and assume that it applies to radiation propagating upwards

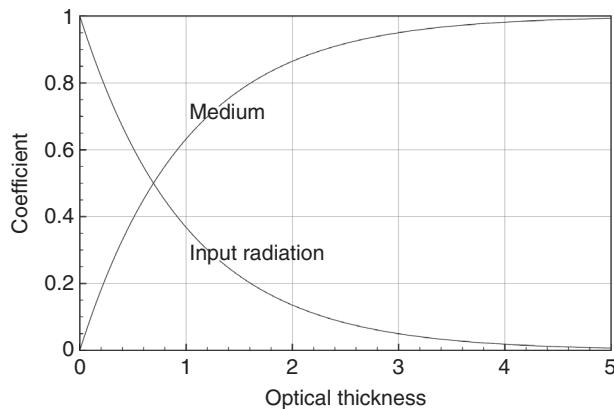


Figure 3.28. Contributions to the output brightness temperature from the input brightness temperature and the physical temperature of the propagation medium, as functions of the optical thickness of the medium.

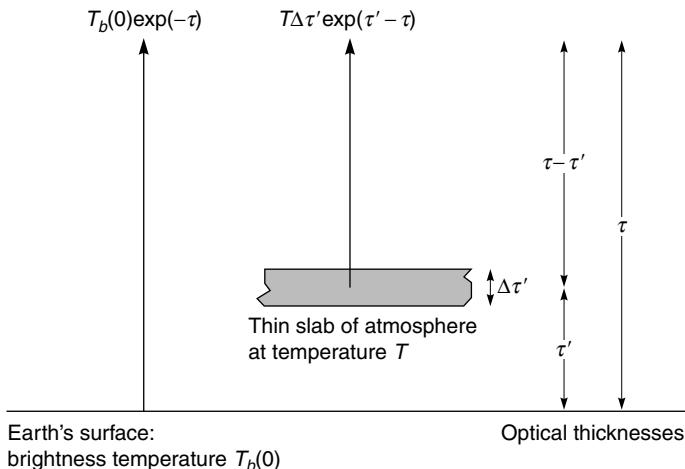


Figure 3.29. Contributions to the upward-propagating brightness temperature of the atmosphere. Heights above the Earth's surface are measured using the optical thickness τ .

from the Earth's surface (where z and τ are zero), through the atmosphere, for some total distance z_{\max} (see Figure 3.29). The brightness temperature of the radiation emitted from the Earth's surface is $T_b(0)$, and the total optical thickness of the path is τ , so the first term in Equation (3.96) corresponds to Equation (3.89) and just represents the absorption of the surface radiation. The integrand considers some position z such that the optical thickness between this point and $z = 0$ is τ' . The optical thickness between this point and $z = z_{\max}$ is $(\tau - \tau')$, so we expect the contribution from this point to be reduced by a factor of $\exp(-(\tau - \tau'))$. The optical thickness $(\tau - \tau')$ from the point in the atmosphere from which the radiation originates, to the point at which it is measured, is often called the *optical depth* of the former point. (The terms 'optical thickness' and 'optical depth' are often used interchangeably.)

3.5.3

A simple model of scattering and absorption: the two-stream approximation

In Section 3.5.1 we saw how to define the absorption coefficient, and how it is related to the scattering cross-section. We define the *scattering coefficient* γ_s analogously, and it is related to the scattering cross-section σ_s through

$$\gamma_s = n\sigma_s, \quad (3.97)$$

where n is the number density of scatterers in the medium. However, the differential equation is more complicated than Equation (3.86) because we have to take account of the direction of the scattered radiation, and the possibility that radiation can be scattered back into the original direction of propagation. Because this is conceptually rather difficult, we begin by developing a simpler model that contains the essence of scattering, but in which radiation can propagate in only two directions, forwards and backwards. (We should note here that we are considering only *elastic* scattering, in which the wavelength of the radiation is unchanged by the scattering process. Thus we are neglecting the phenomenon of *fluorescence*, in which radiation is absorbed at one wavelength and re-emitted at another, usually longer. Some minerals, especially sulphides, fluoresce in the visible part of the spectrum when excited by ultraviolet radiation. Plant material also shows a diagnostically useful fluorescence response. However, most fluorescence phenomena are too small to be measured accurately from airborne or (especially) spaceborne observations.)

Figure 3.30 shows radiation propagating the $+z$ and $-z$ directions in three adjacent parallel slabs, each of thickness Δz . We have used the symbols U and D (for ‘up’ and ‘down’) to represent the flux densities propagating in the two directions. When radiation is incident on one of these slabs, a fraction $\gamma_a\Delta z$ is absorbed and a fraction $\gamma_s\Delta z$ is scattered. We assume that *all* of this scattered radiation is scattered backwards, so that the fraction of the radiation that is transmitted through the slab is $1 - (\gamma_a + \gamma_s)\Delta z$. It is clear from Figure 3.30 that the flux density U in the middle slab is contributed to by the transmitted component of the positive-direction flux in the lower slab, and the reflected (scattered) component of the negative-direction flux in the middle slab, so we must have

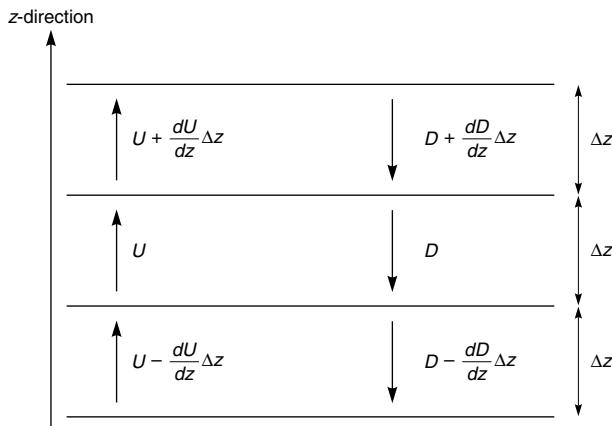


Figure 3.30. Radiation propagating in the $+z$ and $-z$ directions in three parallel slabs of thickness Δz .

$$U = \left(U - \frac{dU}{dz} \Delta z \right) \left(1 - (\gamma_a + \gamma_s) \Delta z \right) + D \gamma_s \Delta z.$$

Ignoring the terms in $(\Delta z)^2$ and rearranging, we find

$$\frac{dU}{dz} = -(\gamma_a + \gamma_s)U + \gamma_s D. \quad (3.98.1)$$

This is an intuitively reasonable equation. It shows that radiation is being lost from the forward direction as a result of both absorption and scattering, but gained from the scattering of backward-travelling radiation. The corresponding equation for the backward-travelling radiation is

$$\frac{dD}{dz} = -(\gamma_a + \gamma_s)D - \gamma_s U. \quad (3.98.2)$$

We can note from these equations the significance of the *sum* of the absorption and scattering coefficients in describing the propagation of radiation where both phenomena are important, since it represents the loss of energy from the forward-propagating radiation. This combination of absorption and scattering is the extinction, or attenuation, as already noted in Section 3.4. The extinction coefficient is given by

$$\gamma_e = \gamma_a + \gamma_s. \quad (3.99)$$

Equations (3.98) can be used to identify some of the important consequences of volume scattering. We consider an infinitely deep slab of material in which the absorption and scattering coefficients are constant, and radiation of unit flux density incident normally on the slab. To keep things as simple as possible, we also assume that the reflection coefficient at the surface of this slab is zero, so that only volume scattering is important. This situation is described by Equations (3.98.1) and (3.98.2) in the region $z \leq 0$, and subject to the boundary conditions that $D(0) = 1$, $U(-\infty) = D(-\infty) = 0$. It is not difficult to show that the solution is

$$D = \exp(\mu z)$$

$$U = \frac{\gamma_a + \gamma_s - \mu}{\gamma_s} \exp(\mu z)$$

where

$$\mu = \sqrt{\gamma_a^2 + 2\gamma_a\gamma_s},$$

so the intensity reflection coefficient is

$$R = \frac{\gamma_a + \gamma_s - \sqrt{\gamma_a^2 + 2\gamma_a\gamma_s}}{\gamma_s}. \quad (3.100)$$

This function, which depends only on the *ratio* of γ_s to γ_a , is shown in Figure 3.31.

Figure 3.31 shows that, if the scattering coefficient is much larger than the absorption coefficient, the volume scattering will be large. This is the reason why many finely divided materials, such as snow, clouds and (for example) table salt, are white. The total optical absorption in a slab of pure ice one metre thick is small, but if the ice is divided up

3.5 The radiative transfer equation

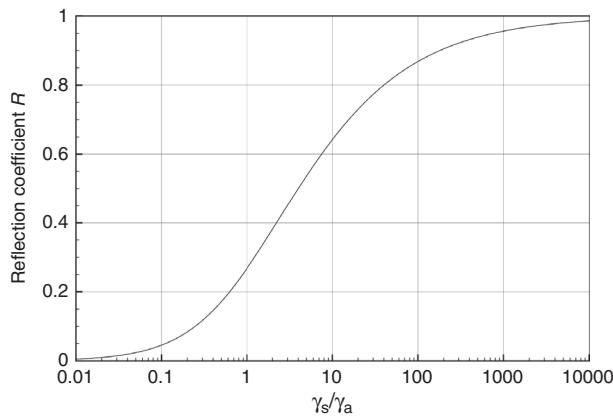


Figure 3.31. Dependence of the intensity reflection coefficient R for volume scattering on the ratio of the scattering coefficient to the absorption coefficient (two-stream model).

(a)



(b)

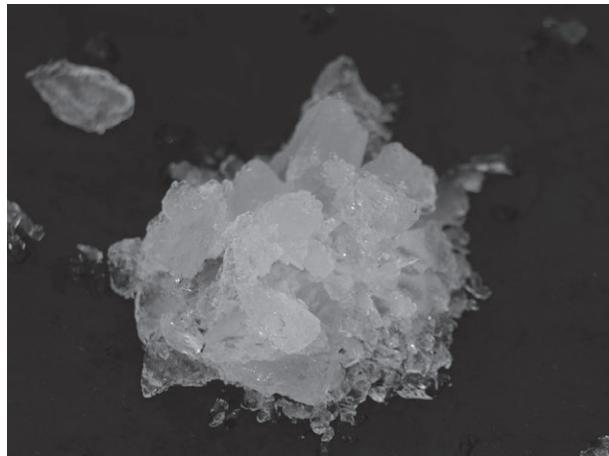


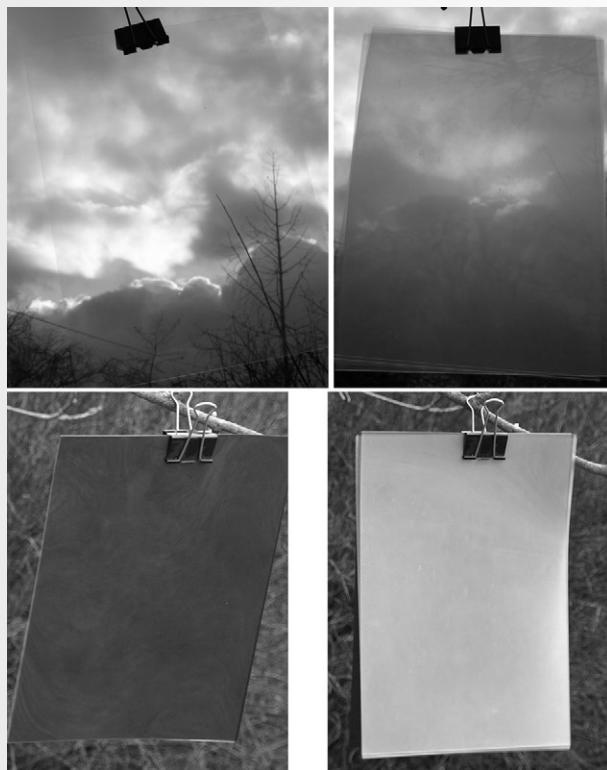
Figure 3.32. (a) A piece of ice is mostly transparent, but when the same piece is finely subdivided (in this case by smashing it with a hammer) (b) it becomes highly reflective.

into snow grains, each of which is only one millimetre across, a ray of light will encounter 2000 ice/air interfaces as it traverses the snow layer. Scattering can occur at each of these interfaces, and although the amount of scattering at each interface is small, the cumulative effect is large (Figure 3.32). Provided the absorption coefficient is small across the whole of the visible spectrum, as it is for ice, water and sodium chloride, the material will therefore appear white.

A simple demonstration of the two-stream model

A simple demonstration of the two-stream model is possible using a stack of the transparent sheets that are used with overhead projectors. Each sheet absorbs very little light, but there is a small but significant reflection at each surface of a sheet. If enough sheets are stacked on top of one another, the stack becomes quite strongly reflective.

The upper photographs show the appearance of the sky viewed through a single transparency and a stack of ten. The decrease in transmission as the number of sheets is increased is clearly apparent. The lower photographs show light reflected by a single transparency and a stack of ten, in each case backed by a sheet of black card so that no light is transmitted through the stack. The increase in reflection as the number of sheets increases is clearly apparent.



3.5.4

Scattering, absorption and emission

We are now in a position to state the full three-dimensional radiative transfer equation and to understand what each of the terms in it means. The form in which we write it shows how the spectral radiance L_f travelling in a particular direction (θ, φ) changes with the distance travelled in that direction:

$$\frac{dL_f(\theta, \phi)}{dz} = -(\gamma_a + \gamma_s)L_f(\theta, \phi) + \frac{\gamma_s}{4\pi} \int_{4\pi} L_f(\theta', \phi') p(\cos\Theta) d\Omega' + \gamma_a B_f. \quad (3.101)$$

We recognise the first term on the right as the loss due to attenuation (extinction, i.e. absorption and scattering combined), similarly to Equation (3.98) derived for the two-stream approximation. The last term on the right corresponds to the emission term in Equation (3.91), and indeed Equation (3.91) can be seen to be the special case of the radiative transfer equation when the scattering coefficient γ_s is zero. The second term on the right of Equation (3.101) describes the radiation scattered into the direction (θ, φ) from other directions specified by (θ', φ') . $d\Omega' = \sin\theta' d\theta' d\varphi'$ is an element of solid angle, and the integration is performed over all directions (i.e. over 4π steradians). $p(\cos\Theta)$ is the phase function of the scattering, which we met in Section 3.4. Its argument is the scattering angle Θ , given (by simple trigonometry) by

$$\cos\Theta = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi - \phi'). \quad (3.102)$$

Equation (3.101) has an understandable structure, but is potentially difficult to apply. Fortunately it is not always necessary to do so, and approximations can often give insight into the physical situation and indeed form the basis of quantitative models. In the last section of this long chapter we consider some practical examples of the interaction of radiation with matter.

Summary

The concepts of scattering and absorption cross-section s , which refer to individual particles, can be related to the corresponding properties of the medium as a continuum through the scattering and absorption coefficients γ . These are related to the corresponding cross-sections through $\gamma = n\sigma$ where n is the number density (number per unit volume) of the particles. The behaviour of radiation in a continuous medium in which scattering and absorption (and therefore also thermal emission) can take place is governed by the radiative transfer equation. This integro-differential equation is not always easy to solve, but some important special cases can be identified that are somewhat easier to understand.

If only absorption is important, the radiation is travelling in a single direction and the absorption coefficient γ_a does not vary spatially, the flux density F of the radiation, or anything proportional to it, decreases with the propagation distance z according to a negative exponential relationship: $F = F_0 \exp(-\gamma_a z)$. This is the Beer–Lambert law. If γ_a does vary spatially, the solution of the radiative transfer equation can be written as $F = F_0 \exp(-\tau)$

where the optical thickness τ is the integral of the absorption coefficient along the propagation path.

If both absorption and thermal emission are important, we can think of radiation entering, passing through, and emerging from a medium. The radiation that emerges is the sum of two components: an attenuated version of whatever radiation enters the medium, and a contribution from the medium itself. Although it is not necessary to use the Rayleigh–Jeans approximation to describe this situation in a way that is mathematically tractable, it is conceptually simpler to do so and it is usually justified in the case of microwave radiation. Using the Rayleigh–Jeans approximation, and also assuming that the physical temperature of the medium is uniform and equal to T , the expression for the brightness temperature of the radiation that emerges from the medium is $T_b(0)\exp(-\tau) + T(1 - \exp(-\tau))$. This can be understood simply as a weighted average of the input brightness temperature $T_b(0)$ and the physical temperature T . The weights are $\exp(-\tau)$ and $(1 - \exp(-\tau))$ respectively, where τ is the optical thickness of the medium. If τ is much smaller than 1 the medium is transparent and nearly all of the contribution to the output is from the input, whereas if τ is much greater than 1 the medium is opaque and nearly all of the contribution is from the medium itself. If the physical temperature of the medium varies spatially, the contribution to the brightness temperature of the emerging radiation from the medium is a weighted average of the physical temperature along the propagation path. This situation is important in understanding techniques for atmospheric temperature sounding.

In situations where both absorption and scattering are important, the behaviour of electromagnetic radiation in the medium depends on the ratio of the scattering coefficient of the absorbing coefficient as well as on the optical thickness of the medium. If the ratio is much greater than 1 it indicates that a photon has a much greater chance of being scattered than of being absorbed, and an optically thick medium will be highly reflective. This multiple-scattering phenomenon is the reason why weakly absorbing materials, when finely subdivided, scatter radiation strongly. It is observed at visible wavelengths for clouds, snow and many powdered materials, in the near-infrared for healthy green-leaved vegetation, and at microwave frequencies for dry snow packs and vegetation canopies.

3.6

Interaction of electromagnetic radiation with real materials

We have now considered the factors that govern the reflection of electromagnetic radiation from solid and liquid surfaces, and its behaviour in inhomogeneous media, in some detail, and we ought now to try to use these ideas to understand the way electromagnetic radiation interacts with some real materials. By extension, we can also consider the emission properties in those parts of the electromagnetic spectrum – namely, the thermal infrared and microwave regions – in which they are important. The treatment in this section is rather brief. A fuller discussion can be found in, for example, Elachi and van Zyl (2006).

3.6.1

Visible and near-infrared region

The visible and near-infrared (VNIR) region of the electromagnetic spectrum, from $0.4\text{ }\mu\text{m}$ to about $2\text{ }\mu\text{m}$, is still the most important for remote sensing of the Earth's surface. The majority of remote sensing systems (with the exception of those designed to probe the Earth's atmosphere) operate in this region, and the data are intelligible to comparatively unskilled users.

Broad-band optical data, in which the reflected radiation is integrated over the whole of the visible part of the spectrum, are needed especially for studies of the energy balance of the Earth system. This is, for example, roughly what is achieved by black-and-white (panchromatic) photography, and it is clear that what is measured is the spectrally averaged albedo. *Albedometers* are designed specifically to measure the broad-band albedo of surface materials, simultaneously integrating the light input from above and reflected upwards from the surface below (Figure 3.33).

Figure 3.34 illustrates typical values of the albedo of various materials.

The lowest albedos are usually shown by water surfaces. The refractive index of water for visible light is roughly 1.33, so for normally incident radiation the Fresnel coefficient is given by Equation (3.35) as 0.14 for the amplitude, and hence 0.02 for the intensity. The low albedo of pure water can therefore be explained in terms of the fact that the refractive index of water is not very different from 1. For 'real', naturally occurring, water, the albedo can be somewhat higher as a result of scattering by suspended particulate matter, or lower because of absorption. The refractive index of pure ice is similar to that of water, but naturally occurring ice generally contains a significant number density of trapped air bubbles, which give rise to volume scattering, and hence a higher albedo.

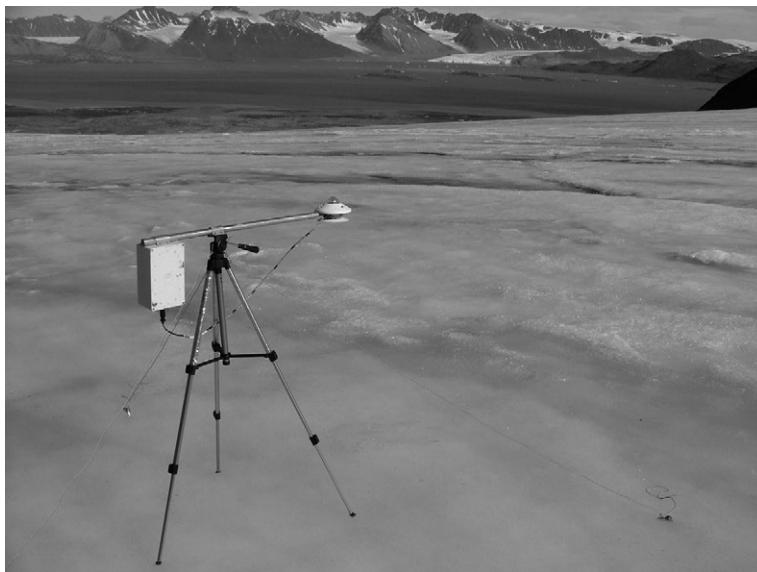


Figure 3.33. Albedometer installed on the surface of a glacier. The instrumentation and battery are housed in the box on the left, while the unit on the boom at the right integrates the downwelling light from the sky and the upwelling light reflected from the glacier's surface.

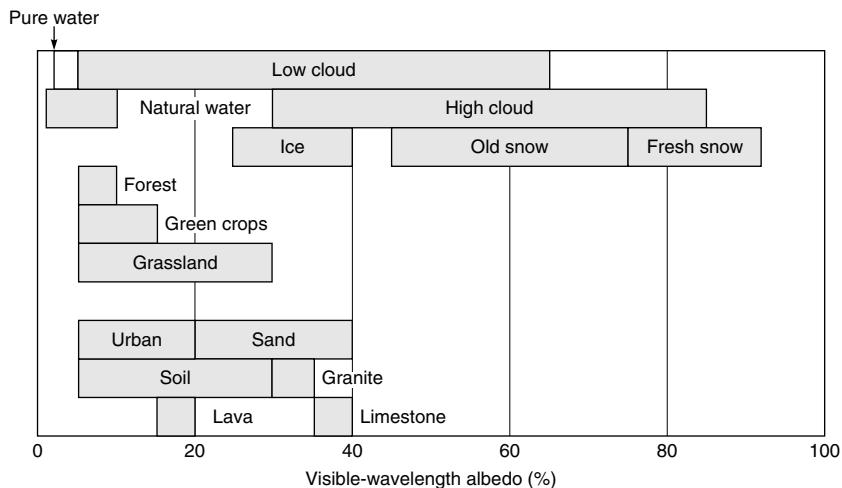


Figure 3.34. Typical values of albedo integrated over the visible waveband for normally incident radiation. (Mostly after Schanda (1986).)

Minerals, soils and the materials from which the visible parts of urban areas are composed show albedos typically in the range 5–40%. The dominant factor in these cases is the refractive index, since the scope for volume scattering is very limited. It is clear that the refractive index must be somewhat larger than that of water, typically between 1.6 and 4.5. If the materials are wet, the refractive index contrast between air and the material is reduced, and the reflectance is also reduced (Figure 3.35). Soils with a high content of organic material, which strongly absorbs light, also show low reflectance.

The albedos of clouds (which are composed of small water droplets or ice crystals) and snow (a mixture of ice crystals, air and, in the case of wet snow, liquid water) are dominated by volume-scattering effects. We saw in Section 3.5.3 that the importance of scattering is governed by the ratio γ_s/γ_a , provided that the medium is optically thick. A typical cumulus cloud might contain a number density $n = 2 \times 10^7 \text{ m}^{-3}$ of water droplets having a typical radius $a = 2 \times 10^{-5} \text{ m}$. The particle size parameter $\chi = 2\pi a/\lambda$ must be very large (several hundred) at optical wavelengths, so we expect geometrical scattering to occur. The scattering cross-section will be of the order of πa^2 , and the scattering coefficient γ_s will therefore be of the order of $n\pi a^2$ or roughly 0.02 m^{-1} . A cloud 1000 m thick would thus have an optical thickness of around 50, i.e. it would be highly opaque. This is quite consistent with our observation of clouds. How can we estimate the absorption coefficient? Figure 3.1 shows that the absorption length for visible light in pure water is greater than 10 m, so the absorption coefficient of pure water is less than about 0.1 m^{-1} . However, most of the cloud is ‘empty’, i.e. air rather than water. One cubic metre of cloud contains only $4\pi a^3 n/3 = 7 \times 10^{-7} \text{ m}^3$ of water, so the average absorption coefficient of the cloud is roughly 10^{-7} m^{-1} . Thus $\gamma_s/\gamma_a \approx 3 \times 10^5$, i.e. hugely greater than 1, and we expect virtually all the radiation that enters the cloud to be scattered out of it. A similar argument applies to snow.

Finally, we consider the case of vegetation. Figure 3.34 shows quite low values of albedo for these materials in the optical region. This is again a consequence of absorption, by light-absorbing pigments (principally *chlorophyll* in the case of green vegetation), and



Figure 3.35. Extract of a Landsat image showing part of the north Norfolk coast, UK. Holkham beach is left centre. Areas of dry sand exhibit significantly higher reflectance than areas of wet sand.

is unsurprising since plants derive energy by absorbing light. However, Figure 3.34 gives a very incomplete picture of the reflectance properties of vegetation in particular. To understand these better, we should examine the variation of reflectance with wavelength throughout the VIR region.

Figure 3.36 illustrates schematically the spectral reflectance (which we use as a general term to denote certain but unspecified conditions of illumination and viewing geometry) of various materials, in the wavelength range from 0.4 to 2.4 μm . Data such as these are often collected, for surface materials anyway, using a reflectance spectroradiometer (Figure 3.37). The data are simplified, and fine detail has been omitted. Nevertheless, it is clear that in many cases the shape of such a curve (often called the *spectral signature*) is characteristic of the material, and indeed we are familiar with the idea that the colour of an object often gives a major clue to identifying it. The manner in which remotely sensed images with spectral content can be analysed to identify the probable distribution of materials represented within it will be discussed in Chapter 11.

Let us comment on each of the spectral signatures of Figure 3.36 in turn. In the case of water, there is little spectral structure, just a decline in reflectance at longer wavelengths as a result of increasing absorption. Limestone also exhibits little spectral structure (at the resolution of the figure), and dry soils (not shown in the figure) broadly follow the same behaviour. However, moist loam exhibits a spectral reflectance that is somewhat lower, as we expect because of the reduced contrast in refractive index, and also shows a number of oscillations in the infrared part of the spectrum. These are due to *absorption lines* (absorption maximum near particular wavelengths) of water molecules. Much of this

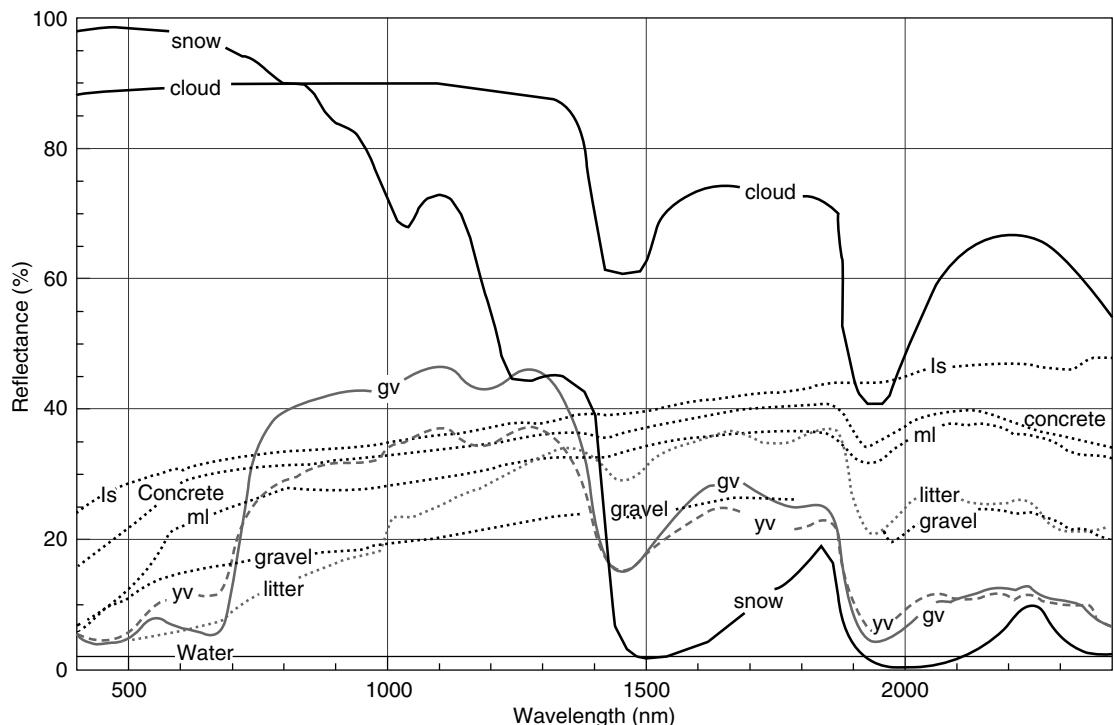


Figure 3.36. Typical spectral reflectances of various materials in the VNIR regions (schematic). gv: green vegetation; yv: yellow (senescent) vegetation; ls: limestone; ml: moist loam.

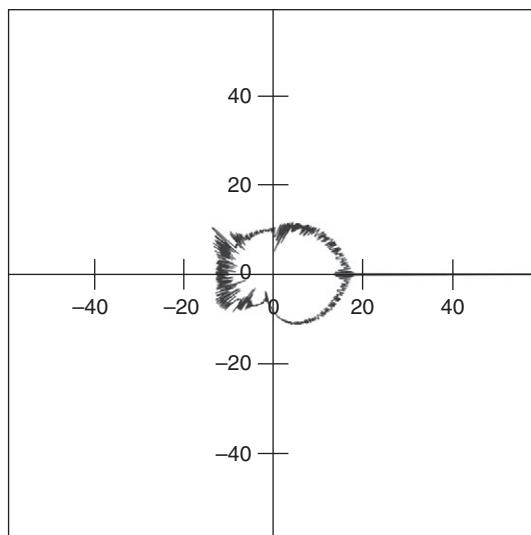


Figure 3.37. Field-portable reflectance spectroradiometer being used to measure the reflectance spectrum of snow. The instrumentation is housed in the user's backpack, connected to the sensor (on the end of a pole) by a fibre-optic cable. The unit is controlled by a laptop computer.

3.6 Interaction of electromagnetic radiation with real materials

same absorption-line structure can also be seen in the spectral signature of snow. Cloud also exhibits an oscillatory spectral signature in the infrared part of the spectrum.

Finally, we turn to the very characteristic spectral signature of vegetation (see also Campbell (2008)), illustrated in Figure 3.36 by the example of a dwarf shrub called *Betula nana*. As we have remarked already, the low reflectance in the optical region is largely governed by the presence of pigments. The most important of these is chlorophyll, which has absorption maxima at 0.45 and 0.65 μm (respectively blue and red), and consequently gives rise to a local maximum in the spectral reflectance at about 0.55 μm . This is the explanation for the green colour of much vegetable matter. The other important botanical pigments are carotene and xanthophyll (which give orange-yellow reflectance spectra) and the anthocyanins (red-violet). These latter pigments are dominant in the autumn, when the chlorophyll decomposes in many species, and give rise to the spectacular colours of autumn leaves (see Justice *et al.* (1985)).

The high reflectance of vegetation in the near-infrared, roughly between 0.7 and 1.3 μm , is a scattering effect. It is principally caused by multiple internal reflections of the radiation from hydrated cell walls in the *mesophyll* of the leaves, with little absorption. The transition from low reflectance in the red part of the spectrum to the much higher reflectance in the near-infrared is rather sharp, and is often termed the ‘*red edge*’. Above about 1.3 μm , the absorption of infrared radiation by water becomes significant (again we see the oscillatory structure that we noted for moist loam and for snow) and the reflectance is reduced. If the vegetation is senescent, stressed or diseased, the cell structure is less well developed or damaged and the high reflectance between 0.7 and 1.3 μm is reduced. Stressed or senescent vegetation may also have a lower chlorophyll content than healthy vegetation, which will tend to increase the reflectance in the red part of the spectrum, so that the steepness of the red edge will be reduced. This phenomenon forms the basis of a number of techniques for assessing the amount and health of the vegetation present in a remotely sensed image, discussed in more detail in Chapter 11.

3.6.2 Emissivities in the thermal infrared region

The thermal infrared region was defined in Chapter 2 as the range of wavelengths from 3 to 15 μm , and we saw in Section 2.6 that objects at normal terrestrial temperatures radiate maximally in this part of the electromagnetic spectrum. The emissivity of a body (loosely speaking, its efficiency of emitting black-body radiation) at a given wavelength is negatively related to the reflectivity at the same wavelength, such that when the reflectivity is zero the emissivity is 1 and conversely. For most naturally occurring materials, with the exception of metals and some minerals, the refractive index in the thermal infrared region is close to 1, which means that the reflectivity is low and the emissivity high. Figure 3.38 shows the emissivities of various materials.

It should be noted that the emissivities shown in Figure 3.38 are somewhat schematic, since the surface roughness of a material will also influence its emissivity. For example, a granite rock has a normal emissivity of about 0.89, but if the same material is highly polished, its emissivity will fall to about 0.80. In fact, the emissivity can be written generally in terms of the BRDF as

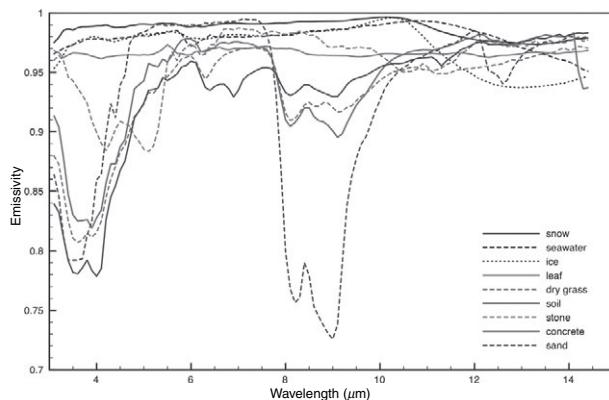


Figure 3.38. Typical emissivities of various materials at normal incidence in the range 3–15 μm . The data have been averaged from a large number of MODIS samples (Anon.). The category ‘concrete’ in fact also includes a wide variety of building materials such as brick, tile and asphalt. The figure excludes metal surfaces, which have much lower emissivities. See also colour plates section.

$$\varepsilon_p = 1 - \int (R_{pp} + R_{qp}) \cos \theta_1 d\Omega_1, \quad (3.103)$$

where ε_p denotes the emissivity for p -polarised radiation and R_{qp} is the BRDF for radiation that is incident in the q -polarised state and reflected in the p -polarised state. θ_1 is the angle between the reflected radiation and the surface normal, and $d\Omega_1 = \sin \theta_1 d\theta_1 d\varphi_1$ where φ_1 is the azimuth angle of the reflected radiation. The integration is carried out over 2π steradians, i.e. $\theta_1 = 0$ to $\pi/2$ and $\varphi_1 = 0$ to 2π . Emissivities can thus be calculated using Equation (3.91) and the discussion of BRDFs presented in Section 3.3.

3.6.3

Emissivities in the microwave region

Thermally generated radiation can also be detected in the microwave region of the electromagnetic spectrum, i.e. at wavelengths between 1 mm and 1 m (frequencies between 0.3 and 300 GHz). As with thermal infrared emission, the physical parameter that determines the quantity of radiation that is emitted at a given temperature is the emissivity, and the principles that determine the emissivity are the same. In practical terms, however, the situation is more complicated than for thermal infrared emission. First, the range of wavelengths over which microwave observations can be made is much larger than for infrared observations. The latter are normally made in a single, rather broad, waveband from (typically) 8 to 12 μm , or perhaps in two narrower wavebands within this range, whereas microwave observations are routinely made at a number of frequencies spanning the range from (typically) 4 GHz to more than 40 GHz – a factor of more than 10. It is therefore necessary to consider the variation of emissivity with frequency. Second, microwave observations are often made at angles away from the surface normal, so it is important to consider the dependence of the emissivity on the viewing direction. Finally, the emissivities are often significantly different for different polarisation states, so that the dependence of polarisation must also be considered. These

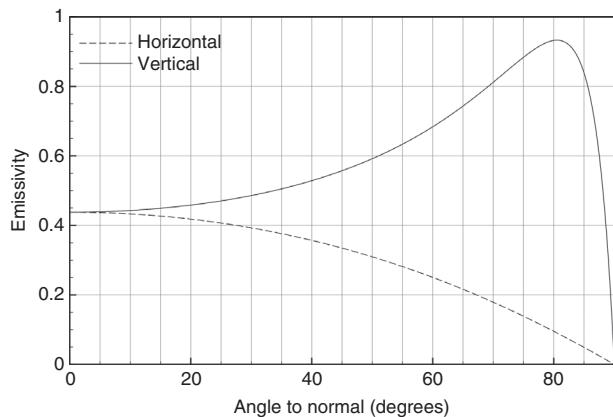


Figure 3.39. Emissivity of a material with a dielectric constant $18.5 - 31.3i$ as a function of observation angle, for vertically and horizontally polarised radiation.

factors greatly increase the difficulty of providing a simple characterisation of the microwave emissivities of ‘typical’ materials. We therefore content ourselves with illustrating the main features. A much fuller discussion is presented by Ulaby, Moore and Fung (1982).

For a homogeneous material (i.e. one in which volume-scattering effects are unimportant) with a smooth surface, the emissivity is given by $1 - |r|^2$, where r is the Fresnel reflection coefficient appropriate to the direction and polarisation of the radiation (this follows from Equation 3.91). Figure 3.39 illustrates this type of behaviour. The figure has been calculated using the Fresnel coefficients for a material with a dielectric constant of $18.5 - 31.3i$, which is appropriate for calm *sea water* at a temperature of 20 °C and a frequency of 35 GHz. It can be noted from the figure that the emissivity for vertically polarised radiation increases to a maximum near 1 at about 80 °C. This is the phenomenon of the Brewster angle that we noted in Section 3.2. The reason that the emissivity does not quite reach a value of 1 is because the dielectric constant is not real but complex. (We may also note in passing that the dielectric constant, and hence emissivity, of a water surface depends on temperature.)

Continuing with our example of sea water, we can consider the effect of varying the *salinity*. Increasing the salinity will increase the electrical conductivity of the water, and hence the Fresnel reflection coefficients, which will in turn lower the emissivity. However, we saw in Figure 3.2 that the dielectric constant of sea water differs significantly from that of pure water only at frequencies below about 5 GHz, so we should not expect any significant dependence of emissivity on salinity above this frequency. We may also consider the effect of a rough surface, such as would be produced by *wind action*. This will lower the reflectance, and hence raise the emissivity, for observing directions away from the normal. Off-nadir observations of the microwave emission from a sea surface therefore have the potential to measure the sea state. The effect is greater at higher frequencies, which is consistent with the scattering models that we discussed in Section 3.3.4, and is greater for horizontally than for vertically polarised radiation.

The microwave emissivity of a bare *soil* surface is dominated by the surface roughness and by the *moisture* content of the soil. The dielectric constant of water in the microwave

region is much higher than that of soil (typically around 3), so increasing moisture content will increase the reflectance and hence decrease the emissivity. Typical emissivities for soil surfaces lie in the range 0.5 to 0.95.

Scattering effects are negligible in soils unless they are extremely dry. However, these effects can dominate in determining the emissivity properties of more open structures such as dry *snow packs* and *vegetation canopies*. In such cases, the emissivity is given by

$$1 - \frac{\gamma_s}{\gamma_e}$$

for an optically thick medium in which multiple scatter is insignificant. Vegetation canopies have microwave emissivities typically in the range 0.85 to 0.99. A deep, dry snow pack has an emissivity of about 0.6. If the medium is not optically thick, the measured emission will include a contribution from the surface below, for example from the soil surface below a vegetation canopy. For a dry snow pack, the optical thickness at normal incidence is proportional to the total mass of ice per unit area of the pack, and this effect can therefore be used to estimate the snow mass. For wet snow, scattering effects are insignificant and the emissivity is much higher, typically 0.95.

3.6.4

Effect of clouds and snow on microwave radiation

As we did in Section 3.6.1, let us also consider the interaction of microwave radiation with a *cloud*. Specifically, we consider the absorption and scattering of 1-cm (30-GHz) microwaves in the same cumulus cloud that we used in Section 3.6.1. The size parameter χ of the water droplets is 0.0126, which is small enough for the Rayleigh scattering expressions (3.68) and (3.70) to be used. The refractive index of water at this wavelength is $5.67 - 2.88i$, so $Q_a = 2.85 \times 10^{-3}$ and $Q_s = 6.16 \times 10^{-8}$. (These values were calculated using Equations 3.68 and 3.70, and are accurate to 0.2% and 0.1% respectively.) Thus the absorption and scattering cross-sections are $3.58 \times 10^{-12} \text{ m}^2$ and $7.74 \times 10^{-17} \text{ m}^2$ respectively, giving absorption and scattering coefficients of $7.2 \times 10^{-5} \text{ m}^{-1}$ and $1.5 \times 10^{-9} \text{ m}^{-1}$ respectively. Absorption is by far the dominant process, but the cloud is still almost transparent since $\tau = 0.07$. The absorption of microwave radiation by cloud is strongly dependent on the frequency of the radiation. This is discussed in more detail in Chapter 4.

Decibels

If the intensity of radiation (or anything else) is reduced by a factor of x , this can equivalently be expressed as reduction by $10 \log_{10}(x)$ decibels. Thus, for example, an absorption coefficient of $y \text{ m}^{-1}$ can also be expressed as $10/\ln(10)y \approx 4.34y \text{ dB m}^{-1}$. The absorption coefficient of the cumulus cloud for 1-cm microwave radiation in the example in this section is therefore about 0.3 dB km^{-1} .

We can also consider the effect of *rain* on microwave radiation. A light rainfall might contain around 100 droplets per cubic metre, with a typical radius of 1 mm. At a wavelength of 1 cm, the size parameter of the droplets is 0.63. This is an intermediate

3.6 Interaction of electromagnetic radiation with real materials

value, small enough neither for the approximations of Mie scattering nor of geometrical scattering to be valid, so we have to evaluate the cross-sections rather more elaborately. We find $Q_a = 0.96$ and $Q_s = 0.56$, so the absorption and scattering cross-sections are $3.02 \times 10^{-6} \text{ m}^2$ and $1.76 \times 10^{-6} \text{ m}^2$ respectively, giving an absorption coefficient of $3.02 \times 10^{-4} \text{ m}^{-1}$ (1.3 dB/km) and a scattering coefficient of $1.76 \times 10^{-4} \text{ m}^{-1}$ (0.8 dB/km). Here scattering, although still (just) smaller than absorption, occurs at a significant level, and this illustrates the possibility of monitoring rainfall from ground-based *rain radars*.

3.6.5 Microwave backscattering coefficients

From a practical point of view, it is as difficult to present characteristic data on the microwave backscattering properties of ‘typical’ materials as for their microwave emissivities. Again, the number of combinations of different observing parameters (frequency, polarisation and incidence angle) is very large, and the influence of natural variations in the physical properties of the scattering materials must also be taken into account. However, by way of example, Figure 3.40 (which is mainly based on material presented by Long (2001)) illustrates the angular dependence of the backscattering coefficient σ^0 of a few representative materials at X-band. A much fuller discussion is presented by Ulaby, Moore, and Fung (1982).

The backscatter from a concrete road surface is almost entirely due to surface scattering processes, and the low values of σ^0 at angles away from normal are consistent with a smooth surface, as expected. Surface scattering also dominates the backscatter from wet snow, the sea (in which case the influence of wind speed on surface roughness is apparent), and urban areas. The dependence of the backscatter from a sea surface on the wind speed means that wind speed over the sea can be inferred from appropriate backscatter measurements (obviously, one would avoid incidence angles close to 10° where the sensitivity is very small). In fact, there is also a small dependence on wind *direction* (the data in Figure 3.40c have been plotted for the case when the radar look direction is upwind), which can be used to infer the wind velocity. The microwave backscattering properties of ocean surfaces are discussed further in Section 9.3.1.1.

The high backscatter from urban areas, not strongly dependent on incidence angle, arises mainly from the presence of large numbers of planar surfaces, oriented at right angles to one another, in such areas. Figure 3.41 illustrates one of the mechanisms by which this occurs. If two reflecting planes are arranged perpendicularly (for example, one plane could be the wall of a building and the other a road surface), incident radiation that strikes the concave region between the planes from any direction in the plane that contains the normals to both surfaces will be reflected back along its incidence direction. This is called a dihedral reflector. In fact, if *three* planar surfaces meet at right angles, in a trihedral reflector, radiation from *any* direction that strikes the concave region between them will be scattered back along its incidence direction. Thus a region that contains a large number of such ‘inside corners’ will give strong specular scattering from most incidence directions.

Dry deserts, vegetation canopies and dry snow packs are examples of media from which volume-scattering effects are important. In general, the backscattering in such cases will show a weaker dependence on incidence angle than is observed in cases where surface scattering dominates. A dry desert can be assumed to be optically thick (physical

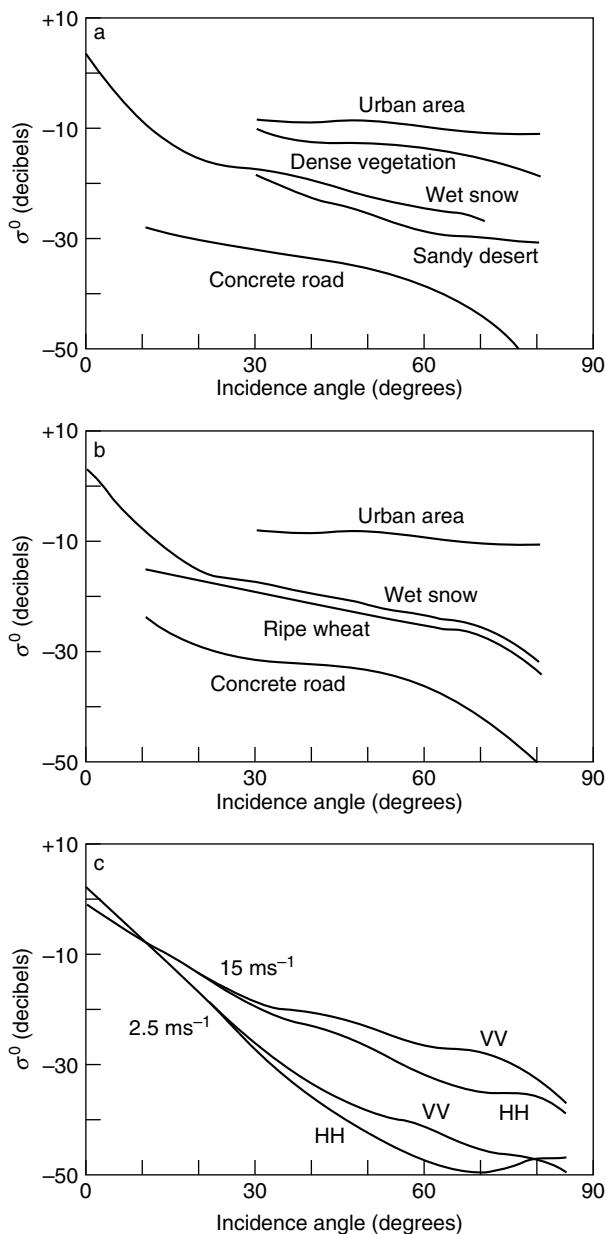


Figure 3.40. Representative X-band backscattering coefficients. (a) Various materials at HH-polarisation; (b) various materials at VV-polarisation; (c) sea surfaces at HH- and VV-polarisations. The curves have been labelled with the corresponding wind speed.

depth much larger than the attenuation length). However, for dry snow and vegetation canopies this assumption may not be true. For example, radiation incident close to vertically on an agricultural crop is likely to be scattered significantly from the soil below. In the case of dry snow, the assumption of optical thickness is often not valid

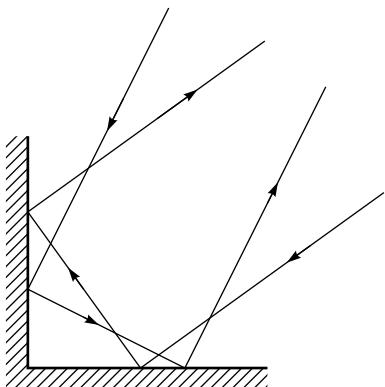


Figure 3.41. Radiation is reflected by two perpendicular planes so that it returns along the incidence direction. This is an important contributor to the strong microwave backscatter from urban areas.

since the attenuation depth may be many metres. Thus, a dry snow cover a few tens of centimetres deep may be practically invisible to microwave radiation.

The data presented in Figure 3.40 show only the co-polarised (like-polarised) back-scattering coefficients, i.e. HH and VV. In general, the cross-polarised components (HV and VH) are extremely small unless multiple volume-scattering occurs.

3.6.6 Modelling microwave backscattering: case study of a snowpack

We saw in Section 3.5 that radiative transfer is in general a mathematically complicated phenomenon. No single model can embrace the range of possibilities that may occur. In this section, however, we shall develop one practical model of backscattering. This was originally derived for the microwave backscatter from a snow pack, and is based on the work of Stiles and Ulaby (1982).

Figure 3.42 shows the situation schematically. Radiation is incident on the snowpack at an angle θ to the normal, and a fraction $T(\theta)$ of the incident power is transmitted across the interface, where

$$T(\theta) = |t(\theta)|^2$$

and $t(\theta)$ is the appropriate Fresnel coefficient (Equation 3.31.2 or 3.31.4). This radiation is refracted towards the normal at an angle θ' , calculated from Snell's law (Equation 3.30).

Once inside the snowpack, the radiation undergoes volume scattering. In deriving Equation (3.100), we saw that the power reflected out of an infinite volume-scattering medium in a simple one-dimensional model depended only on the ratio $\gamma_s/\gamma_a = \sigma_s/\sigma_a$. Plausibly, we may write the power backscattered from an infinite three-dimensional medium as

$$\frac{\gamma_s \cos \theta'}{2\gamma_e} = \frac{\sigma_s \cos \theta'}{2\sigma_e},$$

where the factor $(\cos \theta')/2$ has been introduced to account for the fraction of the total scattered power that is directed in the backscattering direction. However, the medium is not infinite, but instead has a depth d . The optical thickness of the path from the snow surface to the ground beneath is given by

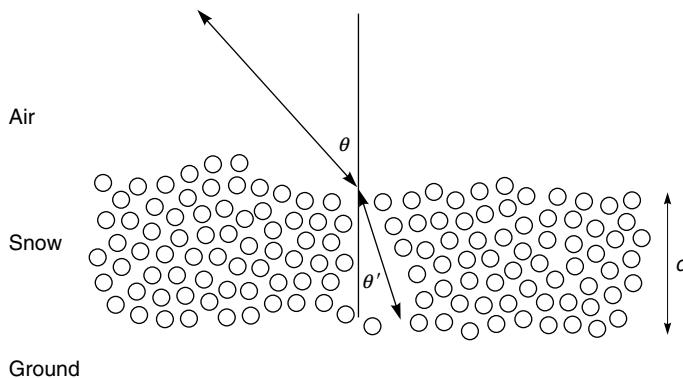


Figure 3.42. Volume scattering of microwave radiation from a snow pack (schematic).

$$\tau(\theta') = \gamma_e d \sec \theta',$$

and the fraction of power remaining after traversing this path twice (once in each direction) is $\exp(-2\tau(\theta'))$. If this fraction is zero, the medium is effectively infinitely deep so that volume scattering will contribute fully, whereas if it is 1 the medium is transparent and volume scattering will not contribute at all. Thus, we may plausibly write

$$\frac{\sigma_s \cos \theta'}{2\sigma_e} \left(1 - \exp(-2\tau(\theta')) \right)$$

for the fraction of the power entering the snow pack that is scattered back to the surface. This radiation must again be transmitted through the snow-air interface, so our expression for the volume-scattered component of the backscatter coefficient becomes

$$T^2(\theta) \frac{\sigma_s \cos \theta'}{2\sigma_e} \left(1 - \exp(-2\tau(\theta')) \right).$$

To complete the model, we need also to consider the surface scattering component and the component that is scattered from the snow-ground interface. We can write the latter term as

$$T^2(\theta) \sigma_{\text{ground}}^0 \exp(-2\tau(\theta')),$$

to take into account that this radiation must pass through the air-snow interface twice and must also traverse the optical thickness $\tau(\theta')$ twice, so our final formula becomes

$$\sigma_\theta^0 = T^2(\theta) \left(\frac{\sigma_s \cos \theta'}{2\sigma_e} \left(1 - \exp(-2\tau(\theta')) \right) + \sigma_{\text{ground}}^0 \exp(-2\tau(\theta')) \right) + \sigma_{\text{surface}}^0(\theta). \quad (3.104)$$

It should be emphasised that this is in fact a very simple volume-scattering model which does not take account of, amongst other things, multiple scattering. Nevertheless, it illustrates the principal considerations involved in such models.

Summary

The visible-wavelength albedos of naturally occurring materials range from a few per cent or less, to almost 100%. Several factors contribute to this wide range of values, including refractive index contrast, multiple scattering, and absorption by pigments. Water contributes both one of the lowest-albedo materials, as large, clear bodies of fresh or sea water, and some of the highest, in the form of snow and cloud. Although the average visible-waveband albedo can discriminate to a limited extent between different materials on the Earth's surface, the spectral reflectance, in which the reflectivity, albedo or some similar measure is plotted as a function of wavelength, provides much greater scope for recognising different materials. Snow, cloud and vegetation have particularly characteristic signatures, especially in the near-infrared part of the spectrum.

In the thermal infrared region, most materials with the exception of metals and minerals have emissivities that are above 0.9. In the microwave region, emissivities are much more variable and their dependence on frequency and on polarisation state must be taken into account. Volume-scattering effects are significant in some cases such as dry snow and vegetation canopies. Microwave radiation interacts weakly with clouds except at the highest frequencies, but somewhat more strongly with rain, and this can be useful in rain radar.

Microwave backscattering is also a complicated situation with a dependence on many variables. In most cases the scattering is dominated by surface effects so that dielectric constant, surface roughness characteristics and imaging geometry are all significant. In the cases of dry deserts, snow packs and vegetation canopies, volume-scattering effects are significant and the dependence of backscattering coefficient on incidence angle may be reduced. Urban areas generally show high levels of backscatter not strongly dependent on incidence angle.

Review questions

Discuss the significance of complex dielectric constants or refractive indices.

What factors govern the dielectric constants of air, water, sea water and (at radio frequencies only) the ionosphere?

Discuss the phenomenon of wave dispersion.

Explain the meaning of Fresnel coefficients and the Brewster angle.

Explain why you would expect dry earth or sand to have a higher albedo than when it is wet.

Why does reflection generally change the polarisation state of electromagnetic radiation?

What is meant by the bidirectional reflectance distribution function?

Explain the difference between Lambertian and specular reflection. What determines which of them is more likely to be a good model of the behaviour of electromagnetic radiation at a surface?

What are the main parameters controlling microwave backscatter from a rough surface?

Draw sketches showing how you would expect the backscattering coefficient to depend on incidence angle for a rough surface and a smooth surface.

Explain what is meant by the cross-section σ , the efficiency Q and the coefficient γ for scattering, and the size parameter χ of a particle. Show how they are related.

Outline the meaning of the scattering phase function, and explain why it is important in considering radiative transfer in a medium in which scattering is significant.

Outline the quantum-mechanical processes responsible for absorption of electromagnetic radiation by atoms and molecules.

Describe the Beer–Lambert law and explain how it can be modified in cases where both absorption and thermal emission of radiation are important.

Explain what is meant by *optical thickness*.

Explain why clouds are white and opaque in visible light.

Discuss the effect of clouds and rain on microwave radiation.

Problems

1. Show that the real and imaginary parts of the refractive index of a non-magnetic material are given by $m^2 = \frac{\sqrt{\epsilon'{}^2 + \epsilon''{}^2} + \epsilon'}{2} \kappa^2 = \frac{\sqrt{\epsilon'{}^2 + \epsilon''{}^2} - \epsilon'}{2}$.
2. A commonly used approximation to the absorption length is $l_a \approx \frac{\sqrt{\epsilon'}}{2\pi\epsilon''} \lambda_0$, where λ_0 is the free-space wavelength. Show that this approximation is accurate to within 1% if $\epsilon''/\epsilon' < 0.28$.
3. Sea water has an electrical conductivity σ of typically about 4 S m⁻¹. The real part of its dielectric constant at radio frequencies of 100 MHz and below is 88.2. By assuming that the imaginary part is given by Equation (3.21.2), find the absorption length of electromagnetic waves at 100 MHz and 100 kHz.
4. Randomly polarised radiation at a wavelength of 3 cm is incident on a plane water surface at an angle of 83° to the normal. Calculate the Stokes vector, and hence polarisation state, of the reflected radiation. The dielectric constant of water at 3 cm is 63.1 – 32.1*i*.
5. (For mathematical enthusiasts.) Show that the diffuse albedo for specular reflection of perpendicularly polarised radiation is $\frac{3n^2 - 2n - 1}{3(n+1)^2}$ and confirm that this expression has the correct limiting forms as the refractive index n tends to 1 and to infinity.
6. A rough surface has a BRDF proportional to $\cos(\theta_0)\cos(\theta_1)$, where θ_0 and θ_1 are respectively the angles of the incident and scattered radiation measured from the surface normal. The BRDF has no azimuthal dependence. Show that, if the albedo for normally incident radiation is 1, the diffuse albedo of the surface is 2/3.
7. Consider scattering from a surface with r.m.s. height variation of $\lambda/2$, where λ is the wavelength. Show that the *stationary phase model* can describe this scattering provided that the incidence angle is less than about 60° and $m < 0.60$, whereas the *scalar model* requires $m < 0.25$, where m is the r.m.s. surface slope variation.

8. Verify the solution given in the text for the two-stream scattering model.
9. A typical pack of freshly fallen snow has a density of 100 kg m^{-3} and consists of ice crystals which can be modelled as spheres of radius 0.5 mm. Show that if the snow pack is sufficiently deep it would be expected to have a visible-band albedo close to 1, and estimate the depth necessary for this to occur. The absorption coefficient for light in ice can be taken as 0.01 m^{-1} , and the density of pure ice is 920 kg m^{-3} .
10. Prove Equation 3.29.
11. According to the Minnaert model of surface scattering, the BRDF R is given by

$$R = A(\cos\theta_0 \cos\theta_1)^{a-1}$$

where A and a are constants; θ_0 and θ_1 are the angles between the incident and scattered rays, respectively, and the surface normal; and the BRDF has no azimuthal dependence. Show that the reflectivity r at angle θ_0 , and the diffuse albedo r_d , are related through

$$r = \frac{r_d(a+1)}{2} (\cos\theta_0)^{a-1}$$

in this model.

12. If scattering is insignificant, show that the radiative transfer equation may be written as

$$\frac{dL_f}{dz} = \gamma_a \left(\frac{2hf^3}{c^2(e^{hf/kT} - 1)} - L_f \right),$$

and if $hf = kT$ this can be rewritten as

$$\frac{dT_b}{dz} = \gamma_a(T - T_b).$$

By comparing these two equations, show that

$$T_b = \frac{c^2 L_f}{2hf^2}.$$

Explain its significance.

13. The absorption cross-section at 20 GHz of a water droplet with a radius of $48 \mu\text{m}$ is about $2 \times 10^{-11} \text{ m}^2$. Hence calculate the absorption coefficient of a cloud consisting of droplets of radius $48 \mu\text{m}$ with a number density of 2 million droplets per cubic metre. Would you expect this cloud to be opaque?
14. Show that Doppler broadening will dominate over pressure broadening of a spectral line for a gas at temperature T and pressure p provided that the wavelength of the spectral line is less than $kT/p\sigma$, where σ is the collision cross-section defined in Equation (3.84).

4

Interaction of electromagnetic radiation with the Earth's atmosphere

In Chapter 3 we discussed principally the interaction of electromagnetic radiation with the surface and bulk of the material being sensed. However, the radiation also has to make at least one journey through at least part of the Earth's atmosphere, and two such journeys in the case of systems that detect reflected radiation, whether artificial or naturally occurring. Each time radiation passes through the atmosphere it is attenuated to some extent. In addition, as we have already seen in Section 3.1.2 and Figure 3.5, the atmosphere has a refractive index that differs from unity so that radiation travels through it at a speed different from the free-space speed of $299\,792\,458\text{ m s}^{-1}$. These phenomena must be considered if the results of a remotely sensed measurement are to be corrected for the effects of atmospheric propagation, or if they are to be used to infer the properties of the atmosphere itself. We have already considered them in general terms in discussing the radiative transfer equation (Section 3.4). In this chapter we shall relate them more directly to the constituents of the atmosphere.

4.1

Composition and structure of the gaseous atmosphere

At sea level, the principal constituents of the dry atmosphere are molecules of nitrogen (about 78% by volume), oxygen (21%) and the inert gas argon (1%). There is also a significant but variable (typically 0.1% to 3%) amount of water vapour, often specified by the *relative humidity* H . This is defined by Equation (4.1)

$$H = \frac{p_{\text{water}}}{p_{\text{sat}}(T)} \quad (4.1)$$

as a fraction between zero and 1 (or more commonly as a percentage), where p_{water} is the *partial pressure* of the water vapour, which can be defined as the product of the total atmospheric pressure with the volume fraction of water vapour, and $p_{\text{sat}}(T)$ is the saturated vapour pressure of water at temperature T . Figure 4.1 shows the variation of the saturated vapour pressure of water with temperature, estimated using Antoine's equation (4.2).

$$\log_{10} p_{\text{sat}} \approx A - \frac{B}{T + C}. \quad (4.2)$$

If T is measured in K and p in Pa, the values $A = 10.196$, $B = 1730.6$ and $C = -39.72$ give an estimate that is accurate to better than 1% over the range from 273.15 K to

4.1 Composition and structure of the gaseous atmosphere

Table 4.1. Gaseous constituents of the Earth's atmosphere. The third column shows the fraction by volume of the gas at sea level, and the fourth column the total mass of gas found in a column through the entire atmosphere

Gas	Chemical formula	Volume fraction	Total mass (kg m^{-2})
Nitrogen	N_2	0.7808	8910
Oxygen	O_2	0.2095	2093
Argon	Ar	9.34×10^{-3}	133
Carbon dioxide	CO_2	3.9×10^{-4}	6.4
Neon	Ne	1.8×10^{-5}	0.13
Helium	He	5.2×10^{-6}	7.7×10^{-3}
Methane	CH_4	1.8×10^{-6}	9.9×10^{-3}
Krypton	Kr	1.1×10^{-6}	3.4×10^{-2}
Carbon monoxide	CO	$0.06\text{--}1 \times 10^{-6}$	$0.6\text{--}11 \times 10^{-3}$
Sulphur dioxide	SO_2	1.0×10^{-6}	2.6×10^{-2}
Hydrogen	H_2	5.0×10^{-7}	4.0×10^{-4}
Ozone	O_3	$0.01\text{--}1 \times 10^{-6}$	5.3×10^{-3}
Nitrous oxide	N_2O	2.7×10^{-7}	4.0×10^{-3}
Xenon	Xe	9.0×10^{-8}	4.0×10^{-3}
Nitric oxide	NO_2	$0.05\text{--}2 \times 10^{-8}$	$0.1\text{--}40 \times 10^{-5}$
Total dry atmosphere		1	1.114×10^4
Water vapour	H_2O	0.001–0.028	6.5–180

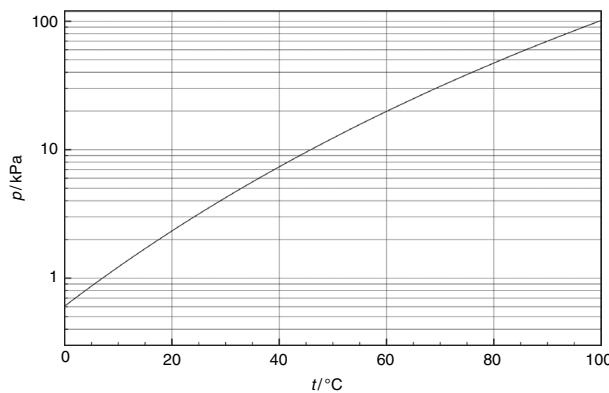


Figure 4.1. Temperature-dependence of the saturated vapour pressure of water.

373.15 K. As an example, at 293.15 K (20 °C) $p_{\text{sat}} = 2.34$ kPa and Antoine's equation estimates it as 2.33 kPa. If the total atmospheric pressure is 100 kPa and the relative humidity is 80%, the volume fraction of water vapour is 1.9%.

In addition to the gases already mentioned, the atmosphere contains a regionally variable quantity of carbon dioxide (currently about 0.039% by volume) and traces, measured in parts per million, of many other gases. Table 4.1 summarises the normal gaseous composition of the atmosphere.

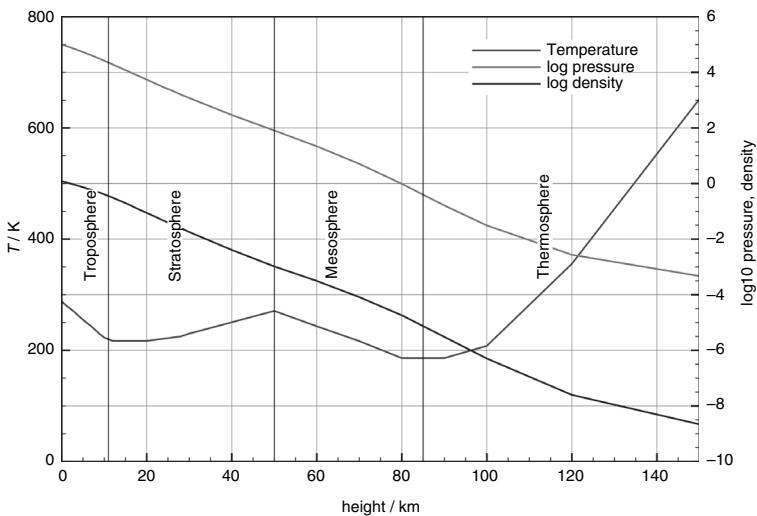


Figure 4.2. Variation with height of temperature, pressure and density of the standard atmosphere.

Atmospheric pressure and density diminish with height above the Earth's surface. This is because the molecules, acted upon by gravity, tend to sink to the surface but are prevented from doing so fully by thermal excitation. The distribution of density with height is thus governed by the Boltzmann distribution and is approximately exponential. There are, however, significant variations from this approximate dependence and it is conventional to divide the atmosphere into several layers. The lowest of these is the *troposphere* (approximately 0–11 km above the Earth's surface), in which the temperature decreases with height, overlain by the *stratosphere* (11–50 km) in which the temperature is roughly constant up to about 35 km, then increases with height, the *mesosphere* (50–80 km) and the *thermosphere* (above about 80 km). The height ranges just specified are for typical conditions at temperate latitudes: there is considerable seasonal and latitudinal variation. Figure 4.2 shows the variation of temperature, pressure and density of the standard (mid-latitude) atmosphere with height.

The absolute temperature T , pressure p and density ρ of the atmosphere can be assumed to be related to one another by

$$\frac{\rho T}{p} = \frac{m_m}{R} \quad (4.3)$$

where m_m is the mass of one mole of atmospheric gas and R is the gas constant (8.314 J K^{-1}). This equation is based on the assumption that the atmosphere behaves as an ideal gas. For heights up to about 100 km, m_m has a more or less constant value of about 0.02896 kg although at greater heights the value of m_m is lower.

The atmospheric pressure p at a height z is a measure of the mass of air, and hence number of molecules, above z . This follows because the Earth's gravitational field strength g may be assumed to be constant over the range of heights for which p is significant, so that

$$p(z) = m_m g \int_z^{\infty} N(z') dz', \quad (4.4)$$

where $N(z')$ is the molar concentration (number of moles per unit volume) at height z' . For example, Figure 4.2 shows that the pressure falls to half its sea-level value at about 5.5 km, and to 1% at about 31 km, so we may state that half of the atmosphere, by mass, is found below 5.5 km and 99% below 31 km. (The troposphere contains about three quarters of the atmospheric mass, and most of the weather. Most aircraft fly within the troposphere, or close to its upper boundary, which is called the *tropopause*.) A spaceborne remote sensing system thus looks through all of the atmosphere, while an airborne system typically looks through only a quarter of it. The variation of pressure with height can be modelled rather crudely as an exponential

$$p(z) \approx p_0 \exp(-z/z_0), \quad (4.5)$$

equivalent to assuming that the logarithmic pressure graph in Figure 4.2 is a straight line, which will be convenient for considering the behaviour of atmospheric sounding systems. For the troposphere, a value of z_0 (the *scale height*) of roughly 7.5 km is appropriate. Equation 4.5 is based on the assumption that the atmosphere is isothermal, but as Figure 4.2 shows, this is not the case. In the troposphere, the temperature decreases uniformly with increasing height, at a *lapse rate* of about -6.5 K per kilometre. In this case, a more accurate expression for the variation of pressure with height is

$$p = (z) = p_0 \left(\frac{T}{T_0} \right)^{5.25}, \quad (4.6)$$

where p_0 and T_0 are the pressure and temperature at height zero, and T is the temperature at height z (which can be calculated from T_0 and the lapse rate). The density can be obtained from Equation (4.3).

Graphs similar to Figure 4.2 can also be drawn for the partial pressures of the individual molecular species listed in Table 4.1 that compose the atmosphere. Provided that we do not consider altitudes much above 100 km, where the heavier molecules are less favoured (this is the reason for the decrease in the mean molar mass m_m at greater altitudes), and that the gas is ‘well mixed’, the graph will follow the shape of Figure 4.2 fairly accurately. Most of the atmospheric gases are well mixed in this sense, with the notable exception of ozone which, as well as being found in the troposphere, is also generated (by the action of solar ultraviolet radiation) in the stratosphere. Satellite measurements of the kind described in Section 6.7.2 have recently shown that carbon dioxide is less well mixed than previously believed (Chahine *et al.* 2008).

Summary

The Earth’s atmosphere is a mixture of gases dominated by nitrogen and oxygen. Water vapour is also a significant component, spatially and temporally variable. The concentration of water vapour can be expressed through the relative humidity. The pressure and density of the atmosphere decrease monotonically with height above the surface, but the temperature varies in a more complicated manner, on the basis of which the atmosphere

can be divided into a number of zones. The most important of these are the troposphere (from 0 to typically 11 km above the surface), containing about three-quarters of the mass of the atmosphere and in which most meteorological phenomena occur, and the stratosphere (11–50 km).

4.2

Molecular absorption and scattering in the atmosphere

In Section 3.4 we considered the physical principles underlying the absorption of electromagnetic radiation by molecules. Now we can consider the effect that molecular species have on atmospheric propagation. Figure 4.3 shows the typical atmospheric transmittance between wavelengths of 250 nm and 2500 nm, i.e. in the ultraviolet, visible and near-infrared regions of the spectrum. In this part of the spectrum absorption is dominated by water vapour, although methane, carbon dioxide and molecular oxygen are also responsible for a few absorption lines. The behaviour in the visible region is dominated by molecular Rayleigh scattering. At the short-wavelength end of the spectrum, in the ultraviolet, absorption by ozone becomes very significant.

We noted in Chapter 3 that the cross-section of a molecule for Rayleigh scattering is given by

$$\sigma_s = \frac{128\pi^5 a^6}{3\lambda^4}, \quad (3.85)$$

where a is the effective radius of the molecule (which is modelled crudely as a conducting sphere) and λ is the wavelength. We can take the radius of an ‘air molecule’ as around 0.2 nm,

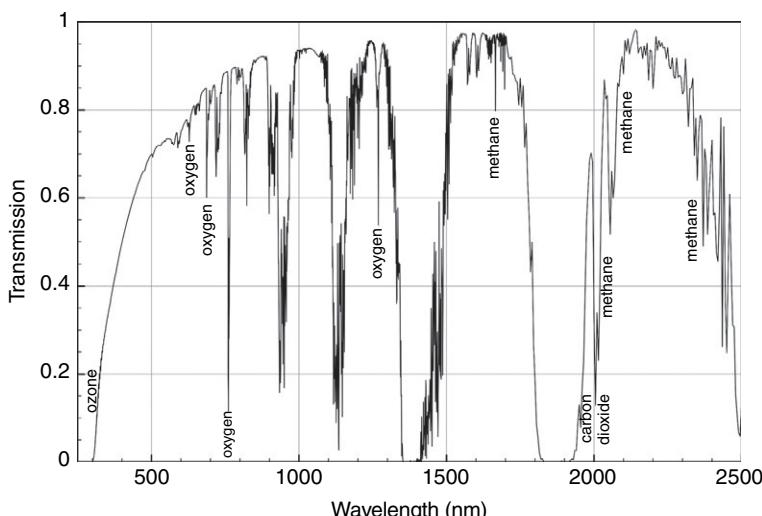


Figure 4.3. Typical zenith atmospheric transmittance between 250 nm and 2500 nm. The main absorption features are labelled, with the exception of water vapour, which is responsible for most of the unlabelled features.

4.2 Molecular absorption and scattering in the atmosphere

so the scattering cross-section is of the order of 10^{-30} m^2 for red light and an order of magnitude larger for blue light. Dividing the total atmospheric mass per unit area from Table 4.1 ($1.11 \times 10^4 \text{ kg m}^{-2}$) by the mean molecular mass ($4.81 \times 10^{-26} \text{ kg}$), we find that the integral of the number density of atmospheric molecules through the whole atmosphere is $2.3 \times 10^{29} \text{ m}^{-2}$. Thus, the optical thickness of the atmosphere for scattering is of the order of 0.1 for red light and 1 for blue light. This fact has a number of important consequences. Most famously, it explains why the clear sky is blue. The fact that we can see the sky at all tells us that it scatters the light from the Sun (since the molecules of which the sky is composed do not themselves emit light, they must scatter it towards our eyes). Since the scattering is an order of magnitude more efficient for blue light than for red light, the scattered light is very much bluer than the incident light. Another way of thinking about the same phenomenon is to recognise that air, when illuminated by visible light, becomes luminous and blue. This light, called *airlight* or *skylight*, explains why shadows are not completely dark. (On the Moon, which has no atmosphere, shadows are completely dark.) It also explains why we should expect a downward-viewing sensor to detect a non-zero radiance from a non-reflecting object, and why this effect will be greater for shorter wavelengths. The explanation of the blue sky also explains why the Sun or Moon, when seen close to the horizon from a point on the Earth's surface, appears much redder than usual. All objects observed through the atmosphere appear redder than they would if the atmosphere were not present, as a result of the preferential

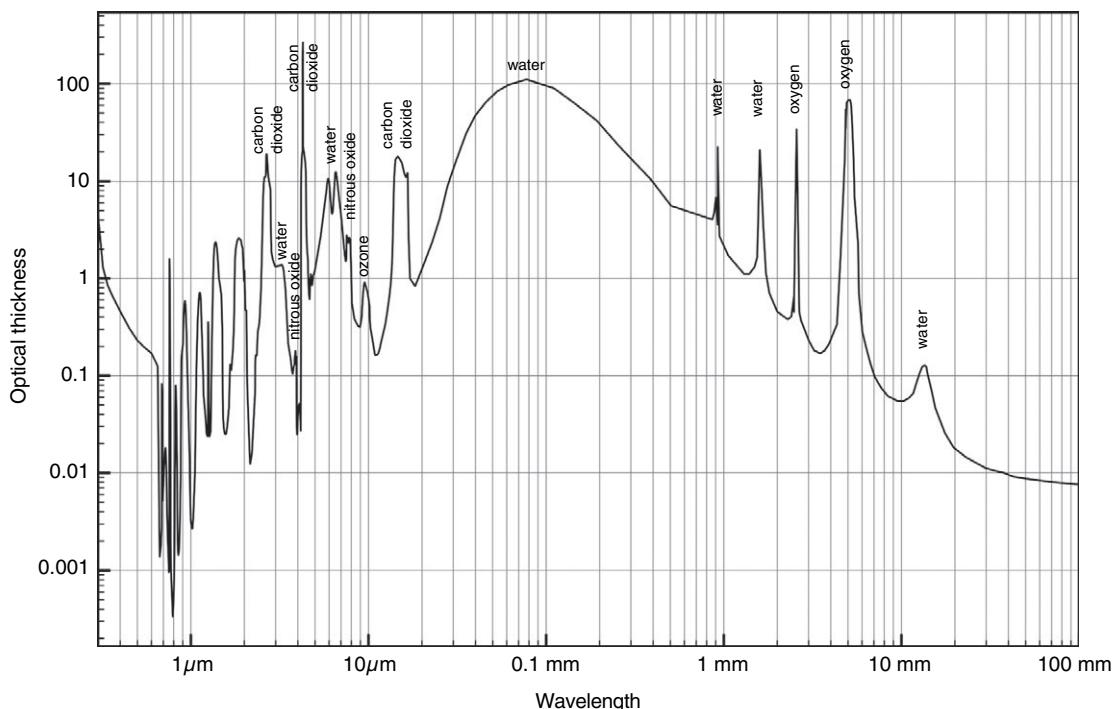


Figure 4.4. Schematic typical optical thickness of the clear atmosphere at the zenith. The main absorption peaks are identified for wavelengths longer than $2 \mu\text{m}$.

scattering of blue light. When an object is seen close to the horizon the path through the atmosphere is much longer and the effect is greater.

Figure 4.4 shows the typical atmospheric transmittance for wavelengths between 300 nm and 10 cm, i.e. covering the whole of the electromagnetic spectrum useful for remote sensing. The figure is rather schematic, since it does not resolve fine spectral details and assumes 'standard' atmospheric conditions. (An alternative presentation of the information in Figure 4.5 is given in Figure 1.5.) The comparative transparency of the optical region and of the microwave region below the 60 GHz oxygen absorption line is clearly apparent, as is the complexity of the infrared region as a result of the large number of molecular transitions. The comparatively transparent regions (optical thickness less than about 1) define the so-called *atmospheric windows*. Roughly speaking, there are two main windows: in the optical–infrared region and in the microwave region. These can, however, be usefully subdivided by the presence of absorption lines.

The general form of Figure 4.4 is dominated by the absorption spectrum of water, with most of the fine structure confined to the visible and infrared region (shown in more detail in Figure 4.3). The increasing optical thickness at shorter ultraviolet wavelengths is caused by Rayleigh scattering, and also by absorption by ozone molecules.

As Figure 4.4 and Table 4.1 both suggest, the presence of water vapour has a very significant effect on remote sensing observations. On average, the total amount of water vapour in the atmosphere is around 1.3×10^{16} kg, which is an average of 25 kg per square

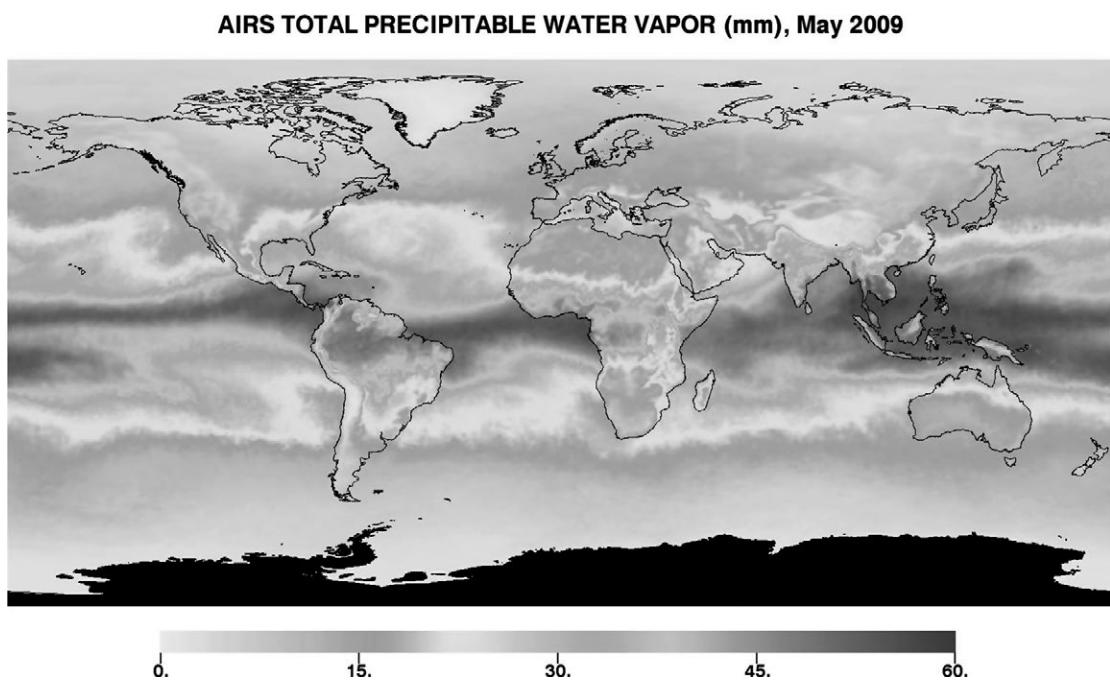


Figure 4.5. Total precipitable water in May 2009, derived from AIRS satellite data. (Figure reproduced by courtesy of NASA. Source: NASA JPL <http://photojournal.jpl.nasa.gov/catalog/PIA12097>.) See also colour plates section.

4.2 Molecular absorption and scattering in the atmosphere

metre of the Earth's surface. This figure is the mean *column integral* of the density of water vapour, i.e. the density integrated over a vertical path through the atmosphere. Another way of specifying the same quantity is through the *total precipitable water* (TPW), which is the depth of the column of liquid water that would be produced by condensing all of the vapour. This is found simply by dividing the column integral of the vapour density by the density of liquid water, so we can say that the global average TPW is around 25 mm. The distribution of water vapour is not uniform over the Earth's surface or over time. Figure 4.5 shows the spatial variation of the TPW for the month of May 2009, derived from satellite remote sensing data. The principle of measurement is discussed in Chapter 7. As Figure 4.5 shows, water vapour is found mostly close to the equator, where there is particularly strong atmospheric convection.

Figure 4.6 shows the optical thickness of the atmosphere for a vertically propagating ray. If the ray travels obliquely through the atmosphere, however, it is clear that it will encounter a larger number of molecules, so we would expect the optical thickness to be increased. We can derive a very simple model of this phenomenon by assuming that the Earth is flat, and considering a ray that makes an angle ϕ to the vertical (this is called the *zenith angle* of the ray). The optical thickness is given by

$$\tau = \int_0^\infty \gamma(x) dz,$$

where $\gamma(x)$ is the attenuation coefficient at distance x along the ray. We know that x is related to the height z in the atmosphere by

$$x = \frac{z}{\cos\phi},$$

so we can rewrite our expression for the optical thickness as

$$\tau = \frac{1}{\cos\phi} \int_0^\infty \gamma(z) dz. \quad (4.7)$$

In other words, the optical thickness is increased by a factor of $1/\cos\phi$ with respect to its value for a vertical ray. This very simple formula is sufficiently accurate for ϕ less than about 75° , i.e. for factors up to about 4. For larger zenith angles, it is necessary to take the Earth's curvature into account. (This is obvious, because the $1/\cos\phi$ formula implies that the optical thickness is infinite for a horizontal ray, whereas we know that such a ray will, because of the Earth's curvature, emerge from the atmosphere and hence propagate within it for a finite, and not infinite, distance.) The required correction depends on the distribution of the attenuation coefficient with height, which we can model as an exponential decay

$$\gamma(z) = \gamma_0 \exp(-z/z_0),$$

where z_0 is the scale height. Note that this is analogous to the simple model of the distribution of atmospheric pressure with height that we introduced in Equation (4.4), and

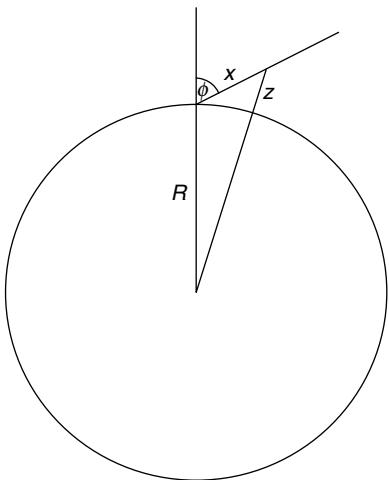


Figure 4.6. Geometry of an oblique ray above a spherical Earth.

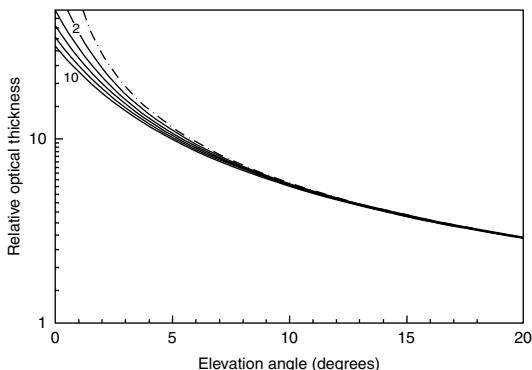


Figure 4.7. The factor by which the optical thickness of the atmosphere is increased for oblique propagation relative to vertical propagation. The dashed curve is the ‘flat-Earth’ result and the solid curves are calculated for a spherical Earth, labelled according to the exponential scale height (in km) of the absorbing material.

that we expect that for most atmospheric species, if they are well mixed in the atmosphere, the appropriate value of z_0 will be about 8 km. We can calculate the effect of an oblique path through the atmosphere as follows. We assume that the Earth is spherical with a radius R , and consider an oblique path making an angle φ with the vertical. Simple trigonometry (Figure 4.6) shows that the relationship between the distance x along the oblique path and its height z above the Earth’s surface can be written as

$$z^2 + 2Rz - x^2 - 2Rx \cos\varphi = 0,$$

so the solution for z is

4.3 Particles in the atmosphere: aerosols

$$z = (R^2 + 2Rx \cos\phi + x^2)^{1/2} - R,$$

and the optical length of the path is

$$\tau = \gamma_0 \int_0^\infty \exp\left(\frac{R - (R^2 + 2Rx \cos\phi + x^2)^{1/2}}{z_0}\right) dx. \quad (4.8)$$

This can be evaluated numerically. Figure 4.7 shows the factor by which the optical thickness for oblique propagation is increased relative to vertical propagation, for various values of the scale height z_0 . The figure is plotted as a function of the ray's elevation angle θ , which is related to the zenith angle ϕ by $\theta + \phi = 90^\circ$. The figure is plotted for values of the exponential scale height of 2, 4, 6, 8 and 10 km, and for elevation angles up to 20° . For larger elevation angles the flat-earth result is sufficiently accurate.

Summary

Transmission of electromagnetic radiation through the clear atmosphere is strongly influenced by molecular (Rayleigh) scattering in the ultraviolet and blue parts of the spectrum. This gives rise to the phenomena of the blue sky and red sunsets, and means that visible-wavelength observations of the Earth's surface will also contain a contribution from the atmosphere itself, especially at the shortest wavelengths. Molecular absorption also strongly influences atmospheric propagation, with many absorption lines in the visible and near-infrared regions, fewer in the thermal infrared and a small number in the microwave region. Water vapour plays a particularly important role (it is responsible for the fact that the atmosphere is highly opaque between about $30 \mu\text{m}$ and $300 \mu\text{m}$, separating the thermal infrared and microwave regions of the spectrum), but other significant molecules are oxygen, ozone, carbon dioxide, methane and nitrous oxide. The total amount of water vapour integrated through the atmosphere is expressed through the quantity of *total precipitable water*. The effect of an absorbing species on propagation through the atmosphere is increased when the path is not vertical.

4.3

Particles in the atmosphere: aerosols

In the preceding sections we have considered the gaseous composition of the atmosphere. However, it also contains solid and liquid components that can have a significant effect on the propagation of electromagnetic radiation. These components generally exhibit much greater spatial and temporal variability than the gaseous components, so they are rather harder to characterise. In this section we discuss aerosols, and in the following section the larger ice and water particles that compose fog, cloud, rain and snow.

Aerosols are suspensions of very small solid particles or liquid droplets, with radii typically in the range 10 nm to $10 \mu\text{m}$. They can be regarded as being suspended in the atmosphere since, because of their small size, the speed at which they fall under gravity is very small. For example, a particle of radius $1 \mu\text{m}$ will fall at something like 10^{-4} m s^{-1} ,

which means that under conditions of absolutely still air it would take of the order of 100 days to fall through 1 km. This definition of an aerosol includes fog and clouds, although we defer their discussion to the next section.

Most aerosols are found in the atmospheric *boundary layer*, the lowest kilometre or so of the atmosphere in which transport processes are dominated by wind turbulence and by atmospheric convection. (Some aerosols are, however, found in the stratosphere.) This indicates that the aerosols are largely generated from the Earth's surface, for example in liberating and then carrying aloft solid microscopic dust particles from the land surface and water droplets from the ocean surface. Consequently, the type and size distribution of aerosols is strongly dependent on the local meteorological conditions and the local nature of the Earth's surface – whether urban, rural, oceanic, volcanic etc. Aerosols compose atmospheric haze, and common names for the different types of aerosol include smoke, dust, ash and soot. Over the ocean surface, spray is also a source of aerosols. Most aerosols are of natural origin although around 10% are a result of human activity.

For visible-wavelength radiation (say a wavelength of $0.55\text{ }\mu\text{m}$), the attenuation coefficient at sea level ranges typically between 0.05 and 0.5 km^{-1} , and the total (vertical) optical thickness of a tropospheric aerosol is typically 0.1 to 1 . If we compare this range of values with the gaseous optical thickness at wavelength $0.55\text{ }\mu\text{m}$ of about 0.2 , as shown in Figure 4.4, we can see that the aerosol component of the atmosphere is radiatively significant. The distribution of atmospheric aerosols varies spatially and over time. Figure 4.8 shows

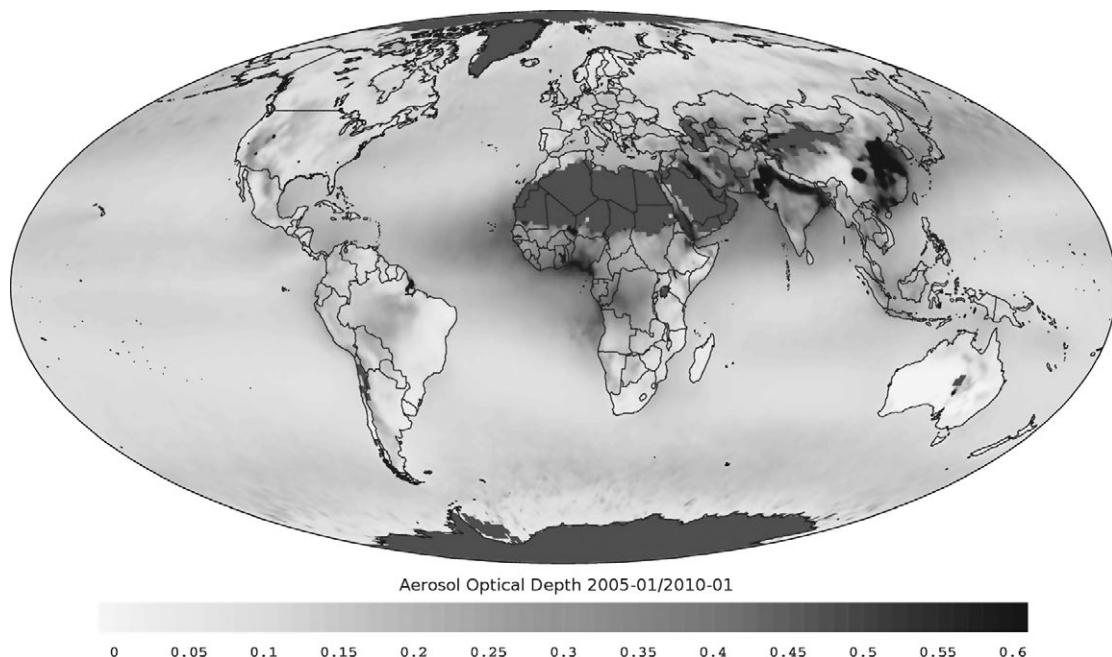


Figure 4.8. Average aerosol optical thickness, 2005–2010, derived from MODIS satellite imagery. No data could be retrieved from the grey areas, as a result of the high reflectance of the earth's surface in these areas. (Source: NASA via Wikipedia. http://en.wikipedia.org/wiki/File:Modis_aerosol_optical_depth.png.) See also colour plates section.

4.4 Fog and cloud

the long-term (5 years) time-average optical thickness of aerosols at wavelength 0.55 µm derived from MODIS satellite imagery.

The attenuation coefficient of an aerosol at optical and infrared wavelengths is dominated by scattering rather than absorption. The dependence of the attenuation coefficient on wavelength is usually represented approximately by the *Ångström relation*

$$\gamma = \gamma_0 \lambda^{-n}, \quad (4.9)$$

where γ_0 is a constant and n is called the *Ångström exponent*. For particles very small compared with the wavelength, the Rayleigh scattering formula implies that $n = 4$, and for very large particles, the fact that the scattering cross-section is of the order of the geometrical cross-section implies that $n = 0$. For aerosols, the Ångström exponent n is usually between 0.2 and 2. Maritime aerosols usually exhibit the largest particle sizes (the modal droplet radius is around 0.2 µm) and the smallest values of n . For most other aerosols, the value of n is around 1, although at high altitudes (stratospheric and higher altitude tropospheric aerosols) the very small particle sizes give $n \approx 2$.

Summary

Aerosols are very small (up to around 10 µm) solid or liquid particles, effectively suspended in the atmosphere because the rate at which they fall through it is very low because of air resistance. They are predominantly of natural origin although anthropogenic aerosols are also significant, and they occur mostly in the atmospheric boundary layer (typically the lowest kilometre or two) although stratospheric aerosols also occur. Their principal effect on electromagnetic radiation is scattering in the ultraviolet, visible and near-infrared region. The dependence of the scattering coefficient on wavelength is expressed through the Ångström exponent, which depends on the size of the particles. The optical thickness due to aerosols is typically 0.5, although it varies spatially and temporally.

4.4

Fog and cloud

At any one time, about two-thirds of the Earth's surface will be covered by cloud (Figures 4.9 and 4.10). Visible-wavelength sensors, and to some extent those operating in other regions of the electromagnetic spectrum, will be limited by the presence of significant amounts of cloud cover, and this can be a serious problem in the case of spaceborne sensors that revisit a particular location comparatively infrequently (see Chapter 10). For example, it has been estimated that a Landsat satellite, which revisits a particular location once every 16 days, will obtain a cloud-free scene of a particular location in Britain only once per year, and a scene with 1 okta of cloud (an okta is one-eighth of the sky obscured by cloud) only twice per year. The probability of less than 10% cloud cover in a single Landsat observation over the continental USA is about 0.05 to 0.4, and the number of observations needed to obtain a probability of 75% of less than 10% cloud cover is between 5 and 100 (Goetz 1979). In high latitude regions, the problems of cloud cover can be significantly worse (Marshall, Dowdeswell and Rees 1994).

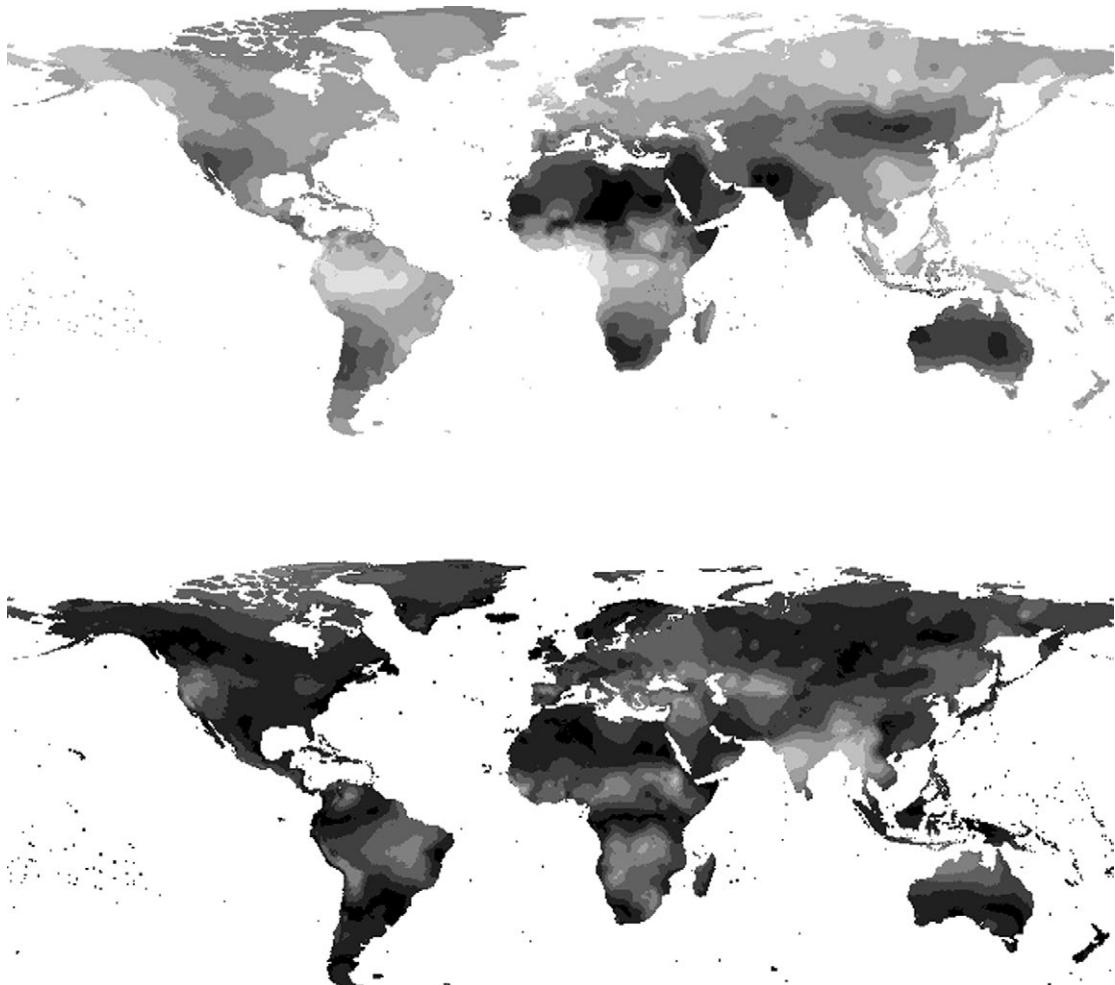


Figure 4.9. Top: average cloud cover over land; bottom: average monthly variability of cloud cover. The darkest shading represents an average cloud cover of less than 20% and a variability of less than 4%, while the lightest shading represents an average cloud cover of 80% or more and a variability of 24% or more. The figures have been calculated from data compiled from terrestrial monitoring stations by the Climatic Research Centre, University of East Anglia, UK; (CRU TS 2: (Mitchell and Jones 2005).)

Fog and low altitude (up to about 3000 m) cloud consists of water droplets, but these are very much larger than the droplets in aerosols, having modal radii from 10 to 50 μm . At visible and infrared wavelengths, scattering is again dominant. We can model the scattering fairly crudely by assuming that all the droplets have the same radius a and that their scattering cross-section is just the geometrical area πa^2 . If the number density of the droplets (the number of droplets per unit volume) is N , the scattering coefficient is therefore given by

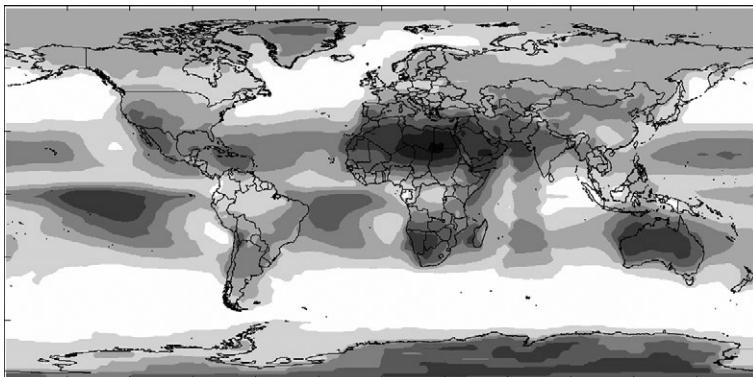


Figure 4.10. Global mean cloud cover according to the International Satellite Cloud Climatology Programme (ISCCP). The darkest shade of grey corresponds to cloud cover up to 20%, while the lightest shade corresponds to cloud cover over 80%. (<http://isccp.giss.nasa.gov>)

$$\gamma_s \approx \pi a^2 N.$$

The mass density of liquid water in the fog or cloud is given by

$$\rho = \frac{4\pi a^3 N \rho_w}{3}, \quad (4.10)$$

where ρ_w is the density of water, and we have assumed that the droplets are spherical, so we see that we can write the scattering coefficient as

$$\gamma_s \approx \frac{3\rho}{4a\rho_w}. \quad (4.11)$$

Since the droplet radius is not strongly dependent on the mass density ρ , this equation implies that the scattering coefficient is roughly proportional to the mass density of liquid water in the cloud. This ranges from about $10^{-4} \text{ kg m}^{-3}$ for fog and thin cloud, to about $10^{-2} \text{ kg m}^{-3}$ for the densest clouds, so we should expect scattering coefficients in the region of 1 to 100 km^{-1} . Cloud layers are typically of the order of 1 km thick, so all but the thinnest clouds (and fog layers, which are typically 50 m thick) are optically opaque, with optical thicknesses in the range 1–100. In fact, the International Satellite Cloud Climatology Project uses the optical thickness as one of the criteria for classifying clouds from satellite data, with three optical thickness classes separated by values of the optical thickness of 2.3 and 23 (Hahn, Rossow and Warren 2001).

Similar remarks apply to the higher altitude clouds that are composed of ice crystals. The crystals are of a similar size and density to the water droplets, and the calculation on which Equation (4.11) is based is too crude to incorporate the fact that the ice crystals are not spherical.

We saw in Section 3.6.4 that absorption dominates hugely over scattering for cloud droplets at microwave frequencies. Putting together Equations (3.66) and (3.67), we see that the absorption cross-section of a small sphere of radius a and dielectric constant $\epsilon' - ie''$ is given by

$$\sigma_a = \frac{12\pi a^3 k \epsilon''}{(\epsilon' + 2)^2 + \epsilon''^2},$$

where k is the wavenumber of the radiation. The dielectric constant of water in the microwave region can be described quite accurately using the Debye equation (3.21), and if we substitute this equation into our expression for σ_a we obtain

$$\sigma_a = \frac{12\pi a^3 k (1 + \omega^2 \tau^2) \omega \tau \epsilon_p}{((\epsilon_\infty + 2)(1 + \omega^2 \tau^2) + \epsilon_p)^2 + \omega^2 \tau^2 \epsilon_p^2} \quad (4.12)$$

for the scattering cross-section at angular frequency ω . Making the substitutions $\omega = ck = 2\pi f$, where f is the frequency, and approximating to the low frequency limit $\omega\tau \ll 1$, we find

$$\sigma_a \approx \frac{48\pi^3 \tau \epsilon_p}{(\epsilon_\infty + \epsilon_p + 2)^2 c} a^3 f^2. \quad (4.13)$$

Although this expression appears complicated, the right-hand side is just a constant multiplied by $a^3 f^2$. Multiplying this by the number density N of the water droplets to obtain the absorption coefficient, and making use of Equation (4.10), we see that this simplified (low frequency) model predicts that the absorption coefficient should be proportional to ρf^2 , where ρ is the mass density of liquid water in the cloud. In fact, a somewhat more accurate approximation over the usual range of microwave frequencies is

$$\gamma_a \approx 0.6 \left(\frac{\rho}{kg\ m^{-3}} \right) \left(\frac{f}{GHz} \right)^{1.9} dB\ km^{-1}. \quad (4.14)$$

Thus we see that even a thick layer of dense cloud is comparatively transparent to microwave radiation, introducing only of the order of 1 dB attenuation at 50 GHz. At frequencies below about 15 GHz, absorption by cloud is clearly negligible.

Summary

Fog and cloud are also aerosols, since they consist of atmospheric suspensions of small water droplets (or ice particles in the case of high altitude clouds). Typical sizes range from 10 µm to 50 µm. Cloud cover is spatially and temporally highly variable, though at any moment of time typically three-quarters of the Earth's surface is covered by cloud. The effect of clouds on visible-wavelength and infrared radiation is dominated by scattering, and optical thicknesses of up to 100 are possible, so that many clouds are extremely opaque. This imposes a major limitation on spaceborne observation of the Earth's surface at these wavelengths. In the microwave region, attenuation is mainly caused by absorption, so that the effect of a cloud at a particular frequency is proportional to the mass of water in the cloud. The absorption coefficient is roughly proportional to the square of the frequency. At frequencies below about 15 GHz absorption by clouds is negligible, while at frequencies above about 50 GHz the thickest clouds can begin to have an appreciable effect.

4.5

Rain and snow

Precipitation can also have a significant effect on the propagation of electromagnetic radiation. The attenuation of electromagnetic radiation by these meteorological phenomena can be considered in both a positive and a negative light. For observations intended to characterise the Earth's surface, it is an inconvenience since it must be corrected for, or may even render the observation impossible. On the other hand, it may be possible to derive useful information about the meteorological phenomenon itself from the effect on radiation, for example in microwave rain radars.

In the case of rainfall, the rain rate is the dominant factor since this largely controls both the size distribution of the droplets and their number density. Rain is liquid precipitation, i.e. the condensation of atmospheric water vapour into drops that are large enough to fall under gravity. A 'typical' raindrop is a few millimetres in diameter (and hence roughly 100 times as large as the droplets that compose clouds), and has a settling velocity of a few metres per second. The global average rainfall is around $5.18 \times 10^{18} \text{ m}^3$ per annum, equivalent to a column of water 1.02 m deep averaged over the whole of the Earth's surface. This means that the global average annual rainfall is 1.02 m, although the figure is somewhat higher over oceans (1.14 m) than over the land surface (0.72 m).

Rainfall is a very dynamic process, as experience tells us and as can also be seen by comparing the global annual rainfall of 1.02 m with the average total precipitable water of 25 mm. The ratio of the latter figure to the former has a value of about 0.025 years or 9 days, and this is the typical *residence time* of water in the atmosphere. Some parts of the world show very strong seasonality in the rainfall. Rainfall is also distributed very inhomogeneously over the Earth's surface (Figure 4.11).

At visible and infrared wavelengths, scattering is dominant over absorption, and since the particle size is very much larger than the wavelength the scattering cross-section of a droplet is of the order of the geometrical cross-section. The angular distribution of the scattered radiation is, however, quite complicated – this fact is obvious from the existence of various rainbow phenomena, as we have already noted.

The radii of the droplets in a particular rain shower are in fact distributed over a rather large range. We can define a droplet size distribution $N(D)$, such that $N(D)dD$ is the number of droplets per unit volume having diameters between D and $D + dD$, and various empirical forms of this distribution have been defined (e.g. Laws and Parsons 1943, Marshall and Palmer 1948, Joss and Gori 1978)). Assuming that the scattering cross-section of an individual drop is given by $\pi D^2/4$, the scattering coefficient is then

$$\gamma_s = \frac{\pi}{4} \int_0^\infty D^2 N(D) dD. \quad (4.15)$$

The fractional volume occupied by drops having diameters between D and $D + dD$ is

$$f(D) = \frac{\pi}{6} D^3 N(D) dD \quad (4.16)$$

and the total mass of water per unit volume of rainfall is given by

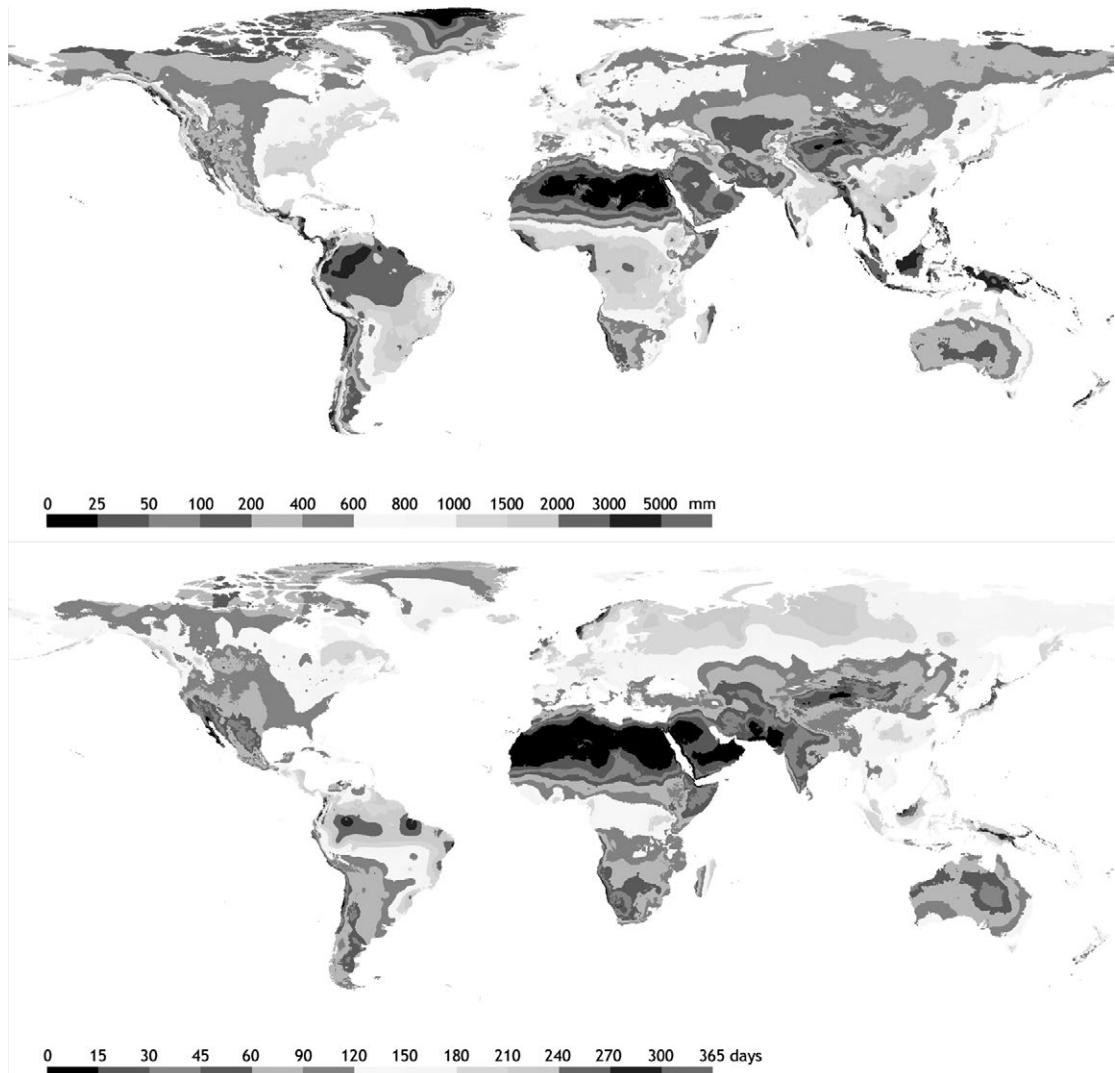


Figure 4.11. Global distribution of average annual rainfall (top) and average number of rainy days (bottom). The figures have been calculated from data compiled from terrestrial monitoring stations by the Climatic Research Centre, University of East Anglia, UK. (CRUTS 2: Mitchell and Jones 2005) See also colour plates section.

$$\rho = \rho_w \int_0^{\infty} f(D) dD, \quad (4.17)$$

where ρ_w is the density of water. The drop size distribution is mainly governed by the *rain rate*, usually specified in millimetres per hour. Table 4.2 illustrates the values calculated from Equations (4.15) and (4.17) for rain rates of 1 mm/hour (a light rainfall) and

Table 4.2. Calculated optical scattering coefficient and mass density of liquid water for two different rain rates

Rain rate (mm/hour)	γ_s (m^{-1})	ρ (kg m^{-3})
1	4.9×10^{-5}	3.3×10^{-5}
100	1.4×10^{-3}	2.2×10^{-3}

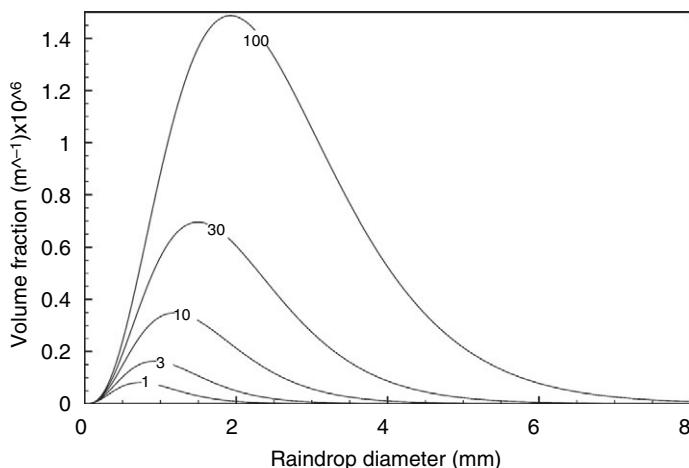


Figure 4.12. Calculated volume fraction of water for rain rates between 1 and 100 mm/hour.

100 mm/hour (a tropical downpour), using the drop size distribution given by Joss and Gori (1978), and Figure 4.12 shows calculated values of $f(D)$.

If we assume that the scattering coefficient is proportional to the rain rate R raised to some constant power, the data of Table 4.3 can be interpolated by Equation (4.18):

$$\gamma_s \approx 4.9 \times 10^{-5} \left(\frac{R}{\text{mm hr}^{-1}} \right)^{0.73} \text{ m}^{-1}. \quad (4.18)$$

Thus we might expect the scattering coefficient in a heavy rain shower ($R = 25 \text{ mm/hour}$) to be of the order of $5 \times 10^{-4} \text{ m}^{-1}$ (about 2 dB km $^{-1}$).

The interactions between microwave radiation and rain are not so easy to calculate, because the droplets are similar in size to the wavelength of the radiation. The dimensionless parameter $\chi = 2\pi a/\lambda$ that determines whether the Rayleigh or Mie scattering formulae should be used can vary over a wide range as a result of the rather broad distribution of droplet sizes (Figure 4.12). Figure 4.13 shows the approximate dependence of the scattering and absorption coefficients on frequency for rain rates of 1 mm/hour and 100 mm/hour, from which it can be seen that (1) both scattering and absorption increase with rain rate, (2) absorption dominates over scattering except at high frequencies and rain rates, and (3) both absorption and scattering can be significant at frequencies above about 10 GHz.

Scattering of microwave radiation by raindrops provides the physical basis for terrestrial *rain radars* (also called *weather radars*). These operate at a typical frequency of

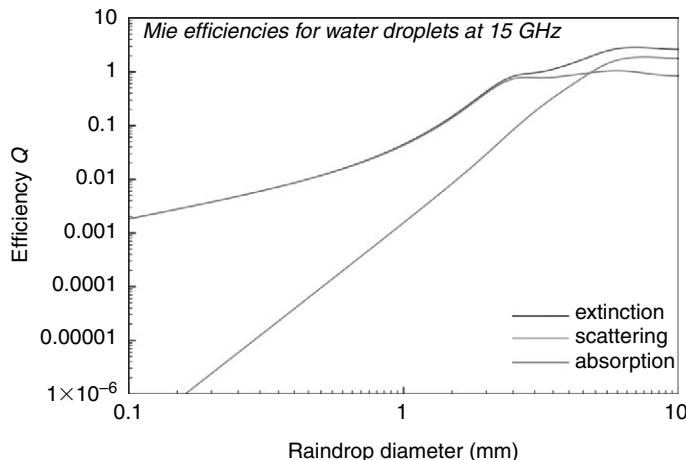


Figure 4.13. Scattering, absorption and extinction efficiencies for spherical water drops at 15 GHz.

5–10 GHz, and are used by meteorological agencies in a number of countries to map the distribution of rain intensity in real time. This technique falls outside our definition of remote sensing since it does not involve a downward view from an airborne or spaceborne platform, but airborne systems do also exist and, since 1997, the Tropical Rainfall Mapping Mission (TRMM) satellite has carried a *precipitation radar* capable of performing a similar function from space.

Summary

Rainfall is liquid precipitation of water in the form of drops with typical sizes of a few millimetres. The average annual precipitation over the Earth's surface is around one metre but the distribution is very non-uniform spatially and temporally. The effect of rain on the propagation of electromagnetic radiation is dominated by scattering in the visible and infrared regions, and generally by absorption in the microwave region although scattering is also significant. The rain rate (intensity of the rainfall) has a strong effect on the scattering and absorption coefficients, and this fact is exploited by terrestrial, airborne and spaceborne rain radars. The effects of snowfall are qualitatively similar to those of rain.

4.6

The ionosphere

The ionosphere is an ionised layer above the Earth's atmosphere, extending from about 70 km to a few hundred kilometres above the surface. The ionisation is produced by extreme ultraviolet and X-radiation from the Sun, and it can have a significant effect on the propagation of radio-frequency electromagnetic radiation.

4.6 The ionosphere

We saw in Chapter 3 that the dielectric constant of a plasma is given by

$$\varepsilon_r = 1 - \frac{Ne^2}{\varepsilon_0 m_e \omega^2}, \quad (3.24)$$

where N is the number density of the electrons, e is the charge and m_e the mass of an electron, and ω is the angular frequency of the radiation. The plasma frequency is

$$\omega_p = \sqrt{\frac{N e^2}{\varepsilon_0 m_e}} \quad (3.25)$$

and the dielectric constant is positive or negative according as ω is greater or less than ω_p . The maximum value of the electron density N in the ionosphere is of the order of 10^{12} m^{-3} , implying that ω_p is about $6 \times 10^7 \text{ s}^{-1}$ ($f_p \approx 9 \text{ MHz}$). Thus, at microwave frequencies (and above) the dielectric constant is positive and very slightly less than 1.

We can use the binomial expansion to approximate the square root of Equation (3.24) and hence obtain the refractive index of a plasma for the case when $\omega \ll \omega_p$:

$$n = \sqrt{\varepsilon_r} \approx 1 - \frac{Ne^2}{2\varepsilon_0 m_e \omega^2}. \quad (4.19)$$

This is purely real, so there is no absorption of radiation. The phase velocity v of electromagnetic waves is given by Equations (3.5) and (3.6) as

$$v = \frac{c}{n},$$

where c is the speed of light *in vacuo*, and in this case it is clearly greater than c . This seems paradoxical, since it appears to contradict Einstein's postulate that the speed of light represents an upper speed limit, until we recall that this speed limit applies to the propagation of *information* and that, as discussed in Section 3.1.3, the information in a wave is propagated at the group velocity and not the phase velocity.

Combining Equations (3.5), (3.6) and (4.21), we may write the dispersion relation for radiation propagating in a plasma at a frequency very much higher than the plasma frequency as

$$\omega = ck \left(1 - \frac{Ne^2}{2\varepsilon_0 m_e \omega^2} \right)^{-1},$$

so we can evaluate the group velocity from Equation (3.27):

$$v_g = \frac{d\omega}{dk} = c \left(1 + \frac{Ne^2}{2\varepsilon_0 m_e \omega^2} \right)^{-1}. \quad (4.20)$$

This is clearly *less* than c , as expected. Furthermore, we can use Equation (4.20) to calculate the time t taken for a pulse of radiation to travel through a finite region (for example, all) of the ionosphere. Since

$$t = \int \frac{dz}{v_g},$$

where z measures propagation distance, we obtain

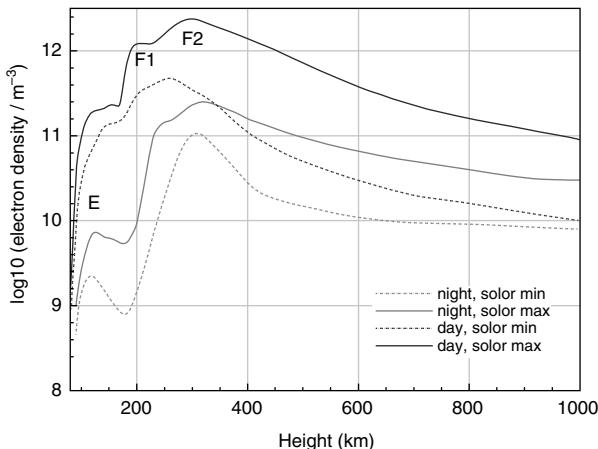


Figure 4.14. Typical electron densities in the ionosphere. The figure also shows the approximate positions of the layers into which the ionosphere is conventionally divided.

$$t = \frac{z}{c} + \frac{e^2}{2\epsilon_0 m_e \omega^2 c} \int N dz. \quad (4.21)$$

The right-hand side of this expression can be interpreted as follows. The first term is just the time taken for light to travel the distance z in *vacuo*, and the second consists of a frequency-dependent constant multiplied by the integrated number density of electrons along the path through the ionosphere.

As an example of the use of Equation (4.21), consider the propagation of a pulse of microwaves at a frequency of 10 GHz vertically through the entire ionosphere. The value of $\int N dz$ in this case might typically be $3 \times 10^{17} \text{ m}^{-2}$, so the second term on the right-hand side of Equation (4.21) has a value of $4.0 \times 10^{-10} \text{ s}$. Thus the pulse is delayed by 0.40 ns, or 0.12 m, compared with a pulse travelling the same distance through free space. A pulse at 5 GHz would be delayed by four times this amount. We shall consider these delays again when we discuss radar altimeters in Chapter 8.

The electron density in the ionosphere is very variable, both temporally and spatially. The ionisation is significantly greater on the Earth's sunlit side than on the night side; it is strongly affected by variations in solar activity; and its spatial distribution is correlated with both altitude and geomagnetic latitude. Figure 4.14 illustrates typical mid-latitude electron densities as functions of altitude for day and night.

As we have seen in deriving Equation (4.21), the integrated electron density $\int N dz$ through the whole of the ionosphere is an important quantity in considering the propagation time of microwave pulses. This quantity is often called the *total electron content* (TEC). It has a typical daytime value of $3 \times 10^{17} \text{ m}^{-2}$, and is usually about ten times less at night.

We remarked earlier that, at frequencies below the plasma frequency, the dielectric constant of the ionosphere is negative. This implies that the refractive index is purely imaginary, and hence that radiation will be absorbed. The ionosphere is thus increasingly opaque as the frequency decreases below about 10 MHz, and this places a lower

4.7 Atmospheric turbulence

frequency limit on spaceborne remote sensing. (It does not apply to airborne techniques, of course, since these do not involve looking through the ionosphere.) However, we can note in passing that the opacity of the ionosphere at sufficiently low radio frequencies does have a beneficial effect, since it allows HF ('short-wave') radio signals to propagate for long distances round the Earth's surface by bouncing between the surface and the ionosphere.

Summary

The ionosphere is a region of charged particles a few hundred kilometres above the Earth's surface, caused by radiation from the Sun. The electrons in the ionosphere interact strongly with microwave radiation, and their most important effect is to reduce the speed of pulses of radiation by an amount that is proportional to the total number of electrons and inversely proportional to the square of the frequency. The delay at a frequency of 10 GHz is typically of the order of 0.1 m during the daytime, and 0.01 m at night, although the activity of the Sun can cause significant variations.

4.7

Atmospheric turbulence

One further and potentially important influence that the atmosphere can have on the propagation of electromagnetic radiation is as a result of atmospheric turbulence. This is always present to a greater or lesser extent in the lower atmosphere, and causes variations in the density, and hence refractive index, of the air. The phase of an electromagnetic wave is corrupted by these variations, and this adversely affects the behaviour of an imaging system.

The most useful way to describe the effects of this kind of phase fluctuation, which is of course statistical rather than deterministic, is by specifying the *structure function*. This is usually defined as the variance of the phase difference between two points that would, in the absence of such effects, be in phase. It is therefore measured in radians squared, and is usually a function of the wavelength of the radiation as well as of the separation between the two points. The time-scale over which the phase variance is measured is also often important.

So far as the effect on the resolution of an imaging system is concerned, an approximate idea can be obtained by replacing the turbulent medium by a notional aperture whose size is equal to the separation at which the structure function reaches a value of 1 radian squared. For visible light travelling through the whole of the Earth's atmosphere, this separation is typically 0.2 m, corresponding to an angular resolution (calculated from the diffraction formula, Equation 2.44) of about 3×10^{-6} radians or about 1 second of arc. This scattering angle, which we denote by $\Delta\theta$, is the limiting angular resolution that can be achieved by an *upward-looking* observation (such as an astronomical observation) through the whole atmosphere. However, for a *downward-looking* observation we also

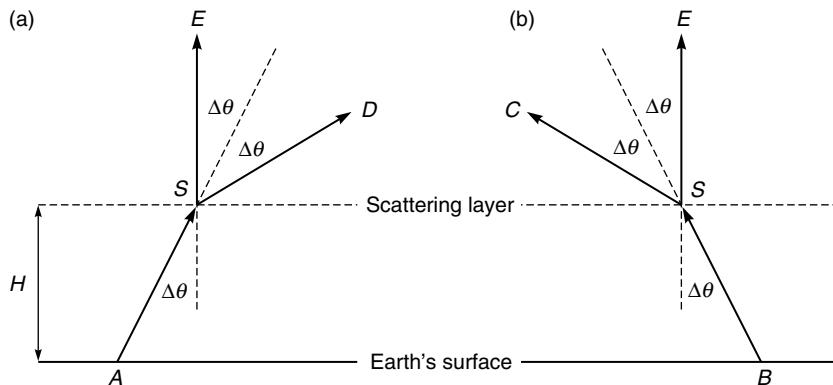


Figure 4.15. Schematic representation of the scattering of light travelling through the atmosphere. (a) The ray AS is scattered into the cone of directions between SE and SD ; (b) the ray BS is scattered into the cone of directions between SC and SE .

need to consider the effective height at which the scattering occurs. We take this to be the scale height of the atmosphere (of the order of 8 km) which we denote by H , and assume that the observation is made from a height much greater than H . In our very simplified model, we assume that all the scattering takes place at the height H .

Figure 4.15a shows a ray AS propagating obliquely through the atmosphere at an angle $\Delta\theta$ to the vertical. At S , it is scattered into a cone of directions from SE , which is clearly vertical, to SD . Figure 4.15b similarly shows the scattering of the ray BS , where S is the same point as in Figure 4.15a. We can see, therefore, that a downward-looking observation from directly above the point S will collect radiation originating from any point between A and B . The distance between these points is approximately $2H\Delta\theta$, so this is the limiting *linear* resolution that is achievable. Substituting the values for H and $\Delta\theta$, we find that this limiting resolution is of the order of 5 cm.

Turbulence in the lower atmosphere causes similar effects at radio frequencies. However, for observations from satellites at radio frequencies, the *ionosphere* poses a potentially more serious problem. It is not really possible to quote a typical value for the ionospheric structure function because of the great variability alluded to earlier, but we may note that the phase variance will be proportional to λ^2 (because of the plasma dispersion relation, discussed in Section 4.5), and that it will be greater near the geomagnetic poles and equator, and during the day time.

Summary

Natural turbulence in the atmosphere causes random fluctuations in the direction of electromagnetic radiation passing through it, and these will in general limit the spatial resolution of a downward-looking imaging system to a few centimetres if it looks through the whole atmosphere.

Review questions

Describe the composition of the Earth's atmosphere, including the variation with altitude of the temperature, pressure and density.

Discuss the propagation of

- (i) visible and near-infrared
- (ii) thermal infrared
- (iii) microwave

radiation through the Earth's atmosphere.

When is the sky blue, and why? What is the relevance of this fact to atmospheric correction of visible-wavelength remotely sensed imagery?

What are aerosols? What is their significance in remote sensing?

Outline the importance of cloud and rain to remote sensing in the visible and near-infrared, thermal infrared and microwave regions of the electromagnetic spectrum.

What is the significance of the ionosphere in remote sensing?

Problems

- Assuming that the temperature T of the troposphere is given by the expression

$$T = 288 - 0.0065z,$$

where T is measured in kelvin and z is the altitude in metres, and that the pressure p in kPa is given by

$$p = 101(T/288)^{5.25},$$

calculate the density of the troposphere at an altitude of 10 km.

- The attenuation coefficient of a typical tropospheric aerosol is 0.1 km^{-1} at sea level, and the total optical thickness of the aerosol for a vertical path through the atmosphere is 0.2. By assuming that the density, and hence the attenuation coefficient, of the aerosol obeys a negative exponential distribution with height (i.e. similar to Equation 4.5), calculate the scale height of the aerosol layer.
- The meteorological visibility in fog is defined as the distance that gives an optical thickness of 4. If a typical fog has a mass density of 10^{-3} kg water per cubic metre and gives a visibility of 100 m, estimate the size of the water droplets.
- Use Equation (4.12) to show that, at low microwave frequencies, the absorption coefficient of a cloud of water droplets is given approximately by

$$0.5 \left(\frac{\rho}{\text{kgm}^{-3}} \right) \left(\frac{f}{\text{GHz}} \right) \text{dBkm}^{-1},$$

(i.e. independent of the droplet size), where ρ is the mass of water in kg per m^3 of cloud. Assume that, for water at microwave frequencies, $\varepsilon_p = 75.9$, $\varepsilon_\infty = 4.5$ and

$t = 9.2$ ps. Hence, by assuming that this relationship holds throughout the microwave region, discuss the statement that clouds are transparent to microwave radiation. Consider only absorption, i.e. ignore any scattering from the droplets. Cloud water content ranges typically from 10^{-6} kg m⁻³ for haze to 10^{-2} kg m⁻³ for cumulonimbus.

5. Use the data of Table 4.3 to estimate the effective raindrop radius, number density and sedimentation rate for rain rates of 1 and 100 mm/hr, assuming that all the drops are spherical and have the same radius.

5

Photographic systems

Aerial photography, as we remarked in Chapter 1, represents the earliest modern form of remote sensing system. Despite the fact that many newer remote sensing techniques have emerged since the first aerial photograph was taken over 100 years ago, aerial photography still finds many important applications, and there are many books that discuss it in more detail than will be possible in this chapter. The interested reader is referred, for example, to Berlin and Avery (2003) and a detailed treatment of photogrammetry and stereogrammetry is give by Kraus (2007). Aerial photography is familiar and well understood, and is a good point from which to begin our discussion of types of imaging system. In particular, it provides a convenient opportunity to introduce some of the imaging concepts that will be useful in discussing some less familiar systems in later chapters.

Photography responds to the visible and near infrared parts of the electromagnetic spectrum. It is, in the context of remote sensing, a passive technique, in that it detects existing radiation (reflected sun- and skylight), and an imaging technique in that it forms a two-dimensional representation of the radiance of the target area. In this chapter we shall consider the construction, function and performance of photographic film, especially its use in obtaining quantitative information about the geometry of objects. Although film-based aerial photography is still dominant, digital photography is beginning (at the time of writing) to rival it, so the chapter includes a brief comparison of the two methods. The chapter then discusses the effects of atmospheric propagation, and concludes by describing the characteristics of some real instruments and giving a brief account of the applications of the technique.

5.1

Photographic film

The traditional photographic process is based on a chemical reaction: the conversion of a salt of silver into metallic silver by the absorption of a photon. Photographic film consists of very many small crystals of a halide of silver (for example, silver bromide) embedded in gelatin and supported on a plastic base. The gelatin layer is typically about 10 µm thick, and the crystals, or grains, of silver halide are typically 0.1 to 5 µm in size. Absorption of a sufficiently energetic photon converts a grain into metallic silver, which is opaque, and unexposed grains are later removed by a chemical process. The result is called a *negative*,

because if it is viewed by passing light through it, areas that received light during the exposure stage will appear dark, and those that did not will appear light.

The simplest type of film is black-and-white, or *panchromatic*, film. The photochemical reaction on which the photographic process depends requires that the photon has some minimum energy and hence maximum wavelength. For the halide ions in the crystalline state, these maximum wavelengths lie in the range 0.4–0.5 μm, so photographic film should respond only to blue, violet and ultraviolet radiation (and in fact to shorter wavelengths, as far as X-rays). It is the opacity of the optical glass and the material of the film itself that prevent these shorter wavelengths from causing problems). In fact, the range of sensitivity is extended by the use of sensitising dyes. Panchromatic (i.e. ‘all colours’) black-and-white film normally has an upper sensitivity limit of about 0.70 or 0.72 μm. The lower limit is in practice set by the opacity of the glass lenses used in the system, at about 0.35 μm, although if an extended response into the ultraviolet is required, quartz lenses can be used to give a cutoff at about 0.30 μm. The sensitivity range of panchromatic film can be extended further into the infrared region by the use of appropriate sensitising dyes, to give an upper cutoff wavelength of about 0.9 μm. In practice, this kind of film is often used in conjunction with a filter that eliminates wavelengths below about 0.7 μm, thus giving a uniform response between about 0.7 and 0.9 μm. This procedure is often referred to as *true infrared photography*.

Colour films are constructed with three layers of emulsion instead of just one. By the incorporation of suitable dyes and filters, the layers are made sensitive to blue, green and red light respectively. *False-colour infrared* (FCIR) film has a similar construction to colour infrared film, the layers being sensitive to near-infrared (0.7–0.9 μm), green and red light. After exposure and development to produce a positive transparency, the parts of the film that were exposed to infrared radiation transmit red light, those that were exposed to red radiation transmit green light, and those that were exposed to green radiation transmit blue light. This is termed the *colour shift*. False-colour infrared film was originally developed during the Second World War to assist in the location of military equipment hidden by camouflage netting. Real vegetation has a high reflectance in the near-infrared region of the spectrum (as discussed in Section 3.6.1), and hence appears bright red in FCIR film, whereas material painted to simulate the visible appearance of vegetation has a low reflectance in the infrared region, and appears blue in the FCIR film. Since the 1940s, many civilian uses have been developed for FCIR film, most of which are derived from its sensitivity to vegetation.

5.1.1

Performance of photographic film: speed, contrast and spatial resolution

The response of a photographic film is normally defined in terms of *photometric units* rather than the radiometric units that we have already met in Chapter 2. Photometric units are weighted with respect to the nominal spectral sensitivity of the human eye, using the function $V(\lambda)$ shown in Figure 5.1.

The photometric quantity corresponding to irradiance is called the *illuminance*, and it is defined by Equation (5.1):

$$E_v = K \int_0^{\infty} E_{\lambda} V(\lambda) d\lambda. \quad (5.1)$$

5.1 Photographic film

Table 5.1. Corresponding radiometric and photometric quantities

Radiometric quantity	Unit	Photometric quantity	Unit
Radiant power	watt (W)	Luminous flux	lumen (lm)
Radiant intensity	W sr^{-1}	Luminous intensity	candela (cd) = lm sr^{-1}
Radiance	$\text{W m}^{-2} \text{sr}^{-1}$	Luminance	cd m^{-2}
Irradiance	W m^{-2}	Illuminance	$\text{lux} = \text{lm m}^{-2}$
Radiant exitance	W m^{-2}	Luminous exitance	$\text{lux} = \text{lm m}^{-2}$

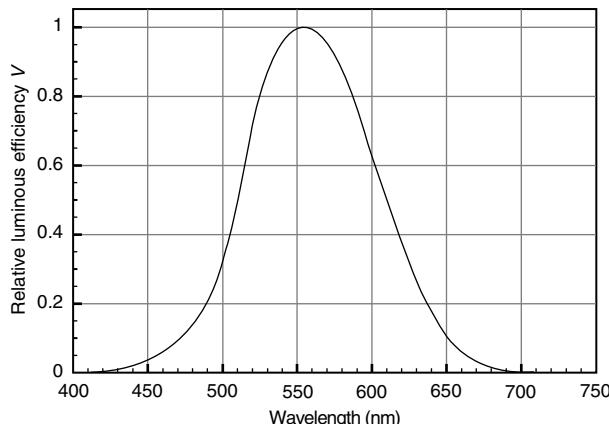


Figure 5.1. The function $V(\lambda)$ that defines the nominal sensitivity of the light-adapted human eye, and hence the photometric units.

In this equation, E_v (the ‘v’ stands for ‘visible’) is the illuminance, E_λ is the spectral irradiance, and K is a constant with the value 680 lumens per watt. All of the radiometric quantities defined in Section 2.5 have their photometric equivalents, defined analogously to Equation (5.1). Table 5.1 summarises the correspondences between these quantities and the names of their units.

The response of a photographic film to incident radiation is characterised most simply by three parameters: the speed, contrast and spatial resolution. The *speed* of a film refers to the length of time for which it must be exposed to light of a given illuminance in order to achieve a significant change in its opacity after processing. The speed is usually quantified by an ASA (American Standards Association), ISO (International Standards Organisation) or DIN (Deutsche Industrie Norm) number, with larger numbers corresponding to faster films and hence shorter exposure times, although for films intended specifically for aerial photography the AFS (aerial film speed) index is used.

The speed of a film describes its response to light of a single illuminance. The *contrast*, on the other hand, describes the effect of changing the illuminance (or the exposure time). If a small change produces a large change in the opacity of the processed film, the film is said to have a high contrast, and conversely. The grain size controls both of these parameters to a large extent (the chemical processing to which the film is subjected also

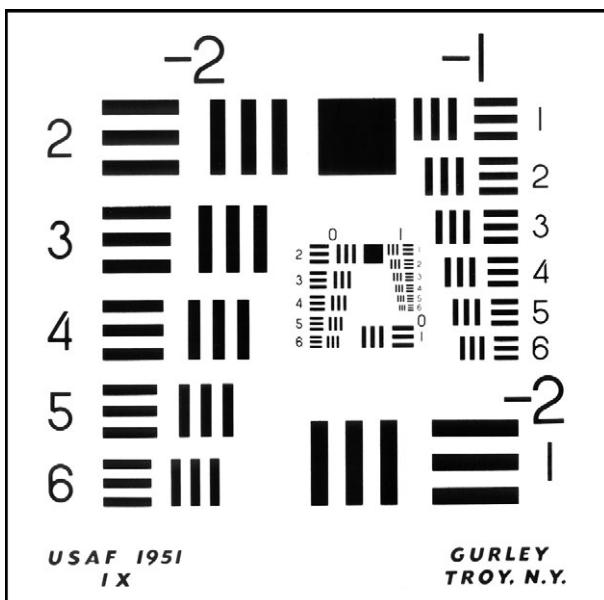


Figure 5.2. Typical object used to determine the spatial resolution of a photographic system in terms of line-pairs per unit length. (This is the 1951 pattern test chart used by the United States Air Force.)

has an effect), high speed films being associated with large grains and high contrast films with a range of grain sizes.

Spatial resolution is, roughly speaking, the ability of a remote sensing system to distinguish an extended object from a point. It is one of the most important parameters describing the performance of a system, and we shall return to it again in subsequent chapters. For photographic systems, the resolution is normally expressed in line-pairs (lp) per unit length. That is to say, a bar pattern resembling Figure 5.2 is photographed, and said to be resolved if the bar pattern is recognisably reproduced on the negative. The spatial resolution is then the greatest number of these bars, per unit length, that can be resolved *on the negative*. If we denote this resolution by r , which has the dimensions of 1/length (so that, for example, a resolution of 100 lp/mm corresponds to $r = 10^5 \text{ m}^{-1}$), this is conventionally regarded as being equivalent to the ability to resolve two *points* on the negative separated by a distance δx , where

$$\delta x = \frac{1}{2r}. \quad (5.2)$$

Thus, a typical mapping film with a resolution of 50 lp/mm can resolve objects 0.01 mm wide on the negative.

The spatial resolution of a film depends on the grain size which, as we have remarked, is typically 1 to 10 μm . Higher resolutions will require smaller grains and, as we have seen, this will result in slower film speeds. The highest spatial resolutions available for aerial photographic films are typically 200 lp/mm, with corresponding low film speeds of about 10 AFS units. Such films are used for reconnaissance. At the other extreme, a fast mapping film might have a spatial resolution of about 20 lp/mm and a film speed of 1000 AFS units.

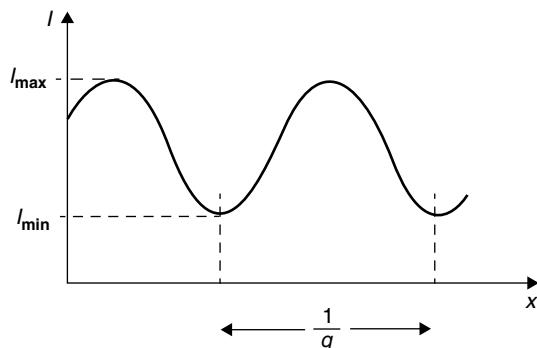


Figure 5.3. A sinusoidal variation of intensity I with position x . This form of variation is characterised by its modulation $m = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$, and its spatial frequency q .

The use of line-pairs per unit length as a measure of film resolution is a comparatively crude measure. A more informative representation is given by the *modulation transfer function* (MTF). This describes the ability of the film to record sinusoidal variations in intensity, as a function of their spatial frequency. Because of the fact, mentioned but not proved in Section 2.3, that any ‘reasonable’ function can be constructed from a (possibly infinite) set of sinusoids of different amplitude, phase and frequency, this approach contains in principle all the information about the spatial response of the film.

Figure 5.3 shows a sinusoidal variation of intensity with position. Apart from its phase, which we shall ignore, this function is characterised by its spatial frequency q (cycles per unit length) and its modulation m , defined as

$$m = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}. \quad (5.3)$$

The MTF of a photographic system is defined as the ratio of the output modulation, i.e. that produced on the film, to the input modulation, i.e. on the target. It is a function of the spatial frequency q , and for photographic systems this is conventionally defined as the spatial frequency in the film plane. Figure 5.4 shows typical MTFs for coarse- and fine-grained films.

5.1.2

Digital photography

Although most aerial photography collected for quantitative purposes is still (at the time of writing, in 2011) acquired using photographic film as the detecting element, digital photography is beginning to challenge it. The technology of digital photography overlaps with that of electro-optical imagers, discussed in Chapter 6, so in this section we shall just make a few rather general remarks.

The main reason for the continued predominance of film photography until now has been spatial resolution. We saw in Section 5.1.1 that a typical mapping film might have a spatial resolution of 50 line-pairs per millimetre, assumed to be equivalent to the ability to resolve an object 0.01 mm wide on the negative. A 230-mm square negative can thus resolve something like $(230/0.01)^2 \approx 500$ million point-like objects. In the language of

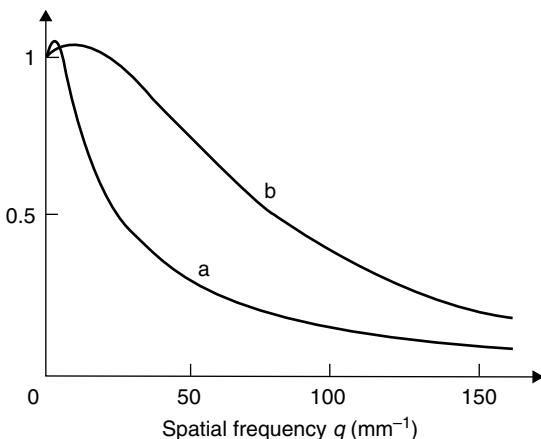


Figure 5.4. Modulation transfer functions for typical photographic films: (a) low resolution; (b) high resolution.

digital photography, we might therefore describe this as a 500-megapixel system. The comparison is not precise, for a number of reasons. Most digital cameras record colour images using three wavebands. They have, however, a single array of sensors, so that some of the sensors are used to record red light, some green light and some blue light. In other words, the ability of the sensor array to resolve spatial variations in colour is somewhat poorer than its ability to resolve variations in illuminance. Most digital cameras also apply considerable amounts of image processing, often including irreversible image compression (see Section 11.6.1) to the image data.

Digital cameras used exclusively for aerial photography are generally designed to avoid these difficulties, so the correspondence between numbers of pixels and photographic resolution is somewhat better. We can therefore take our figure of 500 megapixels as a rough guide to the kind of performance that would be required from a digital camera to rival an aerial film-based mapping system. At the time of writing, no 500-megapixel digital camera yet exists, and although some specialised aerial mapping cameras can achieve resolutions of around 200 megapixels, a figure of around 50–100 megapixels is more common.

Summary

Photographic film uses a photochemical process in which exposure to light causes grains of a silver halide to be converted to metallic silver. Film can be manufactured to give sensitivity from the ultraviolet to the near-infrared regions of the electromagnetic spectrum, and colour film combines sensitivity to three different wavebands: normally red, green and blue in the case of true-colour film or infrared, red and green in the case of colour infrared film. Black-and-white (panchromatic) film roughly duplicates the spectral sensitivity of the human eye, and a parallel system of units, termed photometric units and based on the lumen rather than radiometric units based on the watt, is normally used to describe the radiometric properties of this kind of film.

The spatial resolution of film can be specified very simply using the number of line-pairs per millimetre resolvable on the negative, or in more detail by describing the modulation transfer function. High resolution films, suitable for cartographic applications, have small grains of silver halide and are comparatively slow (require longer exposure times) but can achieve resolutions of the order of 0.01 mm on the negative. For a typical film format of 230 mm square, this is equivalent to a digital photograph having around 500 million pixels, a performance that is currently beyond what can be achieved by digital cameras. Film remains dominant over digital photography for high resolution mapping applications.

5.2

Photographic optics

In this section we shall consider briefly the optics of photographic systems. It is assumed that the reader is already familiar with the simple theory of image formation by lenses; if not, any elementary textbook on optics can be consulted.

Let us consider first a system with a single lens of focal length f (Figure 5.5). The object distance u and image distance v are related by

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}, \quad (5.4)$$

and it is clear that an object of height x subtending an angle θ will produce an image of height $v \tan \theta$. As the object distance u becomes very much larger than the focal length f , the image distance v tends to f and the image size to fx/u . Clearly, in all practical cases of aerial photography this will be a justifiable approximation.

Suppose now that we consider an object of uniform exitance, such that the luminance incident at the lens is L , and that the object subtends a small solid angle Ω . The irradiance at the lens is thus ΩL and the total flux intercepted by the lens is

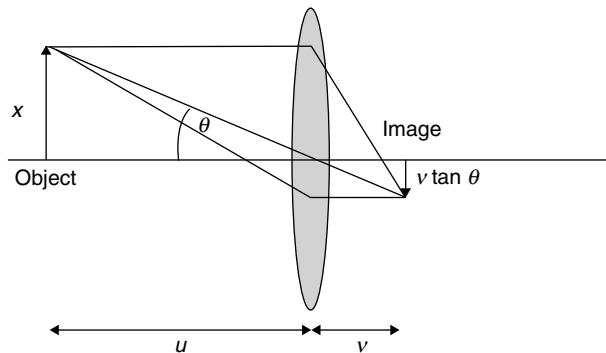


Figure 5.5. Formation of an image by a single converging lens.

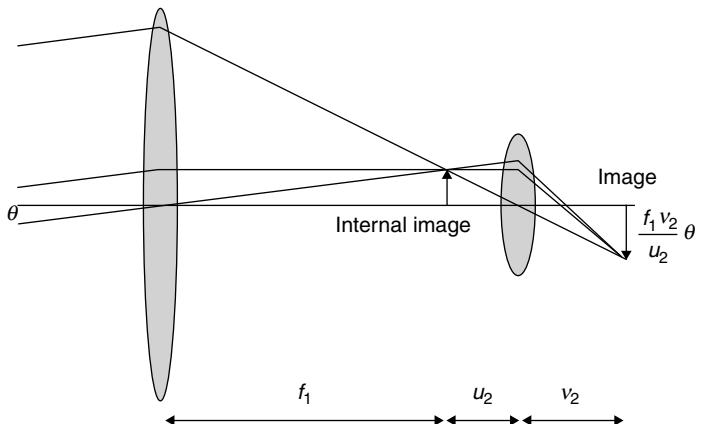


Figure 5.6. Formation of an image by a compound lens. The object is at infinity and subtends a small angle θ .

$$\pi \left(\frac{D}{2}\right)^2 \Omega L,$$

where D is the diameter of the lens. Assuming that there are no losses in the lens, all of this flux will be distributed over an area Ωf^2 on the film, giving an illuminance at the film plane of

$$E_{film} = \frac{\pi D^2 L}{4f^2}.$$

Thus the ratio of the illuminance at the film to the luminance at the lens is given by

$$\frac{E_{film}}{L} = \frac{\pi}{4} \left(\frac{D}{f}\right)^2, \quad (5.5)$$

i.e. it is determined by the ratio f/D , which is called the *f/number* of the lens. The smaller the *f/number*, the larger the lens and the brighter the image. Lenses can be constructed with *f/numbers* as small as about 1, although most lenses used in aerial mapping have *f/numbers* in the range 5 to 10. The *f/number* of a lens can be increased by ‘stopping down’, i.e. by reducing the diameter of an aperture placed just behind the lens.

Most aerial photographic systems in fact use *compound lenses*, which have the advantage of giving increased focal length without making the lens assembly physically larger. Figure 5.6 illustrates a compound lens using two converging lenses, with focal lengths f_1 and f_2 respectively. We again assume that the object is located at infinity (which in practice means at a distance very much greater than f_1), so that the first lens (the *objective lens*) forms an image of it in its focal plane. By comparing Figure 5.6 with Figure 5.5, it is clear that the combined effect of the two lenses is equivalent to that of a single lens of focal length $f_1 v_2 / u_2$, that is, the focal length of the first lens multiplied by the magnification of the second. All of the remarks we have made about the single-lens system remain valid so long as we substitute this effective focal length for the simple focal length f . In fact, an even greater saving of space is made if the second lens of the

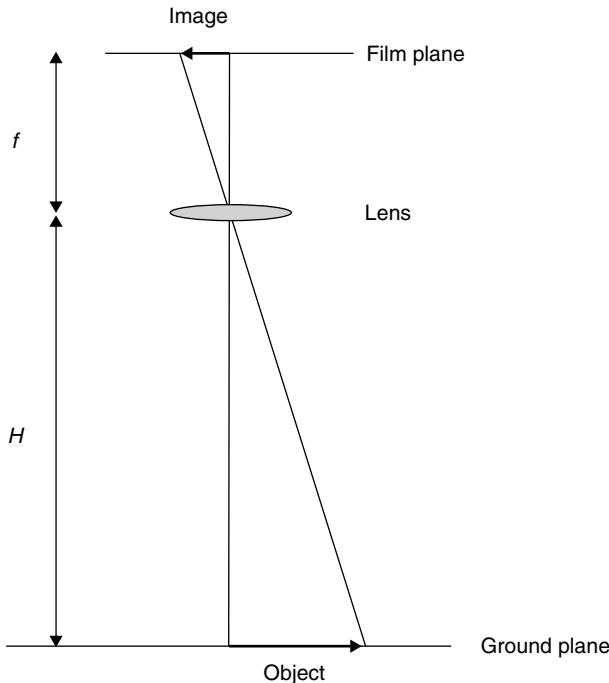


Figure 5.7. Schematic illustration of the formation of a vertical aerial photograph.

compound system is of the diverging, rather than the converging, type. This is the usual construction of telephoto lenses.

The *scale* of a map and, by extension, of an aerial photograph is the number less than unity that expresses the ratio of the size of the representation of an object on the map to the size of the real object. ‘Large scale’ is usually taken to mean greater than 1/50 000 and ‘small scale’ less than 1/500 000.

Figure 5.7 shows schematically the simplest geometry of a vertical aerial photograph, i.e. one in which the optical axis of the camera is directed vertically, normal to the ground surface. We are again assuming that the distance H is much larger than the focal length f , so that the distance from the lens to the film plane can be taken as f . It is clear from simple geometrical considerations that the scale of the image formed at the film plane (i.e. the scale of the negative) is given by

$$s = \frac{f}{H}, \quad (5.6)$$

although of course prints having larger scales can be prepared by photographic enlargement from the negative. It is also clear that, if the negative has a width w , the width of the corresponding region on the ground – the coverage of the aerial photograph – will be given by

$$\frac{w}{s} = \frac{wH}{f}. \quad (5.7)$$

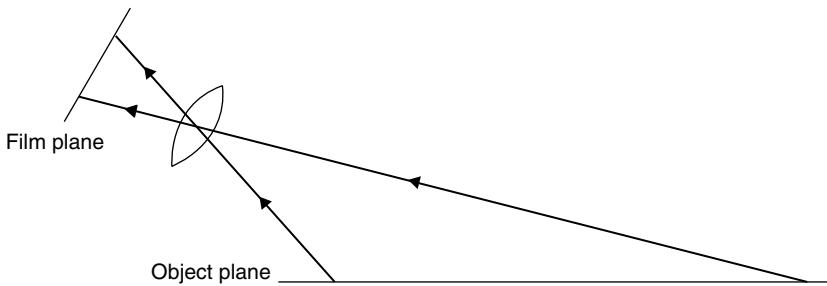


Figure 5.8. The oblique aerial photograph obtains greater coverage at the expense of variable scale and resolution.

The spatial resolution at the ground is also determined by the imaging geometry, and by the resolution of the film. In Section 5.1.1 we introduced the film resolution in terms of the maximum number of line-pairs per unit length resolvable on the negative. As before, we denote this resolution by r , so that (using Equation 5.2) the spatial resolution on the ground (defined as the minimum separation of two resolvable points) is given by

$$r_g = \frac{H}{2rf}. \quad (5.8)$$

Comparison of Equations (5.7) and (5.8) shows that, other things being equal, higher spatial resolution (smaller r_g) requires a longer focal length f and will hence give a smaller coverage.

In fact, the spatial resolution of a photographic system is a combination of the performances of the film and of the optics. We saw in Chapter 2 that diffraction at an aperture of diameter D broadens plane parallel radiation into a cone of angle $\approx \lambda/D$ radians, where λ is the wavelength, and this also sets a limit on the resolution. For example, a system operating at a nominal wavelength of $0.5\text{ }\mu\text{m}$ with an objective lens of diameter 1 cm will be diffraction-limited to an angular resolution of about 5×10^{-5} radians. A point-like object will be imaged as a blurred spot on the negative. If the focal length of the lens is 150 mm , the radius of this spot will be about $7.5\text{ }\mu\text{m}$. Thus, if the resolution of the film is greater than about 70 lp/mm , it is the lens, rather than the film, that will determine the system resolution. The combined effect of the different components of the photographic system on the spatial resolution can conveniently be described using MTFs. The MTF can be defined for each component, and these are then multiplied to calculate the MTF for the system as a whole.

Note that all of our considerations of ground scale, resolution and coverage have been derived for the case of *vertical* aerial photography. *Oblique* photography (Figures 5.8 and 5.9) gives much greater coverage, but at the expense of variable resolution and scale. For this reason it is generally unsuitable for quantitative analysis.

5.2.1

Lens distortion

Our analysis of photographic geometry so far has been based on the assumption of rectilinear propagation, or the ‘pinhole’ model of the camera lens. Real lenses may show



ISS013-E-77377

Figure 5.9. An oblique photograph, showing wide coverage and variable scale. The photograph (ISS013-E-77377) was acquired from the International Space Station on 5 September 2006 and shows the Jungfrau, Mönch and Eiger mountains in the Bernese Alps, Switzerland. Photograph credit: NASA. (http://www.nasa.gov/mission_pages/station/expeditions/expedition17/earthday_imgs.html).

some deviation from this assumption, especially at short focal lengths. Such *lens distortion*, if calibrated, can be corrected for. The simplest assumption is that the distortion is radial, depending only on the distance from the principal point. Radial distortion can be represented by the function $L(r)$:

$$\left(\frac{u^*}{v^*}\right) = L(r) \left(\frac{u}{v}\right) \quad (5.9)$$

where

$$r^2 = u^2 + v^2 \quad (5.10)$$

and where (u, v) are the Cartesian coordinates of a point in the image as actually measured, while (u^*, v^*) are the image coordinates that would have been observed if the lens were truly rectilinear. If $L(r) = 1$ for all values of r , there is no distortion and

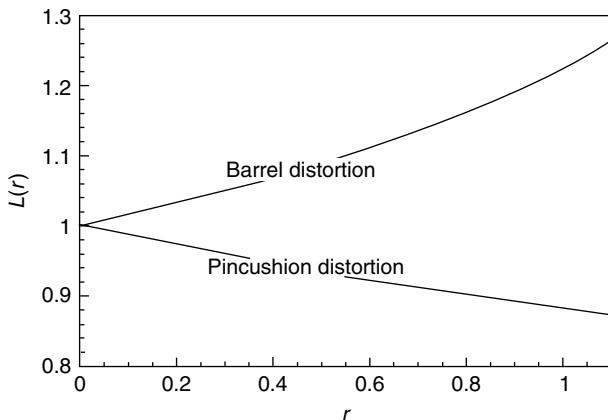


Figure 5.10. Lens distortion curves.

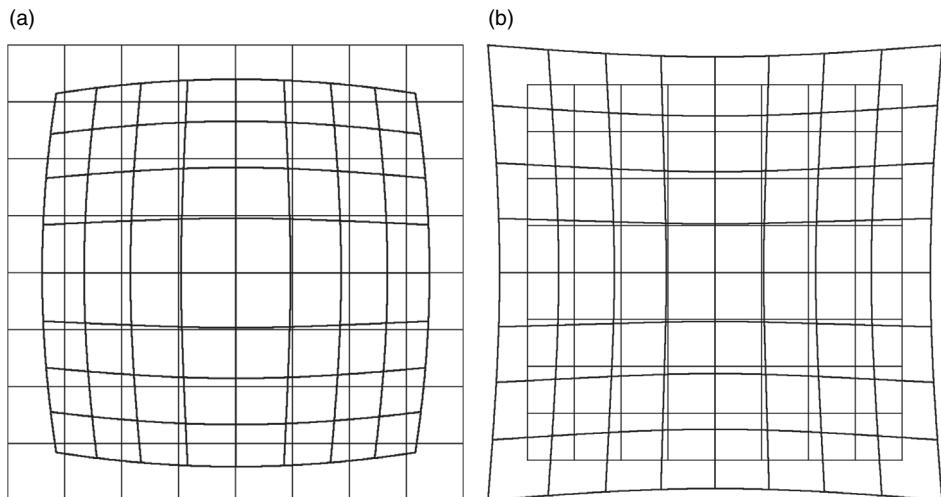


Figure 5.11. Barrel and pincushion distortion, corresponding to the distortion functions in Figure 5.10. In each case the heavier lines show the position actually recorded in the image, while the lighter lines show the position corrected to where it would have been in the case of rectilinear propagation.

straight lines in the real world are projected as straight lines in the image. If $L(r)$ decreases with r it gives rise to *pincushion distortion*, while if it increases with r it gives rise to *barrel distortion*. Figure 5.10 shows two examples of lens distortion functions and Figure 5.11 the corresponding distortions. More complicated forms of distortion are possible.

Summary

The focal length f of the lens of an aerial photography camera controls (other things being equal) the scale, coverage and spatial resolution of the photograph and the amount of light reaching the negative (and hence the exposure time). A large value of f gives rise to a large scale, small spatial coverage and high spatial resolution. It also requires longer exposure times.

In the case of a vertical aerial photograph the scale of the negative is given by f/H , where H is the height of the camera above the ground plane. The spatial coverage is given by wH/f . A typical mapping system operated with $H = 1000$ m would give a scale of the order of 10^{-4} , a coverage of the order of 10^3 m and a ground resolution of the order of 10^{-1} m.

The geometry of aerial photographs can often be defined very simply on the assumption that rays of light pass through the centre of the lens without deviation. This is the rectilinear or pinhole model of the lens. Deviations from this behaviour give rise to image distortions such as barrel or pincushion distortion, which can be calibrated and corrected.

5.3

Photogrammetry and stereogrammetry

One of the most important groups of applications for which aerial photographs are used is based on measuring the geometrical properties of the image. These properties are especially simple if it can be assumed that the camera optics are described by rectilinear propagation (pinhole optics), or at least that an accurate model of the lens distortion is available. It is very well justified for mapping cameras, less well so for reconnaissance systems.

We need to develop a mathematical model of the imaging geometry and it is convenient to use Cartesian coordinates for this purpose. The coordinates of a point in the real world that is imaged in the photograph can be denoted by (x, y, z) and those of the corresponding point in the image by (u, v) . The relationship between the image coordinates and the real-world coordinates can be expressed fairly simply in matrix-vector notation. This is conveniently factorised into two parts. The first step describes the position and orientation of the camera itself. This is called the *exterior orientation* and its effect is represented by some matrix \mathbf{S} :

$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = \mathbf{S} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}. \quad (5.11)$$

Clearly, since \mathbf{S} transforms one 4-element vector into another it is a 4×4 matrix. The second step describes the image formation within the camera itself. Symbolically, we write

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \frac{1}{z'} \mathbf{T} \begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix}, \quad (5.12)$$

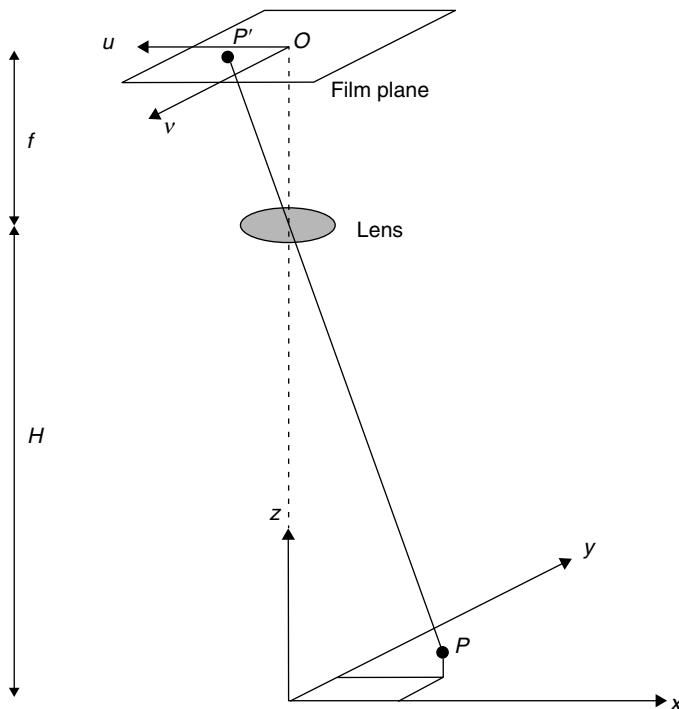


Figure 5.12. Geometry of a vertical aerial photograph in Cartesian coordinates, assuming rectilinear propagation. O is the principal point. An object at P is imaged at P' .

where the 3×4 element matrix \mathbf{T} is referred to as the *interior orientation matrix*. In the case of rectilinear propagation, this matrix has the form

$$\mathbf{T} = \begin{pmatrix} fm_u & 0 & 0 & u_0 \\ 0 & fm_v & 0 & v_0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad (5.13)$$

The non-zero terms in this matrix are u_0 and v_0 , which give the image-plane coordinates of the focal point, f , which is the focal length, and m_u and m_v , which are scale factors in the u - and v -directions and which will normally be equal. The scale factors could be used, for example, to change the units of the image coordinates from units of length into pixel units in the case of a digital photograph. In subsequent analysis, however, we take $m_u = m_v = 1$. We also take $u_0 = v_0 = 0$ since this simplification does not sacrifice any generality.

The simplest geometry of the exterior orientation is the situation when the camera is located at (x_c, y_c, z_c) , is viewing vertically downwards, and is oriented so that the u -axis of the image is parallel to the x -axis and the v -axis is parallel to the y -axis (Figure 5.12). In this case the exterior orientation matrix is simply

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & -x_c \\ 0 & 1 & 0 & -y_c \\ 0 & 0 & 1 & -z_c \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (5.14)$$

5.3 Photogrammetry and stereogrammetry

so that application of (5.12) gives

$$\begin{aligned}x' &= x - x_c, \\y' &= y - y_c, \\z' &= z - z_c.\end{aligned}$$

And application of (5.11) then gives

$$u = \frac{f(x - x_c)}{z - z_c} \quad (5.15a)$$

$$v = \frac{f(y - y_c)}{z - z_c} \quad (5.15b)$$

This now gives us a sufficiently general geometrical framework. First we consider the case where the camera is located at $(0, 0, H)$. This gives

$$u = \frac{fx}{z - H} \quad (5.16a)$$

and

$$v = \frac{fy}{z - H} \quad (5.16b)$$

Setting $z = 0$ in these equations shows that the image of a horizontal surface has a constant scale of f/H , as we have already seen through Equation (5.6). This fact is the basis of simple photogrammetry, in which the lengths, areas and orientations of shapes on horizontal ground can be determined from a photograph, as illustrated by Figure 5.13.

5.3.1

Relief displacement

Equations (5.16) show that an aerial photograph also contains some information about the heights of objects. For example, we suppose that there is a vertical object located at a distance r from the principal point so that the base of the object has coordinates $(r \cos \theta, r \sin \theta, 0)$. This point is imaged at

$$u = \frac{fr \cos \theta}{-H}$$

$$v = \frac{fr \sin \theta}{-H}.$$

If the height of the object is h , so that its top has coordinates $(r \cos \theta, r \sin \theta, h)$, this point is imaged at

$$u = \frac{fr \cos \theta}{h - H}$$

$$v = \frac{fr \sin \theta}{h - H}.$$



Figure 5.13. A vertical aerial photograph of horizontal terrain (a village in Cambridgeshire, UK). A 500-metre grid and a simple map of the road network have been overlaid to show the correspondence between photograph and map coordinates. (Cambridge University Collection: Copyright reserved.)

These two points are not coincident unless $r = 0$, and in general they are separated by a height-dependent amount. This is the phenomenon of *relief displacement*, illustrated in Figure 5.14. Qualitatively, objects appear to lean away from the principal point of the photograph, as can be seen in Figure 5.14.

The length of the projection of the object onto the image, h' , is called the relief displacement. It is given by

$$h' = \frac{fr}{H - h} - \frac{fr}{H} = \frac{hfr}{H(H - h)}. \quad (5.17)$$

Thus, the height of a vertical object can be determined from a single vertical aerial photograph, provided we know the height from which the photograph was obtained, and



Figure 5.14. Relief displacement. The bottom and top of a vertical object (in this case a chimney stack) are imaged at points P_1 and P_2 . (Photograph courtesy of Cambridge University Collection: Crown copyright reserved.)

provided the object is not located at the principal point. We can also see that the distance r' from the image's principal point to the point P_1 is

$$r' = \frac{fr}{H},$$

so the relief displacement can also be written as

$$h' = \frac{hr'}{H - h}. \quad (5.18)$$

If we differentiate Equation (5.17) with respect to h , and make the approximation that $H \gg h$, we find that

$$\frac{\partial h'}{\partial h} \approx \frac{r'}{H}.$$

Thus, if the smallest change in h' that can be resolved in the image is $\Delta h'$, the smallest corresponding change in height is given by

$$\Delta h \approx \frac{H \Delta h'}{r'}.$$

This shows that the accuracy with which heights can be determined will improve with distance from the principal point. Taking r' to have a representative value of $w/2$ (where w is the width of the image), this gives the optimum height resolution as

$$\Delta h \approx \frac{2H \Delta h'}{w}. \quad (5.19)$$

For a typical mapping system, $w = 230$ mm and $\Delta h'$ is of the order of 0.1 mm. This implies that $\Delta h \approx H/1000$, i.e. that heights can be determined from relief displacement to an accuracy that is typically one thousandth of the flying height. This is a common rule of thumb for film photography.

5.3.2 Stereophotography

The method of relief displacement, relying on the measurement of h' , depends for its success on two conditions being met. The first is that the object should not be located at the principal point, and the second is that both the top and bottom of the object should be visible in the photograph. While this second condition may be met for some buildings, for example, it will often not apply in the case of topographic features. In such cases, however, the height information can be retrieved provided that a *pair* of photographs, from different locations, is available. This is the technique of *stereophotography*, and the procedure for determining the topography is called *stereogrammetry*. Figure 5.15 shows a pair of stereophotographs.

To show formally that topographic information can be retrieved from a stereopair, we can consider two positions of our downward-looking camera. As before, one position of the camera is at $(0, 0, H)$. We refer to this as position 1, so that we can put

$$S_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -H \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and we make the same assumptions as before about the orientation of the camera, so that we have

$$u_1 = \frac{fx}{z - H},$$

$$v_1 = \frac{fy}{z - H}.$$

The second position of the camera is taken as (B_x, B_y, H) . B_x and B_y are the Cartesian components of the *baseline*, i.e. the horizontal separation of the camera positions, and for



Figure 5.15. A pair of stereophotographs. Note the different perspectives at the roundabout in the centre of the images. The images were recorded from a height of approximately 900 m, and have a coverage of about 0.9×0.9 km. They show a motorway interchange in Birmingham, UK. (Cambridge University Collection: Copyright reserved.)

simplicity we are assuming that the two photographs are obtained from the same height. While this is not necessary for the stereophotographic method to work, it does simplify the calculations a little. With these assumptions, then, we have

$$S_2 = \begin{pmatrix} 1 & 0 & 0 & -B_x \\ 0 & 1 & 0 & -B_y \\ 0 & 0 & 1 & -H \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and hence

$$u_2 = \frac{f(x - B_x)}{z - H},$$

$$v_2 = \frac{f(y - B_y)}{z - H}.$$

If we assume that f, B_x, B_y and H are known, we need to show that we can find the real world coordinates x, y and z from the image coordinates u_1, v_1, u_2 and v_2 . This is straightforward:

$$x = Cu_1 \tag{5.20a}$$

$$y = Cv_1 \tag{5.20b}$$

$$z = H + fC \tag{5.20c}$$

where

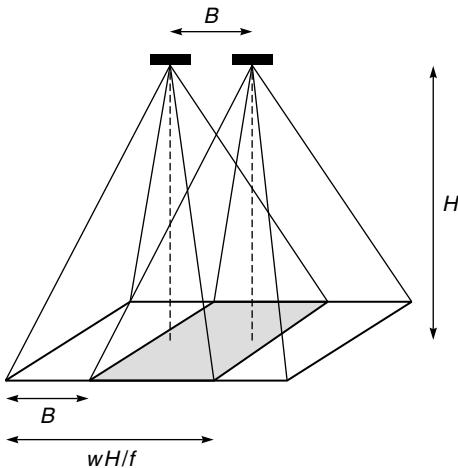


Figure 5.16. Overlap area (shaded) between two vertical aerial photographs.

$$C = \frac{B_x^2 + B_y^2}{(u_1 - u_2)B_x + (v_1 - v_2)B_y}. \quad (5.21)$$

The preceding analysis shows that the difference in the relief displacement between the two photographs is $Bf/(H-h)$. Thus, we expect the accuracy achievable in determining the height to increase as the baseline B is increased. However, this will also have the effect of reducing the overlap between the two photographs, i.e. the area common to both, as shown in Figure 5.16.

From the figure, it is clear that the width of the overlap is given by

$$c = \frac{wH}{f} - B, \quad (5.22)$$

where H is the height, f the focal length, B the baseline and w the film width. A reasonable compromise between the requirements to maximise both c and the height accuracy is to set

$$B \approx 0.4 \frac{wH}{f},$$

giving an overlap width

$$c \approx 0.6 \frac{wH}{f},$$

i.e. 60% of the width of the region imaged by a single photograph. In this case, the accuracy Δh with which heights can be determined is given by

$$\Delta h \approx \frac{H\Delta h'}{0.4w},$$

where $\Delta h'$ is the accuracy with which differences in relief displacement can be measured from the negatives. Taking typical values of $w = 230$ mm, $f = 150$ mm and $\Delta h' = 0.1$ mm, we find that a typical value of the *base-height ratio*, B/H , will be about 0.6, giving an

5.3 Photogrammetry and stereogrammetry

overlap width c of approximately $0.9H$ and a height accuracy of about $H/900$, similar to the result we obtained for relief displacement in Section 5.3.1.

The actual analysis of heights from a pair of stereophotographs may be made by measuring from prints, or by scanning them to convert them to digital form and then analysing in a computer. Formerly, most quantitative analysis of stereophotographs was made using optical-mechanical instruments and these are still used, but the ability to process digital stereophotographs to generate topographical data is now included in at least one commercial image-processing package. One goal of such processing is to produce a *digital elevation model* (DEM) which, as its name implies, is a digital representation of the spatial variation of the height of the Earth's surface. The DEM can be used to correct the original photographs for the variations of scale caused by the relief; such a corrected photograph is called an *orthophotograph*. DEMs find many other applications than the creation of orthophotographs, and can be created by many other means than stereophotography.

A common method of analysing stereophotographs where quantitative height information is not required is to view them through an instrument called a stereoscope, which presents one image to each eye. The human brain is familiar with the interpretation of the slightly different perspectives seen by the two eyes; indeed, this is the normal functioning of binocular vision. The object of the stereoscope, then, is to fool the brain into reconstructing a sensation of three-dimensionality. Figure 5.15 can be viewed in this way. An alternative approach to viewing a pair of stereophotographs is to use them to create an *anaglyph*, in which the two photographs are superimposed but rendered in different



Figure 5.17. Anaglyph created from the stereophotographs of figure 5.15. If the anaglyph is viewed with a red filter over the left eye and a cyan filter over the right eye, the three-dimensional effect can be seen. See also colour plates section.

colours. By viewing the anaglyph through suitably coloured filters, each eye sees only one of the photographs. An example of an anaglyph is shown in Figure 5.17.

Most people's eyes are about 7 or 8 cm apart, and focus comfortably on objects held about 50 cm distant. Because of this, stereophotographs obtained with this base-height ratio (approximately 1/6) will appear to have the correct perspective when interpreted by the brain. Conversely, if the base-height ratio is greater than 1/6, the reconstructed image will appear to have a proportional vertical exaggeration. This vertical exaggeration is merely an artefact of the way the brain interprets parallax, but it is often useful since it enhances subtle variations in relief.

Summary

The well-defined geometry of the process of image formation means that quantitative three-dimensional information can be obtained from photographs, especially vertical photographs. The position of an imaged point on the negative depends on both the horizontal (x and y) and vertical (z) coordinates of the corresponding object (i.e. the scale depends on the height z). In the case of a single photograph this causes the phenomenon of *relief displacement*, in which the image of the top of a vertical object is displaced radially away from the centre of the photograph with respect to the image of the object's base. This phenomenon can be used to measure the heights of vertical objects, with an accuracy that is typically $H/1000$ for a photograph obtained from height H .

In *stereophotography*, vertical photographs from two known locations, separated by the *baseline* B , are combined. By measuring the coordinates of the same feature in the two photographs, its three-dimensional coordinates (x , y and z) can be calculated. Recognising the same features in the two photographs can be carried out manually or, increasingly, automatically within a digital image processing environment. The choice of baseline B controls the accuracy of the method and also the degree of overlap between the two photographs: long baselines give greater vertical accuracy at the expense of smaller overlap. A typical compromise for an aerial mapping camera uses $B \approx 0.6H$, in which case the accuracy with which heights can be determined is again typically $H/1000$.

Stereophotography is extensively used for constructing digital elevation models (DEMs). These can in turn be used to correct for the phenomenon of relief displacement in individual photographs. The result of this process is an *orthophotograph*, in which the horizontal scale is constant.

5.4

Atmospheric propagation

In Chapter 4 we discussed in general the effects of the atmosphere on electromagnetic radiation propagating through it. For a photographic system, the most important of these effects are likely to be the limiting spatial resolution imposed by atmospheric turbulence (although this will only be significant for high magnification systems of high intrinsic resolution), and the reduction in image contrast as a result of atmospheric scattering. Here we consider the latter phenomenon.

The contrast C of a scene, or of some part of it, is commonly defined as

$$C = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}}, \quad (5.23)$$

where L_{\max} and L_{\min} are respectively the maximum and minimum luminances (or, in radiometric terms, radiances) of the scene. Other definitions of scene contrast are also used, for example L_{\max}/L_{\min} , but the definition of Equation (5.23) is convenient as it varies between zero (corresponding to completely uniform luminance) to one (corresponding to a minimum luminance of zero). Absorption of radiation by the atmosphere will reduce the luminances, whereas the corresponding reradiation from the atmosphere will increase them.

Let us write E_s for the illuminance at the Earth's surface due to skylight (and also to direct sunlight, if it is a clear day), E_a for the (upwards) illuminance of the atmosphere itself, T for the intensity transmittance of the atmosphere and r for the reflectivity (defined in Section 3.3.1) of the scene. The contrast of the scene, measured just above it so that the effects of the atmosphere can be ignored, is then

$$C = \frac{r_{\max} - r_{\min}}{r_{\max} + r_{\min}}.$$

The luminous exitance immediately above a part of the scene having reflectivity r is just rE_s . However, above the atmosphere this value will be reduced by a factor of T as a result of atmospheric attenuation, and increased by the upwelling skylight contribution E_a , giving a total upwelling illuminance of $rE_sT + E_a$. Thus the scene contrast detected above the atmosphere becomes

$$C' = \frac{r_{\max} - r_{\min}}{r_{\max} + r_{\min} + \frac{2E_a}{E_s T}}. \quad (5.24)$$

The scene contrast is therefore reduced, and it may be necessary to use a film with a high contrast to compensate for the correspondingly small range of illuminance present at the camera.

Figure 5.18 shows typical sunlight and atmospheric illuminances as a function of the solar elevation angle. The figure shows the sunlight, skylight and total illuminances for downwelling radiation at the Earth's surface, and the upwelling illuminance at altitudes above approximately 8000 m.

As an example, let us consider the scene contrast between two objects with reflectivities of 0.1 and 0.2. The intrinsic scene contrast is clearly 0.33. For a solar elevation of 30°, Figure 5.18 indicates that $E_s \approx 49$ klx and $E_a \approx 2.5$ klx, so if we take the atmospheric transmittance $T = 0.8$, we find that the scene contrast C' above the atmosphere will be reduced to about 0.23. When the solar elevation is 5°, $E_s \approx 8.5$ klx and $E_a \approx 1.0$ klx, so the scene contrast C' will be reduced to 0.17.

Equation (5.23) shows the importance of the atmospheric transmission T in determining the scene contrast above the atmosphere. When T is small, the contrast will be reduced. We have seen in Section 4.2.2 that the optical thickness of the atmosphere increases as the wavelength decreases through the visible band, so we expect the scene contrast to be greater for red light than for blue light. In Section 4.3 we discussed the influence of haze, and noted that the optical thickness of the tropospheric aerosol layer can vary between about 0.1 and 1, giving values of T between 0.9 (clear sky) and 0.4 (hazy).

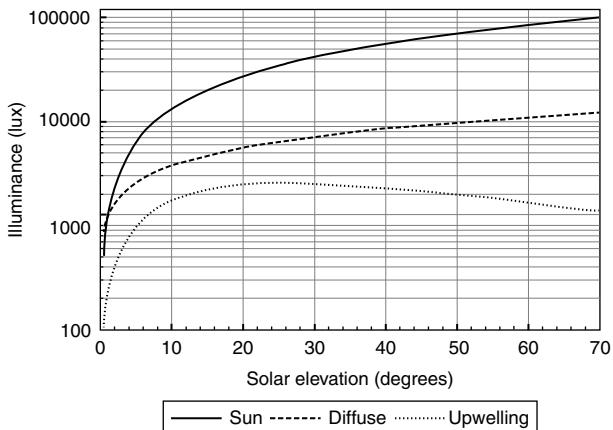


Figure 5.18. Illuminance at the Earth's surface due to direct sunlight and diffuse skylight, and upwelling illuminance of the atmosphere (typical values).

Atmospheric propagation also influences the spatial resolution that can be achieved in an aerial photograph. We saw in Section 4.6 that atmospheric turbulence introduces random changes of phase that normally limit the ground resolution of a high altitude or spaceborne vertical photograph to a few centimetres; photographs acquired from lower altitudes will be proportionately less affected by this phenomenon.

Summary

The contrast in a scene is defined in terms of the spatial variation in the luminance (or radiance) of the light within it. This is contributed to by not only the spatial variation in the reflectance of the scene, but also by the effect of atmospheric propagation. Upwelling radiation from the atmosphere itself reduces the contrast. This can be significant in the case of high-altitude or space photography, and may dictate the use of films more sensitive to contrast. The effect is strongly dependent on wavelength, being more significant at the blue end of the spectrum.

Atmospheric turbulence normally limits the effective ground resolution of a vertical photograph to a few centimetres if it is acquired through most or all of the atmosphere; the effect is proportionately smaller for photographs acquired from lower elevations.

5.5

Some instruments

In this section we describe some actual photographic systems, to illustrate the principles discussed earlier in the chapter. The first example we consider is an airborne mapping camera. There is a wide variety of such cameras, so rather than select the product of a particular manufacturer we describe a typical instrument. This might have a film format

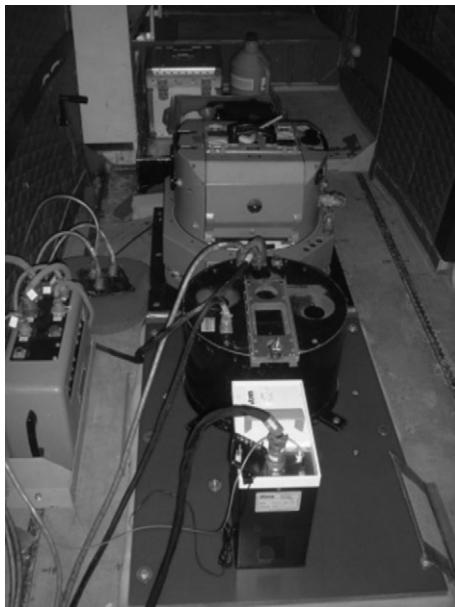


Figure 5.19. Typical aerial mapping camera (Wild RC-10) mounted on a small aircraft (Dornier 220). The camera is the larger assembly at the back.

of 230×230 mm, and operate at a focal length of 150 mm, achieving a spatial resolution of 50 lp/mm and very small distortion. Let us suppose that such a camera is operated from a height of 3000 m (approximately the greatest flying altitude possible without pressurisation of the aircraft). Equation (5.5) shows that the ground coverage achieved will be $4.6 \text{ km} \times 4.6 \text{ km}$, and Equation (5.6) shows that the ground resolution will be roughly 0.2 m. The scale of the negative will be 1/20 000, but it could be enlarged to a scale of perhaps 1/2000 (which would correspond to a spatial resolution of 0.1 mm on the print). The height accuracy achievable from an altitude of 3000 m will be given by the approximate rule $H/1000$ as about 3 m. Figure 5.19 shows a mapping camera mounted in a small aircraft, and Figure 5.20 shows a typical aerial photograph produced by such a camera.

Airborne photographic cameras generally hold very large rolls of film, capable of recording typically 500 to 1000 images. Since flights are of comparatively short duration, these can be rapidly returned to a laboratory for processing. Spaceborne photography obviously poses a greater problem in this regard. One solution is to carry the camera on a short-duration manned mission. This was the approach adopted for the *Metric Camera*, which was flown on board Spacelab-1, a short mission of the Space Shuttle flown in winter 1983 at an altitude of 250 km. The Metric Camera had a focal length of 305 mm, a film format of 230×230 mm (it was modified from a standard aerial mapping camera) and an effective spatial resolution of about 35 lp/mm. Its coverage was thus approximately $190 \text{ km} \times 190 \text{ km}$, with a ground resolution of about 12 m, suitable for mapping at scales up to 1/50 000.

The Metric Camera, and the *Large Format Camera* flown on Space Shuttle missions in 1981 and 1984, were short-duration missions providing limited continuity of data. Other photographic instruments have, however, been flown on manned spacecraft, notably the



Figure 5.20. Mapping camera photograph showing the upper part of a glacier in Svalbard. Photograph reproduced by courtesy of Natural Environment Research Council, UK. See also colour plates section.

Russian space station Mir. Amongst other instruments, this carried a camera designated *KFA-1000*, which had a film format of 300×300 mm and a focal length of 1000 mm. From a nominal altitude of 350 km, this achieved a coverage of approximately 100×100 km and a ground resolution of 20 m. Spaceborne photography from manned missions continues through the International Space Station (e.g. Figure 5.9).

Finally, we should mention the *Corona*, *Argon* and *Lanyard* American military reconnaissance satellites flown from 1959 to the early 1970s. These were short-duration unmanned satellites in extremely low altitude orbits, chosen to give very high ground resolution. They carried cameras codenamed ‘Keyhole’. Canisters of exposed film were ejected from the satellites and retrieved by high altitude aircraft. As an example, Keyhole 4B was carried on a series of Corona satellites between 1967 and 1972, at a nominal minimum altitude of 150 km. The film format was 55×760 mm and the focal length 610 mm, giving a coverage of 14×190 km, and the effective spatial resolution was 160 lp/mm, giving a ground resolution of about 1 m. Since 1995 much of this imagery (over 800 000 images) has been declassified (Macdonald 1995, Campbell 2008). Figure 5.21 shows an enlarged portion of a Keyhole image of central Moscow. The Soviet Union, and latterly the Russian Federation, operated a broadly similar series of satellite photography missions termed *Zenit*, from 1960 to 1994. From about 1963 onwards, these probably achieved spatial resolutions of around 1 m (maybe better).



Figure 5.21. Extract of a Keyhole photograph of central Moscow, recorded on 22 November 1967. The extract shows an area approximately 5×5 km.

Summary

Photography used as a remote sensing technique is almost entirely airborne. A typical airborne mapping camera has a film format of 230 mm and is capable of imaging an area of 1–10 km on a side with a ground resolution of the order of 0.1–1 m. Spaceborne photography is technically challenging since it requires either a manned space mission or elaborate procedures to return exposed film to Earth. Photography was carried out from various manned space missions, especially the Space Shuttle and Mir, and continues from the International Space Station. Very large numbers of photographs were acquired from unmanned US military reconnaissance satellites in the 1960s and 1970s, with ground resolutions of around 1 m, and many of these are now declassified and publicly accessible.

5.6

Applications of aerial and space photography

The applications of aerial photography are in general well known, and (with the exception of the use of infrared film) the correspondence between a photographic image and the perception provided by the human eye and brain is sufficiently great that most applications in any case have an intuitive feel about them. The main advantages of photography as a remote sensing technique are that it is controllable and comparatively inexpensive, and that photographic optical systems can be made with sufficient precision and lack of distortion that quantitative spatial information can be obtained from the images. In this way, aerial photography has found very widespread application in mapping and surveying, for example in geology and hydrology, in terrain analysis, archaeology, field mapping, regional planning, the study of crop types and diseases and so on. Colour film, though more expensive and complicated to process, is widely used, especially in cases where vegetation is studied (for example in agriculture, forestry and ecology) but also in geomorphology, hydrology and oceanography. Black and white near-infrared film has proved especially useful in studying soil moisture and erosion, and in archaeological surveying. False-colour infrared film has found applications in the classification of urban areas, in monitoring soil moisture, in disaster assessment, and especially in vegetation mapping and monitoring.

Summary

Aerial photography is widely used in mapping and surveying, largely because of its high spatial resolution and wellcontrolled geometry. Colour and infrared photography is particularly useful in a wide range of environmental applications.

Review questions

Outline the photochemical basis of photography

Explain the distinction between photometric and radiometric quantities.

Describe the principal advantages and disadvantages of film photography as a remote sensing technique.

Explain what is meant by *relief displacement*.

Outline the principles of stereophotography.

Explain how propagation through the atmosphere decreases the contrast in an aerial photograph.

Problems

1. Estimate the illuminance at the Earth's surface due to solar illumination when the Sun is 45° above the horizon. Use Equation (5.1), but approximate the function $V(\lambda)$ shown in Figure 5.1 by a function that is 1 for λ between $0.51 \mu\text{m}$ and $0.61 \mu\text{m}$, and zero everywhere else. Compare your answer with Figure 5.18.
2. A typical 35-mm camera with a standard lens and normal outdoor film can be assumed to have the following parameters: focal length $f = 50 \text{ mm}$; spatial resolution $l = 40 \text{ lp/mm}$; film format $w = 25 \times 35 \text{ mm}$. Estimate the performance of this system when used to obtain vertical aerial photographs from an altitude of 5000 m .
3. A vertical aerial photograph reveals a tall building. The foot of one corner of the building has (x, y) coordinates $(30.5, 62.0)$ (both measured in mm from the lower left-hand corner of the negative), and the top of the same edge appears at $(19.0, 58.0)$. The corresponding coordinates for an adjacent edge of the building are $(30.5, 73.0)$ and $(19.0, 71.5)$. Given that the camera's focal length was 88 mm and the aircraft's altitude 212 m , find (i) the coordinates of the photograph's principal point (directly below the camera), (ii) the height, and (iii) the width, of the building.
4. An aerial photographic system is used for stereophotography. The film negative is a square of side s , the focal length is f , the baseline is b and the photographs (which are acquired with the camera pointing vertically downwards) are obtained from a height H above sea level. The stereophotographs are used to determine the height h of topographic variations above sea level. It can be assumed that $f \ll h \ll H$, and that the spatial resolution measurable on the negative is a .
 - (i) Show that the accuracy Δh with which heights can be resolved by this system is approximately
$$\frac{H^2 a}{bf}.$$
 - (ii) Show that the width w of the area on the ground that is imaged in both photographs, measured in the direction of the baseline, is
$$\frac{sH}{f} - b.$$
 - (iii) Hence show that the value of b/H that maximises the value of $w/\Delta h$ is $s/2f$, and that the corresponding height accuracy is $2Ha/s$.
 - (iv) Comment on the results in (iii) in the light of the commonly adopted rule that heights can be determined from stereophotographs with an accuracy of about $H/1000$.

6

Electro-optical systems

In Chapter 5 we discussed photographic systems, and although these provide a familiar model for many of the concepts to be addressed in this and subsequent chapters, they nevertheless stand somewhat apart from the types of system to be discussed in Chapters 6 to 9. In the case of photographic systems, the radiation is detected through a photochemical process, whereas in the systems we shall now consider it is converted into an electronic signal that can be detected, amplified and subsequently further processed electronically. This clearly has many advantages, not least of which is the comparative simplicity with which the data may be transmitted as a modulated radio signal, recorded digitally and processed in a computer.

In this chapter, we shall consider electro-optical systems, interpreted fairly broadly to include the visible, near-infrared and thermal infrared regions of the electromagnetic spectrum. The reason for this is a pragmatic one, since many instruments combine a response in the visible and near infrared (VNIR) region with a response in the thermal infrared (TIR) region, and much of the technology is common to both. Within this broad definition we shall distinguish imaging systems, designed to form a two-dimensional representation of the two-dimensional distribution of radiance across the target, and systems used for profiling the properties and contents of the atmosphere. It is clear that an imaging system operating in the VNIR region has much in common with aerial photography, and systems of this type are in very wide use from both airborne and spaceborne platforms. We shall therefore begin our discussion with these systems.

6.1

Visible and near-infrared imaging systems

6.1.1

Detectors

We saw in Chapter 5 that the detection process in photography involves a photochemical reaction between the incident radiation and ions of silver halide. In electro-optical systems, the radiation falls on a suitable detector, which generates an electrical signal dependent upon the intensity of the radiation. In this section we consider the main types of detector, although a discussion of the numerical values of their sensitivities is beyond our scope.

6.1 Visible and near-infrared imaging systems

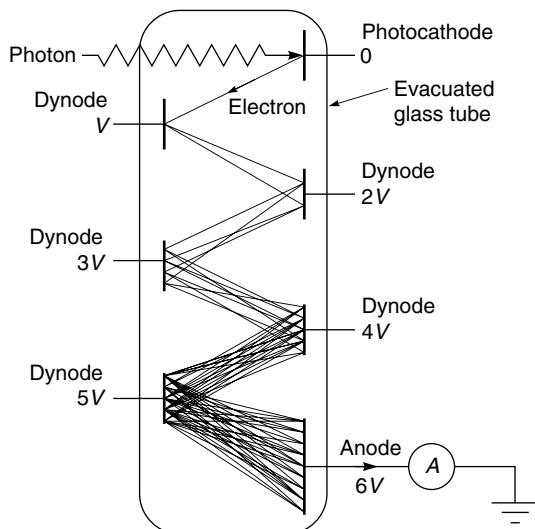


Figure 6.1. A photomultiplier tube (schematic).

One of the earliest types of detector, still sometimes used in the VNIR region (for example in the OLS (Operational Linescan System) instrument carried on the DMSP (Defense Meteorological Satellite Program) series of satellites), is the *photomultiplier*. This is a vacuum-tube device, shown schematically in Figure 6.1. It consists of an evacuated glass vessel containing a number of electrodes that are maintained at different electric potentials. A photon falls on the electrode at the most negative potential (the *photocathode*), where it causes the ejection of an electron by the *photoelectric effect*. This electron is accelerated towards an intermediate electrode (*dynode*) at a more positive potential, and the increased kinetic energy that it acquires causes it to eject more than one electron from the dynode. This process is repeated several times, the number of electrons increasing at each dynode, until the flow of electrons reaching the most positive electrode (the *anode*) represents a measurable current. The size of this current depends on the intensity of the incident radiation.

The minimum photon energy that can be detected by such a device is the *work function* of the material from which the photocathode is made. The work function W is defined as the energy difference between an electron *in vacuo* and an electron within the material, i.e. the energy required to eject the latter. For metals, values of W lie typically in the range 2 to 5 eV, so the maximum wavelength to which such devices are sensitive is about 0.6 μm . Mixtures of alkali metals, however, can have significantly smaller work functions, and sensitivity up to about 1 μm can be obtained.

The photomultiplier is a very sensitive device with a rapid response time of the order of 1 ns. Its main disadvantages are its mechanical fragility, its comparatively large size, and the fact that it needs a high operating potential of about 1 kV.

Radiation in the near-infrared region is usually detected using a *photodiode*. This is a semiconductor junction device, usually indium antimonide (InSb) or lead sulphide (PbS), in which an incident photon generates a current or voltage across the junction. The signal is proportional to the light intensity. The theory underlying the operation of photodiodes is quite complicated, and is outlined below.

A semiconductor diode consists of two abutting pieces of semiconducting material. One piece has been *doped* with a trace of impurity that gives rise to an excess of electrons, the other with an impurity giving a deficit of electrons. These are referred to as n-type and p-type material respectively, since the effective charge carriers are, respectively, negatively (electrons) and positively ('holes', i.e. absences of electrons) charged. At the junction between the two materials, holes from the p-type material diffuse into the n-type material, where they combine with the free electrons. A corresponding effect also occurs in the opposite direction, with electrons diffusing into the n-type material, and this gives rise to a *depletion region* of very low conductivity, typically about 1 μm in width. Because there is now an excess of positive charge in the n-type material and an excess of negative charge in the p-type material, there is an electric field from the n-type to the p-type across the depletion region, and this inhibits the further diffusion of charges.

If an external electric field is applied from the p-type to the n-type material (*forward bias*), the internal field is overcome to some extent and the depletion region is made narrower. A current flows, and its magnitude increases approximately exponentially with the applied voltage difference. If however the external field is applied in the opposite sense (*reverse bias*), the depletion region is made wider and a very much smaller current is able to pass.

If now the diode, with no external bias, is subjected to incident electromagnetic radiation, a photon may be able to create an electron-hole pair in the p-type material. If the electron diffuses into the depletion region, it will be accelerated by the internal field into the n-type material, and the work done will appear as a voltage, proportional to the intensity of the radiation, across the diode. This is called the *photovoltaic mode* of operation. Similarly, of course, if the electron-hole pair is created in the n-type material, the positively charged hole will be accelerated into the p-type material. If, on the other hand, the diode is reverse-biased, increasing the width of the depletion region, the external field maintains the field within the depletion region and the charge carriers will generate a current. This is called the *photoconductive mode* of operation, and it gives a much faster response (of the order of 1 ns) compared with the photovoltaic mode (about 1 μs). The reason for this is that the response time is determined by the capacitance of the depletion region, which is inversely proportional to its width. The photoconductive operation of a photodiode is illustrated in Figure 6.2.

The maximum photon wavelength (i.e. minimum energy) that can be detected by a photodiode is determined by the energy required to create an electron-hole pair. This is termed the *band-gap* of the semiconductor. Semiconductors such as germanium have relatively large band-gaps, giving sensitivity up to only 1.7 μm , but PbS responds up to about 3 μm and InSb to about 5 μm .

Photomultipliers and photodiodes are both single-element detectors, which means that, in order to use them to produce images, they must either be combined in large arrays of detectors, or scanned over the target. Clearly, the construction of an array of detectors is more practical for the photodiode (which can be made very small) than for the photomultiplier, but another possibility is represented by the *charge-coupled device* (CCD), now familiar through its use in digital cameras. The CCD is another semiconductor device, consisting in its simplest form of many (of the order of 1000) identical elements in a linear array. Each of these elements is light-sensitive, developing and then storing free charges as a result of illumination. By suitable manipulation of the voltages applied

6.1 Visible and near-infrared imaging systems

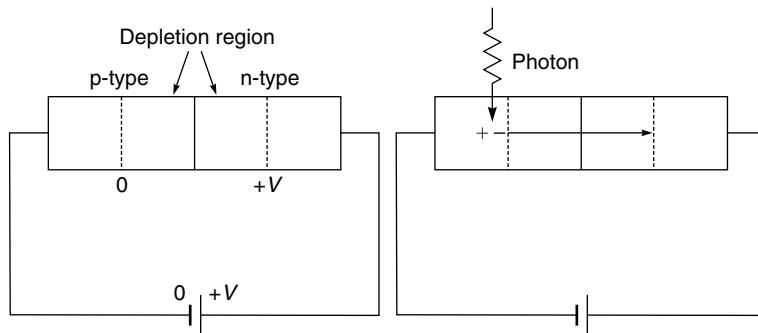


Figure 6.2. Operation of a photodiode (schematic). The p-type region contains free positive charges and the n-type region free negative charges, except in the depletion region. An incident photon creates an extra pair of charge carriers in the p-type material, and the negative charge is accelerated by the potential difference V into the n-type material, thus generating a current. A current in the same direction will be produced if the photon creates an electron-hole pair in the n-type material, in which case a positive charge carrier is accelerated into the p-type material.

to the elements of the array, these charges can be moved from one element to the next, until they are finally ‘read out’ at the end of the array. In other words, the device operates as a shift-register. Silicon-based CCDs are sensitive to wavelengths up to about $1.1\text{ }\mu\text{m}$.

CCDs can be constructed in both one-dimensional (linear) and two-dimensional (planar) forms. The two-dimensional CCD has two major advantages. The first follows from the fact that the controlling voltages, used to shift the charges from one element to the next, are applied only to the edges of the array and not to each individual element. Thus, an array consisting of $n \times m$ elements needs only $(n+m)$ connections, so that arrays of the order of 1000×1000 elements are feasible. Each element is of the order of $10\text{ }\mu\text{m}$ wide, so the whole array is only a centimetre or so in width. The second advantage is that, since the array does not need to be scanned in order to form an image, each element can ‘look at’ (i.e. collect photons from) the target for a longer time than is the case for a scanned system, which increases the potential sensitivity.

6.1.2 Imaging

Having reviewed the principal methods by which VNIR radiation is detected, we now consider the means by which images, i.e. two-dimensional representations of the scene radiance, can be built up.



A two-dimensional detector such as a planar array CCD generates a two-dimensional image. All that is necessary is to ensure that the detector views the scene for long enough to acquire and process sufficient photons from it. However, we should note that, if the platform carrying the detector is in motion relative to the target, it may be necessary to compensate for this motion in order to avoid ‘motion blurring’ of the image. This mode of operation can conveniently be thought of in terms of the detector ‘staring’ at a scene, then moving on to stare at the next scene, and so on. For this reason, it is sometimes called *step-stare imaging*, and is illustrated in Figure 6.3. However, if the detector is one- or zero-dimensional (a linear array or a single detector), some form of mechanical scanning is required.

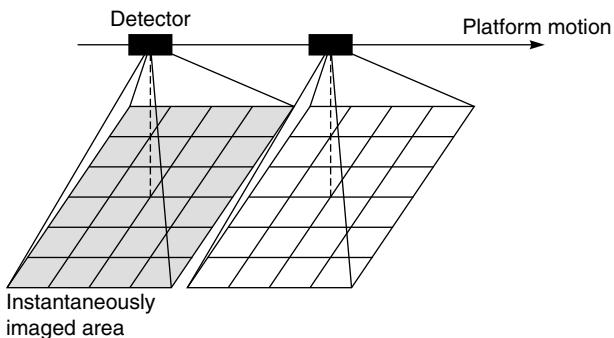


Figure 6.3. ‘Step-stare’ imaging by a two-dimensional detector. The elements within the instantaneously imaged area are called resolution elements or *rezels*; the corresponding elements in the image are called picture elements or *pixels*. In practice, the number of rezels or pixels is of the order of 1000×1000 .

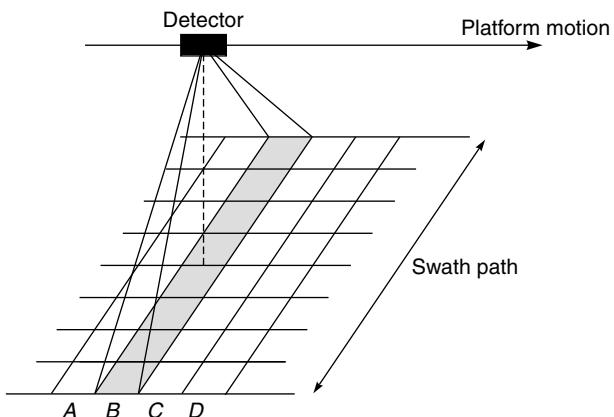


Figure 6.4. ‘Push-broom’ imaging by a one-dimensional detector (linear array). The detector is currently imaging the shaded strip B of rezels; strip A was imaged previously, and strips C and then D will be imaged when the platform has moved forward a sufficient distance.

In the case of a linear array of detectors, this scanning can be achieved using the motion of the platform. A strip of rezels, oriented perpendicularly to the direction of motion, is imaged instantaneously, and the adjacent strip of rezels is imaged when the platform has moved a distance equal to the width of the rezels. In a pleasing domestic analogy, this mode of imaging is often called *pushbroom imaging*, and it is illustrated in Figure 6.4.

This type of imaging is used, for example, by the HRV (high resolution visible) sensor carried on the SPOT satellites.

In the case of a single (zero-dimensional) detector, it is clear that some form of mechanical scanning will also be required, since the forward motion of the platform can provide scanning in only one direction. The usual way of achieving this mechanical scanning is through a rotating or oscillating mirror, which scans the *instantaneous field of view* (IFOV) from side to side, in a direction approximately perpendicular to the forward

6.1 Visible and near-infrared imaging systems

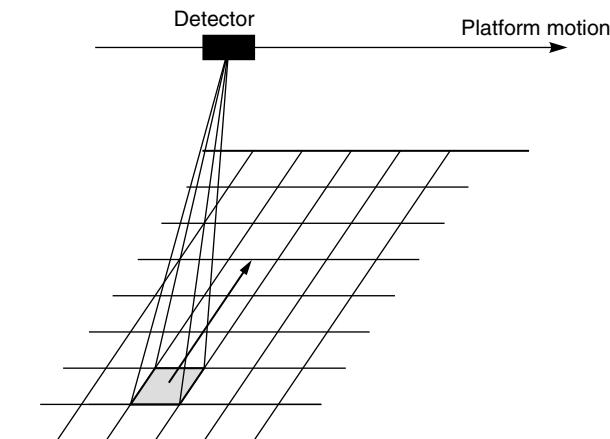


Figure 6.5. ‘Whisk-broom’ imaging by a line scanner. The instantaneous field of view (shaded) is scanned sideways by a mirror within the instrument. Forward scanning is achieved by the forward motion of the platform.

motion. An instrument operating on this principle is usually called a *line scanner*, and the mode of operation is often called *whiskbroom imaging*. It is in widespread use for airborne and spaceborne instruments, for example the TM (thematic mapper) and ETM+ (enhanced thematic mapper) instruments carried on recent Landsat satellites. It is illustrated in Figure 6.5.

It will be apparent that care must be taken to match the speed of the side-to-side scanning to that provided by the forward motion. If the former is too small, or the latter too large, some strips of the surface will not be imaged. The ideal relationship between the two is governed by the width Δx of the IFOV measured in the direction of platform motion. If the platform is not to advance by a distance of more than Δx during the time ΔT taken for one line to be scanned, the scan time must satisfy

$$\Delta T \leq \frac{\Delta x}{v}, \quad (6.1)$$

where v is the platform velocity. The Landsat ETM+ sensor, for example, has an effective IFOV width $\Delta x = 30$ m, and an equivalent ground speed $v = 6.46 \times 10^3$ m s⁻¹, so Equation (6.1) implies that the scan time must be at most 4.6 ms. In fact, the sensor scans 16 lines simultaneously in a time of 71.4 ms, giving an effective scan time for each line of 4.5 ms.

If the platform that carries the sensor is *not* in motion relative to the Earth’s surface, then of course it is necessary to provide mechanical scanning of a single detector in *two* directions. This is the case for geostationary imagers (geostationary orbits are discussed in Chapter 10). Scanning in the east–west direction is normally achieved by spinning the satellite about its north–south axis (this also helps to stabilise the orientation of the axis), while north–south scanning is effected using a mirror within the instrument that rotates in a series of steps. This mode of operation, illustrated in Figure 6.6, is usually called *spin-scan imaging*.

The number of steps made by the scanning mirror determines the number of east–west scan lines. For a typical geostationary imager this is of the order of 2000–5000, giving



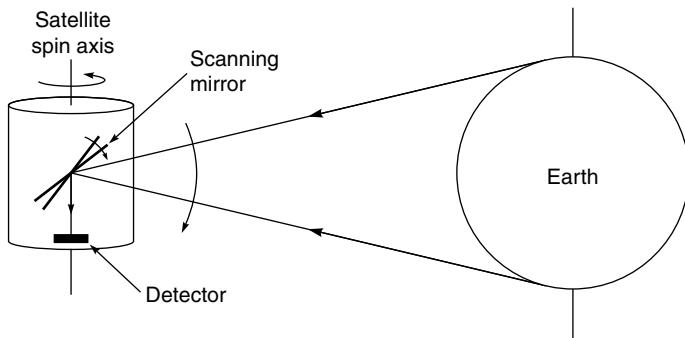


Figure 6.6. ‘Spin-scan’ imaging by a geostationary satellite.

scan lines a few kilometres wide at the Earth’s surface. The angular velocity of the satellite’s spin then determines the time taken to acquire a full image. Higher angular velocities will give shorter imaging times, but also mean that the detector views a given region of the Earth’s surface for a shorter time and hence provides a less sensitive measurement. Typically, the angular velocity is of the order of 100 r.p.m., so the time required to acquire a complete image is of the order of 20–50 minutes.

6.1.3 Spatial resolution

We showed in Section 2.7 that the angular resolution of an imaging system is in general limited by diffraction effects to $\sim \lambda/D$, where λ is the wavelength of the detected radiation and D is the diameter of the objective lens, mirror, or in general the first obstruction encountered by the incoming radiation. Provided that this angular resolution is small (i.e. that the objective lens is many wavelengths in diameter), the corresponding linear resolution at the Earth’s surface can be written approximately as $H\lambda/D$, where H is the distance from the sensor to the surface. However, the design of the instrument may degrade the spatial resolution to a value significantly poorer than this. We have already seen an example of this in Chapter 5, where we noted that the spatial resolution of a photographic system is determined by both the camera (which includes a contribution from diffraction) and the film. In the case of VNIR imaging systems, the spatial response of the film must be replaced by the spatial response of the detector, or array of detectors, but the principle is the same. Since a detector has a finite size, and since the signal derived from it is, in effect, an average of the radiation intensity over the entire detector, it is clear that the spatial resolution cannot be better than the size of the detector *projected through the system’s optics onto the Earth’s surface*. This is the concept of the *rezel* introduced in the previous section. For a simple optical system characterised by a focal length f and a detector size a , the size of the rezel will be $Half$.

Roughly speaking, then, we may state that the spatial resolution of the system will be the larger of the two terms $H\lambda/D$ and $Half$. If the former term is larger, the data are *oversampled*, so that adjacent pixels in the image will be strongly correlated and the image will contain less information than the number of pixels would imply. If the latter term is larger, the spatial resolution of the system could be improved simply, for example by increasing the focal length. Thus a design is in some sense optimal if the two terms are

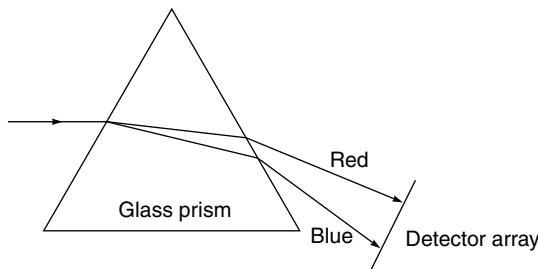


Figure 6.7. Dispersion of broad-band radiation by a prism.

approximately equal, and this is in fact normally the case. A further complication may arise if the scan time ΔT is short enough that adjacent scan lines on the ground actually overlap one another significantly. In this case, the rezel size will be smaller than the IFOV, and again the data are oversampled. In any case, the spatial resolution will be proportional to the height H of the sensor above the Earth's surface. Combining this observation with Equation (6.1), we can see that the height and speed of the platform, and the scan time ΔT , cannot be varied independently. This imposes operational limitations for airborne applications, and design considerations for spaceborne applications (for which the relationship between v and H is fixed). We shall refer to these limitations again in Chapter 10.

The foregoing discussion has outlined some of the issues relating to the spatial resolution of VNIR imaging systems. It should be noted that spatial resolution is in fact a rather elusive concept (see, for example, the discussion by Forshaw *et al.* (1983)). For the present, we note that although the concepts of spatial resolution, rezel size and IFOV are all different, in practice they are of similar magnitude.

6.1.4 Spectral resolution

Most VNIR instruments designed for observing the Earth's surface are capable of discriminating between different wavelength components in the incident radiation. The number of spectral bands provided by such instruments varies from a few to a few hundred. Where there are only a few bands, with comparatively large bandwidths in the range 10 to 100 nm, these are normally defined by filters. Higher spectral resolutions (i.e. smaller bandwidths), to approximately 0.1 nm, are generally obtained by using prisms or diffraction gratings to disperse the spectrum of the incoming radiation onto an array of detectors. Alternatively, a single detector can be scanned mechanically across the spectrum.

Dispersion by a prism is illustrated in Figure 6.7. Radiation undergoes a deviation when it enters and leaves the prism, the deviations being given by Snell's law (Equation 3.30). The refractive index of glass varies somewhat with wavelength, so different wavelengths undergo different deviations. The angular dispersion is not large: for a typical arrangement, the angle between the emerging blue (say $\lambda = 0.40 \mu\text{m}$) and infrared ($\lambda = 1.0 \mu\text{m}$) radiation is of the order of 5 to 10 degrees.

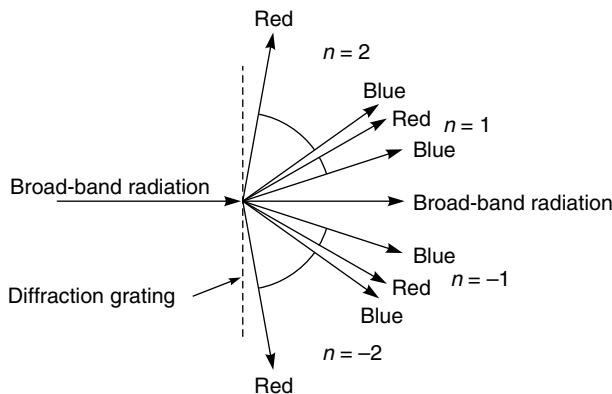


Figure 6.8. Dispersion of broad-band radiation by a diffraction grating. The values of n are the orders of the spectra.

Figure 6.8 illustrates dispersion by a diffraction grating. The grating itself consists of a large number of parallel lines ruled on a transparent (e.g. glass) or reflective (e.g. polished metal) sheet. The lines are regularly spaced at some small separation d . A beam of plane parallel radiation of wavelength λ , striking such a grating at normal incidence, is diffracted into a plane parallel beam that makes an angle θ to the normal, where the value of θ is given by

$$\sin\theta = \frac{n\lambda}{d} \quad (6.2)$$

and n is an integer. Thus, if broad-band radiation is incident on the grating, it will be dispersed in angle, producing spectra corresponding to different values of n , called the *order* of the spectrum. We should note that the value $n = 0$ gives $\theta = 0$ for all wavelengths, so there is a ‘zero-order spectrum’ of undeviated radiation, which has the same spectral content as the incident radiation. This is of course useless from the point of view of spectral discrimination.

Equation (6.2) shows that the angular dispersion of a given range of wavelengths can be increased by making the line spacing d smaller. Values of d as small as $1.3 \mu\text{m}$ are possible, giving an angular dispersion of about 15° for the first-order spectrum between $0.4 \mu\text{m}$ and $0.7 \mu\text{m}$. This is about three times better than can be achieved with a glass prism. The angular dispersion is greater for the higher order spectra but we should note that, for these higher orders, the maximum usable range of wavelengths will be reduced. If the range of wavelengths is too large, different orders of different wavelengths can overlap. Thus, if a diffraction-grating spectrometer is designed for especially high spectral resolution, it is generally necessary to filter the radiation first.

6.1.5

Atmospheric propagation and correction

For many quantitative applications of VNIR imagery, it is desirable to correct the data for the effects of atmospheric propagation. If the data are accurately calibrated, the variable that is measured is the radiance reaching the sensor, but the variable that is wanted is the

6.1 Visible and near-infrared imaging systems

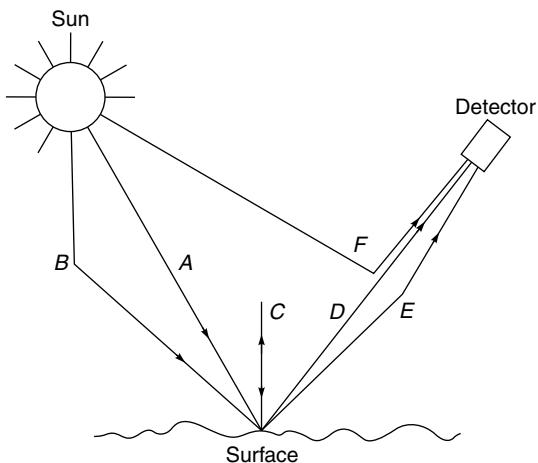


Figure 6.9. Contributions to the radiance measured at a sensor in the presence of atmospheric scattering.

reflectance (for example, the BRDF) of the surface. If there were no atmosphere, the calculation would be a straightforward one, requiring only a knowledge of the radiance at the Earth's surface due to solar illumination, and of the geometry of the Sun–target–sensor arrangement. However, the presence of the atmosphere significantly complicates the relationship between the solar radiance and the radiance detected at the sensor, as illustrated schematically in Figure 6.9.

In this figure, A represents direct illumination of the surface by the Sun, but B shows that the surface can also be illuminated by radiation that has been scattered from the atmosphere – in other words, by skylight. The rays C to E show the possible destinations of radiation scattered from the surface. In C , it has been scattered back to the surface; in D it is transmitted directly to the sensor, and in E it also reaches the sensor, although after being scattered from the atmosphere. Finally, ray F shows that some sunlight can be scattered directly into the sensor. Thus rays A and D represent the simple, ‘no atmosphere’ situation, and the other rays the contributions due to atmospheric scattering. These are clearly less significant for a low altitude airborne observation than for a satellite measurement that views through the entire atmosphere.

Correction of the image data for the effects of atmospheric propagation can be carried out in essentially three ways (Campbell 2008). Physically based methods attempt to model the phenomena illustrated in Figure 6.10. This is the most rigorous approach, and also the most difficult to apply. The atmospheric scattering and absorption characteristics are calculated by a computer model that solves the radiative transfer equation. This approach requires as input data meteorological, seasonal and geographical variables. In practice, these variables may not all be available with sufficient spatial or temporal resolution, and in particular, estimation of the contribution of atmospheric aerosols (see Section 4.3) is difficult.

There are many computer models designed to solve the radiative transfer equation for the Earth's atmosphere (and indeed other planetary atmospheres). The most accurate models take into account the detailed spectral line structure of molecular species in the

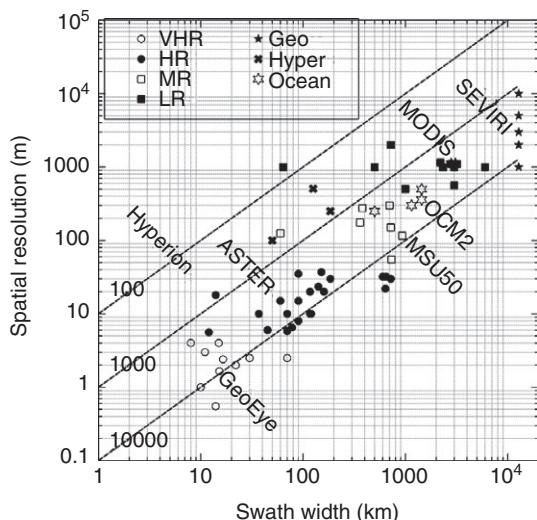


Figure 6.10. Some current (2011) spaceborne VNIR imagers, categorised as very high resolution (VHR), high resolution (HR), medium resolution (MR), low resolution (LR), geostationary imagers (Geo), hyperspectral imagers (Hyper) and ocean-colour imagers (Ocean). The dashed lines show the number of pixels across the swath width.

atmosphere, multiple scattering, and polarisation effects. Models can be characterised by the range of wavelengths for which they are valid, and the spectral resolution, although both of these are normally specified in terms of the wavenumber, using the spectroscopists' definition (see box in Section 3.4.1) of $1/\lambda$. Some well-known models are LOWTRAN (Kneizys *et al.* 1988), MODTRAN (Berk *et al.* 2005) and LibRadTran (Mayer and Kylling 2005).

The second approach to atmospheric correction of VNIR imagery is based on calibration against targets of known reflectance (Chander *et al.* 2010). These targets can be artificially constructed or naturally occurring, but they need to satisfy a number of criteria: (1) their reflectances must be known sufficiently accurately, in the same spectral bands as are used by the imager; (2) the range of reflectances represented by the calibrators must span the range of interest in the scene; (3) each calibrator should cover an area of at least several rezels; (4) the calibrators should be well distributed over the entire scene, so that possible variations of atmospheric conditions from place to place can be assessed and, if necessary, allowed for.

Finally, the simplest, and most widely applied, method of atmospheric correction is based on *dark-pixel subtraction* (Chavez 1988). For each spectral band present in the image, the minimum-radiance pixel is identified, and this minimum radiance is subtracted from all pixels in the image. This method is quite crude: it assumes that the minimum reflectance in each band is zero, that the atmospheric correction can be modelled adequately as an additive effect, and that the correction does not vary from place to place within the scene. To some extent, visual inspection of an image can determine whether these assumptions are likely to be valid. Zero-reflectance rezels can be provided by deep shadows and, in the near-infrared region, by clean water bodies.

Summary

In an imaging system operating in the visible and near-infrared (VNIR) part of the spectrum, incoming radiation is focussed by a telescope onto a detector or array of detectors that convert the radiation into an electronic signal. Different types of detector are available, including photomultipliers, which are bulky but very sensitive, and semiconductor devices such as photodiodes and charge-coupled devices (CCDs). CCDs can be constructed in one-dimensional (linear) or two-dimensional (area) arrays. A two-dimensional array of detectors obtains all the data needed to construct a two-dimensional image essentially simultaneously, in a mode of operation called ‘step-stare’ imaging. In the case of a one-dimensional array of detectors (often called a ‘line scanner’), scanning in the direction perpendicular to the orientation of the array is achieved through the relative motion of the platform carrying the instrument and the Earth’s surface. This mode of operation is called ‘pushbroom imaging’. In the case of a single sensor, or a small number of sensors, two-dimensional scanning is achieved through combining the effects of platform motion with across-track scanning typically by a rapidly rotating mirror. This mode of operation is called ‘whiskbroom imaging’. In the case of an imaging instrument carried by a geostationary satellite, there is no platform motion relative to the Earth’s surface and two-dimensional scanning is achieved by rapidly spinning the satellite about an axis parallel to the Earth’s axis (to give east–west scanning) while slowly rotating a scan mirror (to give north–south scanning). This is often called ‘spin-scan imaging’.

Spatial resolution of a VNIR imager is controlled by diffraction and also by the size of the detector and the magnification of the telescope. Both of these effects give rise to a linear spatial resolution that is proportional to the observing height. The element on the ground that corresponds to a single pixel in the image is called a *rezel* (‘resolution element’). Spectral resolution, i.e. the ability to discriminate between different wavelengths, can be controlled by filters or, if a finer resolution is needed, by a dispersive element such as a prism or diffraction grating.

Radiances detected by VNIR imagers can be significantly altered by the atmosphere, especially in the case of satellite-based instruments, which view through essentially all of the atmosphere. Atmospheric effects are stronger at shorter wavelengths. They can be corrected by physical modelling using the radiative transfer equation, although this approach requires a lot of ancillary data. A simpler approach is possible if calibration targets, of known reflectance, are visible in the image. The simplest method is ‘dark pixel subtraction’, which assumes that the atmosphere contributes additively to the measured radiance and relies on the ability to find pixels in the image corresponding to areas of zero reflectance to estimate the atmospheric radiance.

6.2

Types of VNIR imager

In Section 5.5 we outlined the main types of photographic systems used for remote sensing. In this section we attempt to do the same for electro-optical VNIR imaging systems. Here the task is a much larger one, as there is a far greater diversity of

instruments. For this reason we concentrate on spaceborne systems. With the obvious exception of the very wide swath instruments, these generally have airborne counterparts with spatial resolutions that are finer in proportion to the observing height.

We can roughly categorise spaceborne electro-optical imaging systems according to their spatial resolution and swath width. Figure 6.10 plots about 75 imaging systems, all of which were active in 2011, on this basis. They have been categorised, for the most part somewhat arbitrarily, according to their spatial resolution, although three other categories are included: geostationary imagers, which are carried on geostationary satellites, ocean colour instruments, and hyperspectral imagers (defined here to mean those that record at least 50 contiguous wavebands). The figure identifies one example of each category, each of which is described below.

6.2.1

Very high resolution imagers

For the purpose of Figure 6.10, very high resolution (VHR) imagers are defined as those having a pixel size of 5 m or smaller. Until the end of the twentieth century such resolutions were available from spaceborne imaging systems only for military or similar restricted use, but since the launch of Ikonos in 1999 they have become available commercially.

As an example of this class of instrument we describe GeoEye-1. This is a commercially owned satellite that was launched in 2008, with an orbital altitude of 681 km. It carries a VHR imager that has a swath width of 15.2 km and records in two modes simultaneously: panchromatic mode, with a single waveband at 0.45–0.80 µm and a rezel size of 0.41 m; and multispectral mode, with four wavebands (0.45–0.51 µm, 0.51–0.58 µm, 0.65–0.69 µm and 0.78–0.92 µm) and a rezel size of 1.65 m. The diameter and focal length of the imager's lens are 1.1 m and 13.3 m respectively, and the field of view is steerable up to 60 degrees off nadir, which gives the possibility of stereo observations and of rapid repeat viewing of a particular location. The detectors are linear silicon CCD arrays (so GeoEye uses pushbroom imaging). The panchromatic detector is a single linear array of around 35,000 elements each 8 µm square, while the multispectral detector consists of four linear arrays of around 10,000 elements each 32 µm square. The widths of the detectors are thus around 0.3 m.

An example of GeoEye imagery is shown in Figure 6.11. This kind of instrument is most directly comparable to aerial photography, and it finds similar applications. It is notable that the data volumes can be as high as from aerial photographic mapping systems. Even so, the very high spatial resolution means that the imaged swath is narrow, usually a few tens of kilometres at most.

6.2.2

High resolution imagers

Here we define high spatial resolution to mean rezel sizes between 5 and 50 m. Spaceborne imagers having spatial resolutions in this range have been in operation since the 1980s, and they still constitute the majority of instruments, usually operated by national or international agencies.

As an example of this type of imager we consider the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER). This is a Japanese instrument carried on



Figure 6.11. Extract of true colour GeoEye image showing the centre of Cambridge, UK, recorded on 4 September 2005. (Image reproduced by courtesy of GeoEye.) See also colour plates section.

the Terra satellite launched in 1999. ASTER has been operational since February 2000. It observes ten wavebands in the VNIR, summarised in Table 6.1, and has a swath width of 60 km. (It also measures TIR radiation, as its name implies, in five wavebands.) Two separate subsystems are used to give this spectral coverage: the VNIR subsystem for bands 1–3 and the SWIR ('short-wave infrared') subsystem for bands 4–9. The VNIR subsystem has two telescopes, one pointing vertically downwards (towards the nadir) and focussing radiation onto a three-band detector, and one pointing backwards at an angle of 27.6° to provide stereo viewing. The telescope assembly can be rotated up to $\pm 24^\circ$ about the direction of motion to give the possibility of off-nadir viewing. Spectral separation is provided by filters. The detectors are linear CCD arrays with 5000 elements, providing pushbroom scanning. The SWIR subsystem operates similarly, though it uses a single nadir-viewing telescope which can be rotated up to $\pm 8.5^\circ$ around the direction of motion. The detectors are linear arrays, cooled to 80 K to reduce noise. Again, spectral separation is achieved using filters. ASTER includes on-board calibrators so that the calibrations of the wavebands can be determined.

Figure 6.12 shows an example of ASTER imagery.

Table 6.1. ASTER wavebands in the VNIR region. All bands are observed at nadir except 3b, which is observed at an angle of 27.6° behind nadir

Band	Wavelength (μm)	Rezel size (m)	Note
1	0.520–0.600	15	
2	0.630–0.690	15	
3n	0.760–0.860	15	
3b	0.760–0.860	15	views backwards
4	1.600–1.700	30	
5	2.145–2.185	30	
6	2.185–2.225	30	
7	2.235–2.285	30	
8	2.295–2.365	30	
9	2.360–2.430	30	



Figure 6.12. Extract of FCIR ASTER image showing Tokyo, Japan, on 22 March 2000. (Image by courtesy of NASA/GSFC/METI/ERSDAC/JAROS and the US/Japan ASTER Science Team. <http://asterweb.jpl.nasa.gov/gallery-detail.asp?name=Tokyo>). See also colour plates section.

Table 6.2. Visible and near-infrared bands of MODIS

Band	Wavelength (nm)	Rezel size (m)
1	620–670	250
2	841–876	250
3	459–479	500
4	545–565	500
5	1230–1250	500
6	1628–1652	500
7	2105–2155	500
8	405–420	1000
9	438–448	1000
10	484–493	1000
11	526–536	1000
12	546–556	1000
13	662–672	1000
14	673–683	1000
15	743–753	1000
16	862–877	1000
17	890–920	1000
18	931–941	1000
19	915–965	1000

6.2.3

Medium resolution imagers

Medium resolution imagers (50 to 500 m) are generally similar to the high resolution instruments, but with larger swath widths – typically several hundred kilometres. An example of this type of instrument is the *MSU-50* carried on the Russian Meteor-M satellite. The satellite was launched in 2009 and has an orbital height of 830 km. The instrument records data in three spectral bands, centred at 0.41, 0.48 and 0.63 µm. The detector is again a linear CCD, with 7926 sensors, each of which is 7 µm square. The telescope's focal length of 50 mm thus gives a rezel size of around 116 m and a swath width of around 920 km.

6.2.4

Low resolution imagers

Low resolution imagers have spatial resolutions coarser than 500 m. The disadvantage of comparatively coarse resolution is offset by the advantage of wide-swath imaging and the possibility of obtaining data on continental or ocean-wide scales.

An example of such an imager, although it actually occupies several of the categories that we have defined here, is the Moderate-Resolution Imaging Spectroradiometer (MODIS) that has been carried on board the Terra satellite since 1999 and the Aqua satellite since 2002. Both of these satellites orbit the Earth at a height of 705 km. MODIS has 19 spectral bands in the VNIR region (and a further 17 in the TIR), with spatial resolutions between 250 m and 1000 m and spectral resolutions between 10 and 50 nm. The swath width of the instrument is 2330 km. Table 6.2 summarises the VNIR bands of MODIS.



Figure 6.13. True-colour MODIS image of Morocco, acquired on 3 August 2000. (Image courtesy of Jacques Descloitres, MODIS Land Group, NASA GSFC. Image downloaded from NASA Visible Earth at <http://visibleearth.nasa.gov>.) See also colour plates section.

MODIS is a whiskbroom imager, using a rotating mirror to scan the IFOV over the swath width. The mirror is scanned at 20 r.p.m. The telescope has an objective lens diameter of 178 mm and a focal length of 380 mm. Prisms are used to disperse the incoming radiation, which is then detected by an array of photovoltaic detectors. The detectors themselves are comparatively large, ranging from 135 µm (corresponding to a rezel size of 250 m) to 540 µm (rezel size 1000 m) square. Figure 6.13 shows an example of MODIS imagery.

6.2.5 Ocean colour imagers

Imagers designed particularly to measure the spectral reflectance of the ocean surface are an important sub-category of multispectral imagers. The purpose of these instruments is primarily to measure concentrations of phytoplankton through the presence of chlorophyll, but also to distinguish the presence of suspended sediments in coastal waters. Because the reflectance of the ocean surface is very low (see Section 3.6.1) it is necessary to make rather accurate correction for the effects of atmospheric propagation, especially atmospheric aerosols (Section 4.3). This requires a number of spectral bands with bandwidths rather narrower than are needed for most land surface remote sensing. High spatial resolutions are not generally necessary.

An example of a currently operational ocean colour imager is Ocean Colour Monitor (OCM) 2, carried on board the Indian satellite Oceansat-2. Oceansat-2 was launched in 2009 and orbits the Earth at a height of 720 km. OCM2 has eight wavebands, summarised in Table 6.3, spanning the range from 404 to 885 nm. The spatial resolution is 360 m and the swath width is 1420 km.

In fact, bands 10–19 of MODIS (Table 6.2) are also optimised for ocean colour observations. Figure 6.14 shows MODIS ocean colour imagery.

Table 6.3. Wavebands of OCM-2

Band	Wavelength (nm)
1	404–424
2	431–451
3	476–496
4	500–520
5	546–566
6	610–630
7	725–755
8	845–885

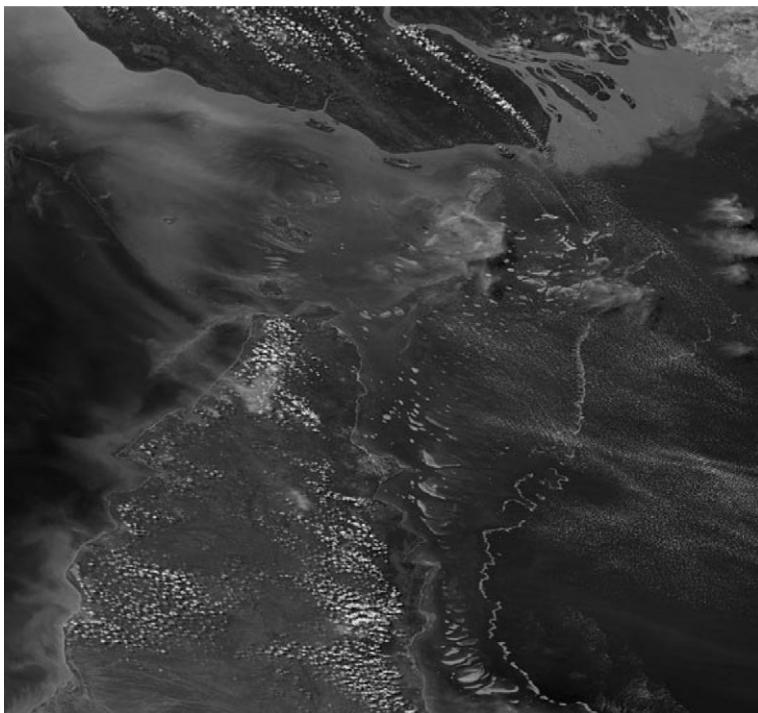


Figure 6.14. Extract of MODIS ocean colour imagery. The image was acquired on 9 August 2011 and shows Papua New Guinea, the Torres Strait, and the Cape York Peninsula, Australia. The Great Barrier Reef is clearly visible to the east of the peninsula. (Original image downloaded from <http://oceancolor.gsfc.nasa.gov/> and reproduced by courtesy of NASA.) See also colour plates section.

6.2.6 Hyperspectral imagers

Hyperspectral imagers provide significantly greater spectral resolution than the categories of imager that we have discussed so far. As a somewhat arbitrary criterion, we define the term ‘hyperspectral’ to mean that the instrument can resolve at least 50 contiguous wavebands in the VNIR region. This means that many more spectral features can be resolved, potentially increasing the scope for classifying materials on the basis of their



Figure 6.15. Hyperion imagery from East Anglia, UK, recorded on 21 October 2008. The strip is 7.7 km wide and 108 km long and extends from the Norfolk coast (left) to near Saffron Walden (right). See also colour plates section.

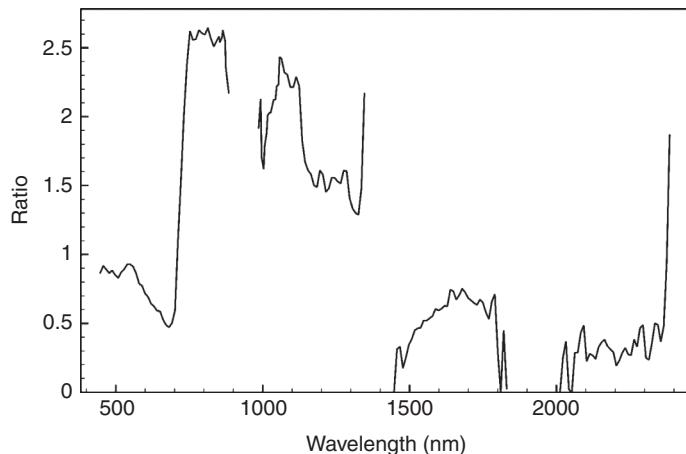


Figure 6.16. Ratio of the radiances recorded in the image of figure 6.15 for a pixel representing a standing agricultural crop and a pixel representing bare earth. The gaps in the data are caused by atmospheric absorption (compare figure 4.3).

reflectance spectra (see Section 11.3.3). Although the applications of hyperspectral imaging are by no means limited only to this, a particularly important use is in the identification of minerals.

As an example of a spaceborne hyperspectral imager we consider the instrument *Hyperion*. This is carried on board the NASA EO-1 satellite, launched in 2000 with an orbital height of 705 km. Although the mission was only scheduled to last for two years, it was formally extended and (at the time of writing) is still operational. Hyperion uses pushbroom imaging, and a diffraction grating to achieve spectral dispersion, with a spectral resolution of around 10 nm between 360 and 2580 nm. It records data in 242 spectral channels (see Figure 6.16). The pixel size is 30 m and the swath width is 7.7 km. Figure 6.15 shows an example of Hyperion imagery, and Figure 6.16 illustrates its spectral content.

6.2.7 Geostationary imagers

Finally, we recall the spin-scan imaging technique described in Section 6.1.2 and commonly employed for sensors carried on geostationary satellites. These sensors, which are used mainly for meteorological imaging, normally combine fairly broad-band coverage of the visible spectrum with TIR channels. A typical example is the Spinning Enhanced Visible and Infrared Imager (SEVIRI) carried on board Meteosat Second Generation (MSG) satellites in geostationary orbit (see Chapter 10) at a longitude of 0°. SEVIRI is a

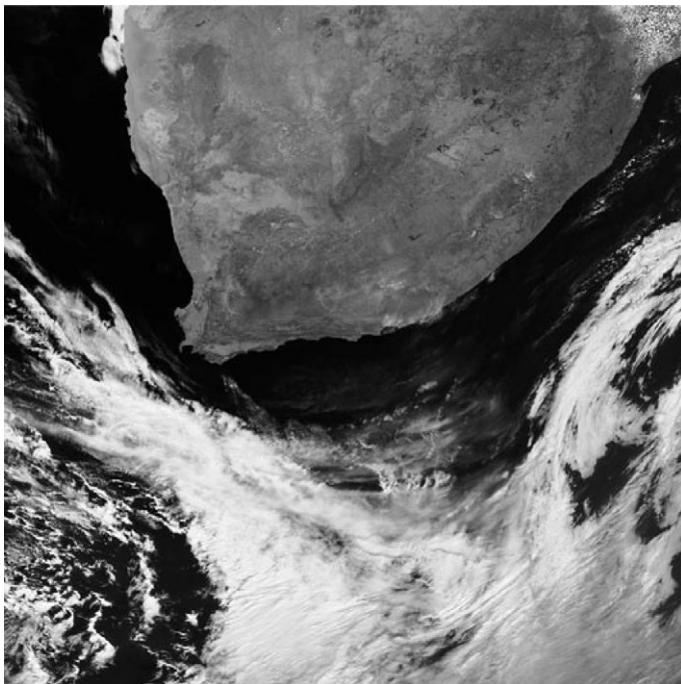


Figure 6.17. Extract of SEVIRI HRV imagery showing southern Africa. The image was acquired on 25 August 2011. (Image courtesy of Dundee Satellite Receiving Station.) Permission needed.

successor to the earlier VISSR (visible and infrared spin-scan radiometer) instrument. It records data in both the VNIR and TIR regions. There are three fairly narrow VNIR bands (0.56–0.71, 0.74–0.88 and 1.50–1.78 µm), each using three photodiodes, and a ‘high resolution visible’ (HRV) band with a spectral response from 0.6 to 0.9 µm, using nine photodiodes. The telescope has an objective diameter of 51 cm and a focal length of 5.37 m.

SEVIRI uses spin-scanning, the satellite spinning at 100 r.p.m. The scan mirror moves in steps of 125.8 microradians between each scan, corresponding to a separation of about 4.5 km between scan lines at nadir. The pixel sizes are thus around 1.5 km for the narrow spectral bands and 0.5 km for the HRV band although these figures should be approximately doubled to give the effective spatial resolutions (Figure 6.17). It requires 1249 rotations of the instrument to image the whole of the Earth’s disc, after which the instrument is calibrated against internal calibrators before the imaging cycle begins again. The repeat cycle is about 15 minutes.

SEVIRI imagery is illustrated in Figure 6.17.

Summary

VNIR imagers show a wide diversity of type, differentiated mainly on the basis of spatial and spectral resolution. Instruments carried on satellites in low Earth orbits (i.e. a few hundred kilometres above the surface) can have spatial resolutions ranging from less than

a metre to a few kilometres. There is a strong correlation between the spatial resolution and the width of the imaged swath, such that the number of pixels across the swath width is generally between about 1000 and 20 000, and this means that the very high resolution imagers view rather small areas of the Earth's surface. Conversely, coarse-resolution imagers can have swath widths of 1000 km or more, suiting them for synoptic observations of the Earth's surface. Because of their much greater height above the surface, VNIR imagers carried on geostationary satellites achieve spatial resolutions of a few kilometres, but they can acquire an image of the whole of the visible Earth disc in around 30 minutes.

Most VNIR imagers operate in a few spectral bands, with a wavelength bandwidth of the order of 100 nm. Instruments designed to measure ocean colour need somewhat higher spectral resolution, with bandwidths of the order of 10 nm. Hyperspectral imagers detect radiation in many (of the order of 100) contiguous spectral bands.

6.3

Major applications of VNIR images

Visible and near-infrared imagery, normally with quantitative or even calibrated radiance values, has found enormous application in remote sensing. Part of its popularity may be ascribed to the greater ease with which such data may be interpreted, since the wavelength range corresponds largely to the sensitivity range of the human eye, but to a great extent also many important processes modulate the image radiance in this range of wavelengths. Since a VNIR image is similar to an aerial photograph, many applications of the former technique can be carried over to VNIR imagery. In this section we cannot discuss any of these applications at great length, but we can at least illustrate something of their range. A good idea of the current applications of VNIR and other images can be obtained by scanning the last few years of the main remote sensing journals.

The most obvious set of applications is probably to land-surface mapping, for example for cartography, the delineation and classification of water bodies (including floods and wetlands; Rebelo, Finlayson and Nagabhatla. (2009)), volcanoes, snow and ice (Rees 2006), coastal, agricultural, forestry (including forest fires) and urban areas, road networks (Charalambos and Suya 2010), archaeology (Parcak 2009), geological and geomorphological characteristics (Smith and Pain 2009), and so on. VNIR imagery has proved particularly useful in the construction of global databases of land cover (e.g. Figure 6.18). The geometrical characteristics of terrestrial imagery can be used to derive topographic data and hence generate DEMs (e.g. Figure 6.19). The development of above-ground vegetation biomass (the total organic matter per unit area of the Earth's surface) can be monitored, and this provides the ability to observe changes in the extent and health of vegetation (e.g. Figure 6.20), for example for the commercial estimation of crop and forestry yields, the prediction of crop failure and drought, and monitoring of the impacts of pollution and climate change (Cohen and Goward 2004). Clearly, the scale at which such mapping can be carried out is dependent upon the spatial resolution of the imagery. However, since the ability to distinguish one type of land cover from another is normally reliant on the ability to distinguish their reflectance spectra, the scale is also governed by the spectral resolution and the sensitivity of the data. In some cases, notably

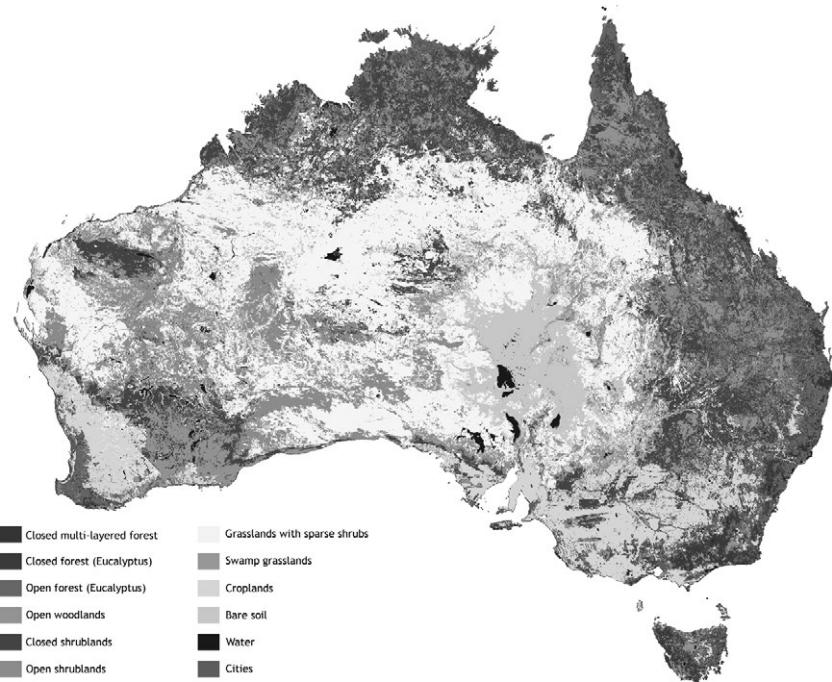


Figure 6.18. Part of the Global Land Cover 2000 dataset, constructed from data collected by the Vegetation instrument on the Spot-4 satellite. The dataset has a spatial resolution of 1 km. (Data downloaded from <http://bioval.jrc.ec.europa.eu/products/glc2000/products.php>. Mayaux and Bossard (2000).) See also colour plates section.

vegetation, water bodies and snow, the reflectance spectra are sufficiently characteristic that their delineation can be carried out with minimal ancillary information (see Section 3.5.1). For example, a pixel that shows low reflectance in the visible part of the spectrum, especially for red light, and a high reflectance in the near-infrared region, is almost certain to contain a high proportion of green-leaved vegetation. Similarly, a pixel that shows high reflectance across the VNIR region is likely to represent optically thick cloud, while a high reflectance in the visible region coupled with a low reflectance around $1.5\text{ }\mu\text{m}$ is likely to represent snow cover. In most cases, however, the association of a particular land-cover type with the measured properties of a pixel is more difficult. This procedure is known as *image classification*, and is discussed in Chapter 11.

Water bodies can also be studied using VNIR imagery. The depth of a shallow water body (river, coastal waters or lake) can be estimated by comparing the reflectance in two or more spectral bands (Gao 2009). This method relies on the variation with wavelength of the attenuation coefficient, so that the relative contributions from the subaqueous surface and the water itself will differ at different wavelengths; although because of uncertainties in the attenuation coefficient and the subaqueous reflection coefficient, the method generally requires calibration. The most useful spectral bands to use are green and near-infrared, since these give the largest difference in attenuation coefficient. Ocean colour measurements (McClain 2009), in which sensitive measurements of the ocean reflectance are made in a number of spectral bands, can be used to identify different

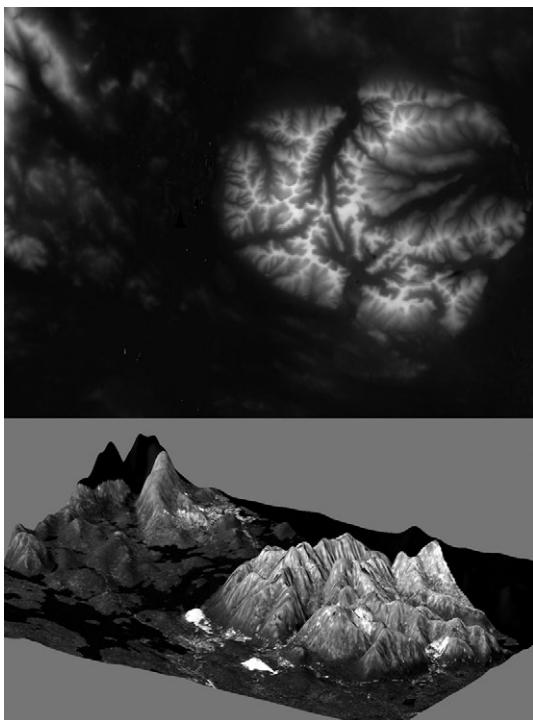


Figure 6.19. Above: Greyscale representation of a DEM derived from ASTER imagery. Below: Terrain visualised by draping a true-colour Landsat image over the DEM. (Rees 2012). See also colour plates section.

bodies of water, for example to distinguish sediment-laden coastal waters from deep ocean waters and to estimate ocean currents from variations in turbidity. Marine phytoplankton concentrations can be estimated by the increased reflectance, caused by the presence of chlorophyll, in the 0.4–0.5 µm band (Figure 6.21). The effect is a small one, so that sensitive instruments and particularly accurate correction for atmospheric propagation effects are needed. Over the open ocean, suspended sediments can be ignored (these are so-called ‘type I waters’) in estimating chlorophyll concentrations, and the calculation is normally accurate to within a factor of 2, but over shallow waters suspended sediments, especially those with significant reflection in the yellow band, severely complicate the problem (‘type II waters’).

Clouds can be delineated and monitored from VNIR imagery, measuring the extent of cloud cover, its height and type, as an aid to meteorological investigations. By tracking the motion of clouds, wind speeds can be estimated. Atmospheric aerosols can also be measured using VNIR imagery. This application shares some characteristics with the measurement of ocean colour, since the aim in both cases is to identify the contribution to the measured radiance from radiation that has been scattered by suspended material. Unless the optical thickness of the aerosol is particularly great, the contribution to the measured radiance from the Earth’s surface can dominate over the atmospheric signal. For this reason, retrievals of aerosol optical thicknesses are easier over dark surfaces (such as

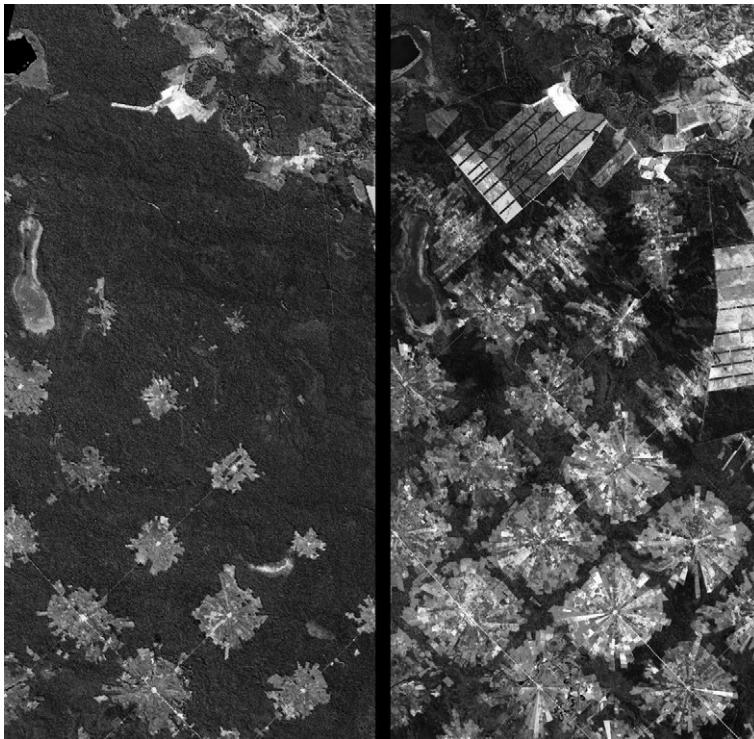


Figure 6.20. Landsat (left) and ASTER (right) false-colour infrared images of the same area of Bolivian forest, in 1986 and 2001. The square areas are villages and soybean plantations. (Images downloaded from <http://asterweb.jpl.nasa.gov/gallery-detail.asp?name=bolivia> and reproduced by courtesy of NASA/GSFC/METI/ERSDAC/JAROS and US/Japan ASTER Science Team.) See also colour plates section.

water bodies). Multispectral observations, with high spectral resolution, improve the ability to detect aerosols since the measured data can be matched to the known spectral reflectance properties of different types of aerosol. Polarisation-sensitive detection is also useful (Waquet *et al.* 2009), although very few such sensors have been deployed from space.

Low resolution VNIR imagery has a very wide range of applications. For example, over land it can be used for measuring albedo, as an input to climate modelling, and a number of satellite instruments have been optimised for this function. Over oceans, applications include the monitoring of large-scale ocean circulation, sediment transport from rivers, and so on.

Summary

VNIR imagery is probably the most widespread and familiar form of remote sensing data and it has a huge range of applications. Over land surfaces, many applications can be loosely classed as surface mapping, and repeated mapping over time allows for change detection. The fundamental property of the surface that can be determined from VNIR imagery is its reflectance, defined spectrally if multi- or hyperspectral imagery is available,

but very many physical parameters can be inferred from the reflectance. Some imagery is also suitable for determining topographic relief and is used for construction of Digital Elevation Models (DEMs). Over water surfaces, multispectral VNIR imagery reveals the colour of the water. If the water is sufficiently deep this is mostly controlled by suspended sediments or chlorophyll and can be used to estimate their quantity. In the case of shallow water, the bottom reflectance also contributes, and this can be used to estimate the depth of the water body. Clouds can be identified and classified in VNIR imagery. The optical thickness of atmospheric aerosols can be measured, at least over sufficiently dark surfaces.

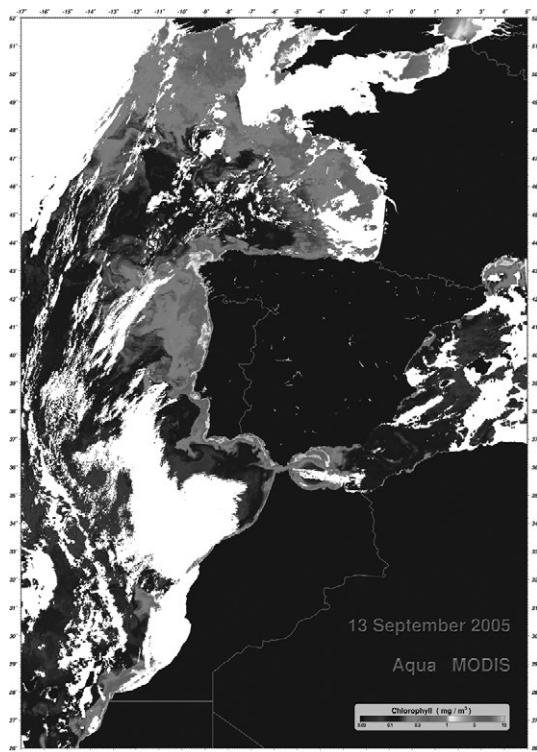


Figure 6.21. Chlorophyll concentrations estimated from MODIS imagery of the eastern North Atlantic. (Image reproduced by courtesy of NASA. Source: http://oceancolor.gsfc.nasa.gov/cgi/image_archive.cgi?c=CHLOROPHYLL.) See also colour plates section.

6.4

Thermal infrared imagers

6.4.1

Detectors

Detectors for TIR radiation can be divided into two classes – quantum detectors and thermal detectors. In a quantum detector, a photon interacts directly with the detector material to change the energy of an electron, whereas in a thermal detector, the radiation

6.4 Thermal infrared imagers

is detected as a result of a change in the electrical properties of the material arising from a change in its temperature.

As we remarked in Chapter 2, the TIR region of the electromagnetic spectrum occupies the wavelength range between about 3 and 15 μm . The energy of a TIR photon is thus about 0.1–0.4 eV, significantly less than for VNIR radiation, and this imposes some difficulty in making suitable photodiodes (quantum detectors). We noted in Section 6.1.1 that germanium photodiodes respond to wavelengths up to 1.7 μm , lead sulphide up to about 3 μm and indium antimonide to about 5 μm . More exotic semiconductors, having even smaller band-gaps, can be devised. Two common ones for TIR detection are mercury cadmium telluride ($\text{Hg}_{0.2}\text{Cd}_{0.8}\text{Te}$, also referred to as MCT) and mercury-doped germanium (Ge:Hg), both of which respond to wavelengths up to about 15 μm . It is, however, usually necessary to cool a quantum detector of TIR radiation, especially one designed to operate at the long-wavelength end of the region. This increases the sensitivity of the detector by reducing the number of photons generated by the sensor itself. Cooling is normally provided using liquid nitrogen, which has a maximum temperature of 77 K, or liquid helium (30 K).

In contrast with the quantum detectors, thermal detectors provide very broad spectral range, but at the expense of much lower sensitivity and longer response times. The three main types of thermal detector are thermistor bolometers, thermocouples and pyroelectric devices. The *thermistor bolometer* is a simple device that consists, in essence, of a material (usually carbon, germanium or a mixture of metal oxides) whose resistance varies with temperature. A *thermocouple* uses the *Seebeck effect*, in which a potential difference is generated across a pair of junctions between dissimilar metals when the junctions are held at different temperatures. To amplify the signal, thermocouples are often connected in series as a *thermopile*. A *pyroelectric device* is a crystal that undergoes a redistribution of its internal charges as a result of a change in temperature. Charge separation occurs at the surfaces of the crystal, resulting in a potential difference that can be amplified and detected.

In practice, thermal detectors are little used in remote sensing instruments except where sensitivity is needed at wavelengths significantly longer than about 15 μm .

6.4.2 Thermal infrared imaging

The principles by which TIR imagers form images are similar to those for VNIR imagers, that is, both mechanical scanning and arrays of detectors are used (see Section 6.2.2). As was mentioned earlier, TIR and VNIR functions are often combined in the same instrument.

6.4.3 Spatial resolution

The factors that determine the spatial resolution of a TIR imager are similar to those for a VNIR imager, so much of the discussion in Section 6.1.3 is applicable. However, we should note that, since the wavelength of TIR radiation is of the order of ten times longer than for VNIR radiation, the spatial resolution is generally somewhat coarser. For example, the Enhanced Thematic Mapper + instrument carried on the Landsat-7 satellite has six VNIR bands, each of which has a rezel size of 30 m (it also has a panchromatic band with a rezel size of 15 m), and a TIR band with a rezel size of 60 m.

6.4.4

Spectral resolution and sensitivity

TIR imaging does not normally require particularly high spectral resolution. In practice, the wavelength range from 3 to 15 μm is usually split into a small number (often only one or two) of channels with a typical width of 1 μm , defined by filters. These channels are generally centred at about 4 μm and about 10 μm , thus avoiding the strong water vapour absorption feature at 6–7 μm (see Figure 4.4).

The intrinsic sensitivity of a radiometric observation, at a given wavelength, to black-body radiation at temperature T , can be defined in terms of the Planck radiation law (Equation 2.31) as

$$S = \frac{T}{L_\lambda} \frac{\partial L_\lambda}{\partial T}, \quad (6.3)$$

i.e. the ratio of the fractional change in the black-body radiance to the fractional change in the absolute temperature. This function is plotted in Figure 6.22 for a temperature of 280 K (a typical terrestrial temperature).

As the figure shows, the sensitivity S is about three times greater at a wavelength of 4 μm than at 10 μm , and for this reason the 3–5 μm band could be preferred to the 8–14 μm band. However, as we recall (e.g. Figure 2.12), the spectral radiance L_λ for a black body at a temperature of 5800 K, i.e. roughly that of sunlight, is about 50 times greater at 4 μm than at 10 μm . Thus, in fact, measurements made in the 3–5 μm band are at risk of ‘contamination’ from reflected solar radiation, and for this reason this band is less useful than the 8–14 μm band except for night-time measurements and for measurements of volcanoes.

Since the purpose of a TIR observation is to measure the brightness temperature T_b of the radiation incident upon it, we should consider how the power detected by the instrument varies with T_b . To some extent this is indicated by the sensitivity S defined in Equation (6.3), but we must also consider the effect of the filter used to define the spectral band. For an instrument that collects radiation from a solid angle $\Delta\Omega$ over

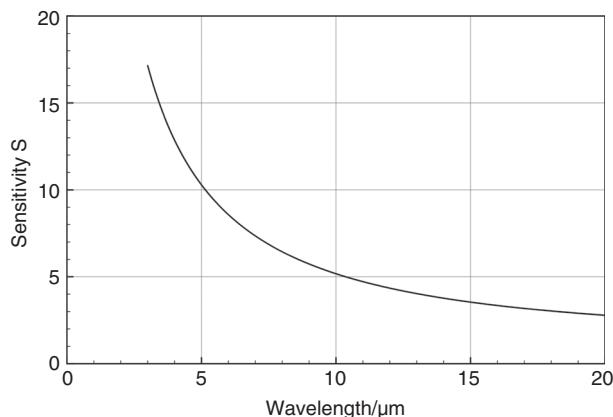


Figure 6.22. The intrinsic sensitivity S of black-body radiation at a temperature of 280 K, defined as the ratio of the fractional change in spectral radiance to the fractional change in temperature.

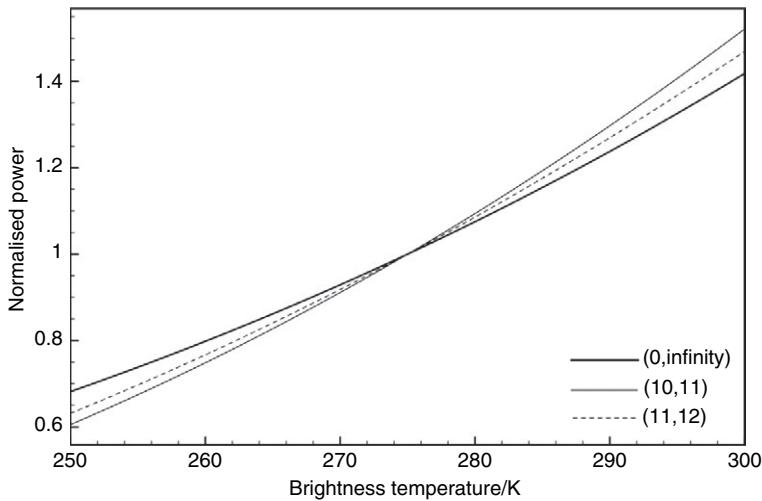


Figure 6.23. Total power received from a black body between fixed wavelength limits. The limits, in μm , are given in the key. In each case the total power is normalised to the value at a brightness temperature of 275 K. The heavy curve is a fourth-power relationship corresponding to the full spectrum of black-body radiation.

an area A , the total power P received from black-body radiation of brightness temperature T_b can be written as

$$P = 2hc^2 A \Delta\Omega \int_0^\infty \frac{f(\lambda)d\lambda}{\lambda^5 (\exp(hc/\lambda kT) - 1)}, \quad (6.4)$$

where $f(\lambda)$ is a filtering function that defines the response of the instrument to radiation of wavelength λ . If $f(\lambda) = 1$ for all wavelengths (no filtering), the integral is just

$$\frac{1}{15} \left(\frac{\pi k T_b}{hc} \right)^4$$

and the received power varies as T_b^4 , in accordance with the laws for black-body radiation. Next we adopt a very simple model for $f(\lambda)$, by supposing that it is equal to 1 for wavelengths between λ_{\min} and λ_{\max} , and zero otherwise. This represents a simple filter. Figure 6.23 shows the variation of P with T_b for some different values of the limits λ_{\min} and λ_{\max} . When the limits are 0 and infinity the instrument receives the whole of the black-body spectrum so that P is proportional to the fourth power of T_b , but for narrower limits the value of P increases more strongly with brightness temperature. This shows that the fraction of the total radiance between the wavelength limits increases with brightness temperature over this range.

6.4.5

Atmospheric propagation and correction

As we have already remarked, a TIR radiometer measures the brightness temperature of the radiation reaching the sensor, whereas the quantity that is usually required is the brightness temperature of the radiation leaving the surface. For spaceborne observations, where the full

effect of the atmosphere is felt, the difference between these two temperatures can be as large as 10 K, mainly as a result of water vapour, and correction is generally necessary. In this section we consider the principal techniques for making this correction.

Atmospheric correction of TIR imagery is generally performed by one of three methods. The first of these is physical modelling, for example using the LOWTRAN (Kneizys *et al.* 1988) or MODTRAN (Berk *et al.* 2005) models introduced in Section 6.1.5. Because the main correction is due to water vapour, which is highly variable both spatially and temporally, physical modelling is rather unsatisfactory unless a detailed characterisation of the atmosphere is available. This can be obtained by local radiosonde observations or by airborne or spaceborne atmospheric sounding (Section 6.7).

The second method is the *split-window technique* (e.g. Wan and Dozier 1996). In this method, which is in widespread use, the brightness temperatures T_{b1} and T_{b2} are measured in two different but closely spaced spectral bands, for example one centred at 11 μm and one at 12 μm . The brightness temperature T_{b0} of the radiation leaving the surface is then modelled as a linear combination of the measured brightness temperatures:

$$T_{b0} = a_0 + a_1 T_{b1} + a_2 T_{b2} \quad . \quad (6.5)$$

The coefficients a_0 , a_1 and a_2 are determined empirically. Because of the contribution from reflected sky radiation, the coefficients have different values during the day and during the night. The technique is reasonably accurate (to within 0.5 K) if the emissivity of the surface is constant, so it is reliable for sea surface temperature measurements, but over land surfaces its usefulness is more limited. With an increased number of spectral bands, such as those available from MODIS, Equation (6.5) is replaced by a more general linear combination of brightness temperatures, and the technique is generally more accurate. As we have already noted, it is possible to estimate surface emissivity from thermal inertia measurements.

Finally, we mention the *two-look technique*. In this approach, the instrument views each rezel twice, at two different incidence angles. For example, the AATSR (advanced along-track scanning radiometer) instrument carried by the ENVISAT satellite, launched in 2002, uses a conical scanning technique such that rezels are viewed both at nadir (i.e. vertically) and at an angle of 52° to the nadir (Smith *et al.* 2012). Since $1/\cos(52^\circ) \approx 1.6$, the oblique path looks through about 1.6 times as much atmosphere as the vertical path and the atmospheric corrections are correspondingly larger. By comparing the two brightness temperatures, the atmospheric correction can be estimated and eliminated.

Figure 6.24 illustrates schematically the principle of two-look correction. We can derive a simplified model as follows. We assume that the brightness temperature of the radiation leaving the surface is T_{b0} , independent of direction, and that reflection of atmospheric radiation is insignificant. The brightness temperatures of the radiation reaching the sensor at positions 1 and 2, T_{b1} and T_{b2} respectively, can be found from the radiative transfer equation (Section 3.5). To simplify this equation we assume that scattering of radiation is unimportant (this is a very good assumption), that the physical temperature of the absorbing material has a constant value of T_a , and that T_{b0} and T_a are sufficiently similar that the dependence of the function B_f on the temperature T (Equation 3.92) can be assumed to be linear. With these simplifying assumptions, and despite the fact that the Rayleigh–Jeans approximation is not valid here, the solution of the radiative transfer equation is given by Equation (3.72). Thus we may put, approximately,

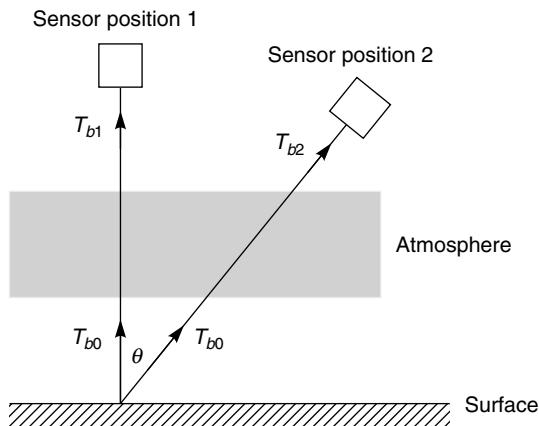


Figure 6.24. Two-look observation for atmospheric correction of a thermal infrared measurement.

$$T_{b1} = T_{b0} \exp(-\tau) + T_a(1 - \exp(-\tau)), \quad (6.6)$$

where τ is the optical thickness of the atmosphere for a ray propagating vertically through it. The corresponding equation for the oblique path is obtained by replacing τ by the optical thickness for this path. Provided that the angle θ is not too close to 90° , this is given by $\tau \sec(\theta)$, and so

$$T_{b2} = T_{b0} \exp(-\tau \sec \theta) + T_a(1 - \exp(-\tau \sec \theta)). \quad (6.7)$$

Treating T_{b1} , T_{b2} and T_a as known variables, these two equations can be solved to find τ and T_{b0} .

Summary

The longer wavelength of thermal infrared (TIR) radiation implies that photons are significantly less energetic than those of VNIR radiation, and more difficult to detect. Semiconductor detectors are usually used, and are often cooled to reduce self-generated noise. The spatial resolution of TIR imagers is generally somewhat coarser than that of corresponding VNIR imagers, again as a consequence of the longer wavelength. Scanning is usually achieved by whiskbroom imaging.

The part of the spectrum that is useful for TIR imaging extends from a wavelength of about $3 \mu\text{m}$ to $15 \mu\text{m}$, although because of the strong water vapour absorption feature at $6\text{--}7 \mu\text{m}$ it is convenient to split it into a short-wavelength region of around $3\text{--}5 \mu\text{m}$ and a long-wavelength region of around $8\text{--}15 \mu\text{m}$. The short-wavelength region is more sensitive to changes in brightness temperature but can be ‘contaminated’ by solar radiation and is generally only used at night. Typical spectral resolutions of TIR imagers are 0.1 to $1 \mu\text{m}$.

As with VNIR radiation, radiances detected by TIR imagers can be significantly altered by the atmosphere. Atmospheric effects can be corrected by physical modelling using the

radiative transfer equation, although again this approach requires a lot of ancillary data. More usual is the ‘split window’ approach, in which measurements of at-satellite brightness temperature are made at two or more wavelengths and used to derive an empirical correction for the atmosphere. Another approach is the ‘two-look method’, in which the same point on the Earth’s surface is viewed both at nadir and obliquely through the atmosphere. This allows the contribution from the atmosphere to be estimated.

6.5

Types of TIR imager

Imaging instruments in the TIR region exhibit much less diversity than VNIR imagers. Since spectral resolution is not particularly important for TIR imagers, they can be characterised by their spatial resolution. Instruments deployed from geostationary orbit typically achieve spatial resolutions of a few kilometres, while those operated from low Earth orbit typically achieve spatial resolutions of the order of 0.1 to 1 km. We consider three examples.

6.5.1

High resolution TIR imager

The example that we consider in this category is the Enhanced Thematic Mapper+ (ETM+) carried on the Landsat-7 satellite, already mentioned in Section 6.3.3 above. Technical details of the satellite and its instrument are given in the *Landsat Data Users’ Handbook* (Irons 2011). The satellite was launched in 1999 in an orbit with a nominal altitude of 705 km, and although the ETM+ suffered a minor technical failure in 2003 it is still (September 2011) collecting data. It is a whiskbroom sensor, recording data in seven VNIR bands as well as in one TIR band. Incoming TIR radiation is collected by a telescope with an objective diameter of 40.6 cm and a focal length of 2.44 m, and focussed onto a ‘cold focal plane’ (maintained at a temperature of 91 K) where there is, amongst other things, an array of eight mercury cadmium telluride photoconductive detectors. These are 0.208 mm square, giving an IFOV of 85 microradians, or 60 m at a distance of 705 km. The IFOV is scanned perpendicularly to the flight direction by a mirror oscillating at 6.997 Hz. Since the satellite moves at a speed of around 6.7 km s^{-1} with respect to the Earth, it advances about 960 m during each full (two-way) swing of the mirror. Thus, during a one-way sweep of the mirror the satellite advances 480 m, equal to projected width ($8 \times 60 \text{ m}$) of the detector array. The scan mirror oscillates through an angle of $\pm 7.7^\circ$, giving a swath width of about 190 km, although data are collected from a swath of 185 km. The spectral response of the TIR detectors is defined by a filter to give a bandpass (between the half-maximum values) of 10.4 to 12.5 μm .

Figure 6.25 shows an example of TIR imagery collected by the Landsat-7 ETM+.

6.5.2

Medium resolution TIR imager

The example that we consider here is the Advanced Very High Resolution Radiometer (AVHRR). This is the name of a series of instruments that have been carried on board the

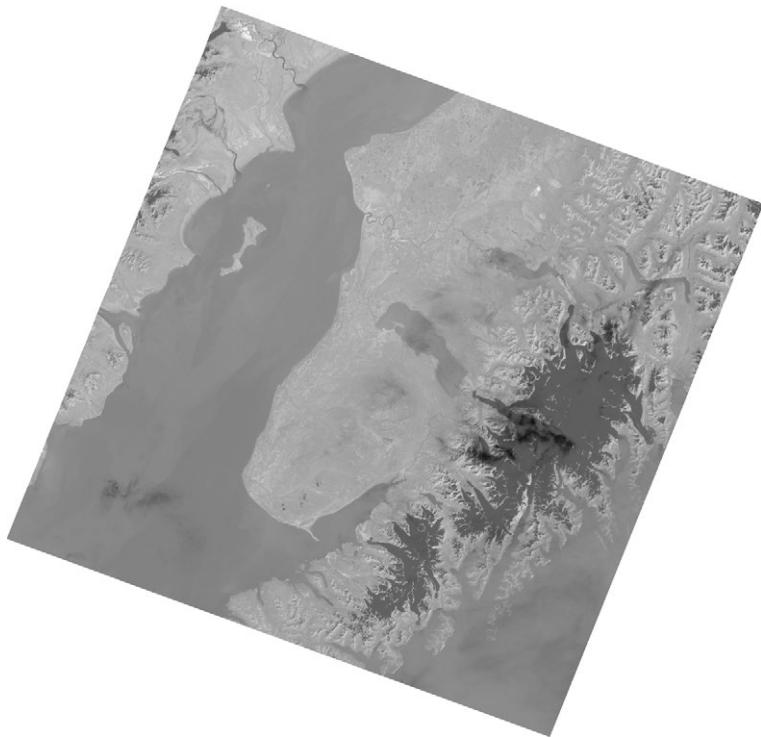


Figure 6.25. Band 6 (TIR) Landsat-7 ETM+ image. The image covers an area of 185 km square and was collected on 30 July 2002. It shows Cook Inlet and the Kenai Peninsula, Alaska. The darker (colder) area towards the right is the Harding Icefield and the darker bands crossing it are clouds.

National Oceanographic and Atmospheric Administration (NOAA) Polar Orbiting Environmental Satellites (POES), and collecting data continuously, since 1981. At least two POES satellites, which orbit the Earth at a nominal altitude of 833 km, are in operation at any time (at the time of writing, six satellites are operational). AVHRR can be regarded as the predecessor of MODIS, although both have been in operation simultaneously since 2000.

The current generation of AVHRR instrument is the AVHRR/3, first deployed on the NOAA-15 POES satellite from 1994. This whiskbroom scanner has three VNIR and three TIR channels. One of the latter responds to wavelengths between 3.55 and 3.93 μm using an indium antimonide detector, while the other two have wavebands of 10.3–11.3 μm and 11.5–12.5 μm respectively, using mercury cadmium telluride detectors. The incoming radiation is focussed onto the detectors by a telescope with an aperture of 203. The field of view is scanned by a mirror rotating at 360 r.p.m., so that the satellite advances approximately 1.1 km between scans, and this is the effective spatial resolution of the instrument. The scan angle of $\pm 55^\circ$ gives a wide swath of around 2500 km (Figure 6.26).

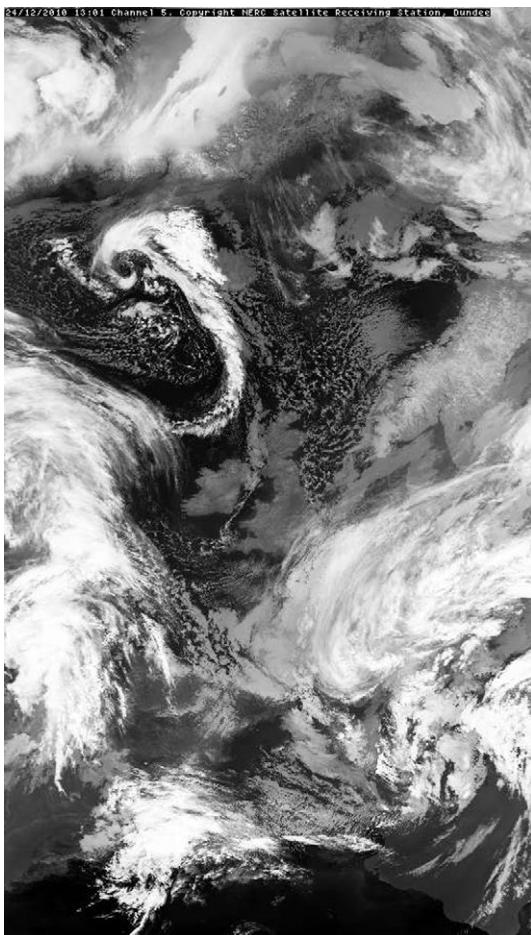


Figure 6.26. Single swath of AVHRR imagery (channel 5: 11.5–12.5 μm) recorded on 24 December 2010. The greyscale in this image is inverted, such that higher brightness temperatures are represented by darker shades of grey. This means that cold high-altitude clouds appear light grey or white, which is helpful for human interpretation of the image since we are used to the idea of clouds appearing white. (Image reproduced by courtesy of Dundee satellite station.)

6.5.3

Geostationary TIR imager

As an example of a TIR imager operated from a geostationary satellite, we consider the Japan Advanced Meteorological Imager (JAMI) carried on the MTSAT-2 satellite. This instrument has three TIR imaging channels (3.5–4.0 μm , 10.3–11.3 μm and 11.5–12.5 μm) as well as an infrared channel at 6.5–7.0 μm and a VNIR channel. Radiation is collected by a telescope with an aperture of 311 mm and a focal length of 895 mm and focussed onto photovoltaic mercury cadmium telluride infrared detectors. The size of the detectors is 50 μm square, giving an angular resolution of 56 μrad and hence a linear

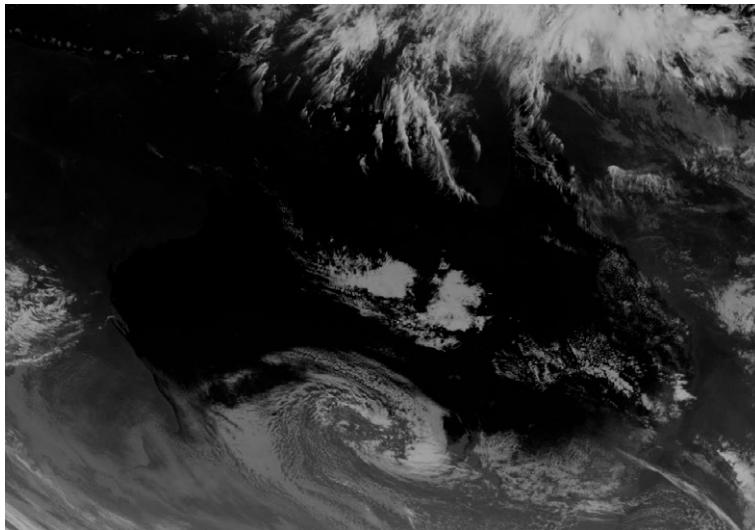


Fig. 6.27. Extract of JAMI thermal infrared imagery showing Australia. 0600 7 September 2011. Again, the greyscale has been inverted. **Permission needed from Dundee.**

resolution of 2 km at nadir. The nadir sampling interval, and effective resolution, is 4 km. MTSAT-2 has a longitude of 145° E. It was launched in 2006, but did not become operational until 2010. Figure 6.27 shows an example of TIR imagery from JAMI.

Summary

TIR imagers show less diversity than VNIR imagers, and can be differentiated mainly by their spatial imaging characteristics. Imagers carried on satellites in low Earth orbit achieve spatial resolutions ranging from around 50 m to 1 km, and resolutions of a few kilometres are achieved from imagers on geostationary satellites. TIR imagers are often combined with VNIR imagers.

6.6

Major applications of thermal infrared images

A calibrated TIR image indicates the brightness temperature of the radiation reaching the sensor. Three factors contribute to this brightness temperature: the physical temperature of the surface being sensed, its emissivity, and atmospheric propagation effects. In most cases, TIR images are used to deduce the surface temperature, and, broadly speaking, we may classify their applications into those in which the surface temperature is governed by man-made sources of heat, and those in which the heating occurs naturally. In the former case, the technique has been used from airborne platforms to determine heat losses from buildings and other structures. It is most useful to perform this kind of survey just before dawn, so that the effects of solar heating will have had the greatest possible

time to decay. The technique has also been used, for example, to monitor plumes of hot water generated by power stations.

In the latter case, the principal applications of TIR images are to the identification of clouds and to the measurement of surface temperatures and thermal inertia. These are discussed in greater detail below. Other applications include the measurement of soil moisture and water stress, identification of frost hollows and crop types, and so on. Thermal infrared imagery can also be used to detect fires, for example forest wildfires, and to estimate the power radiated by them (Wooster, Zhukov and Oertel 2003). One difficulty in the latter task is caused by the fact that the bandwidth of the TIR sensor is generally much narrower than the spectrum over which the fire emits significant quantities of radiation, so it is necessary to estimate the fraction of the radiance that is not detected from the part that is detected. This can be carried out either using an empirical relationship, or by estimating the temperature of the fire, and hence its radiation spectrum, from measurements made in two wavelength bands.

6.6.1 Earth surface temperature

The land surface temperature (LST) and sea surface temperature (SST) are quantities of obvious meteorological and climatological importance, and their determination forms one of the primary motivations for collecting TIR imagery (Jimenez-Munoz *et al.* 2009, Merchant *et al.* 2009). As a first step, it is usually necessary to calculate the brightness temperature of the radiation detected at the instrument. This is quite straightforward if the recorded data are calibrated to radiance values. As we saw in Chapter 2, the spectral radiance at a single wavelength λ of radiation with a brightness temperature T is given by

$$L_\lambda = \frac{2hc^2}{\lambda^5(e^{hc/\lambda kT} - 1)}. \quad (2.32)$$

However, the instrument actually measures the *average* value of the spectral radiance over a finite range of wavelengths and in this case we can, more generally, put

$$\bar{L}_\lambda = \frac{K_1}{(e^{K_2/T} - 1)}, \quad (6.8)$$

where K_1 and K_2 depend on the spectral response of the waveband. For example, the TIR band of the Landsat-7 ETM+ instrument has values of $K_1 = 666.1 \text{ W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$ and $K_2 = 1282.7 \text{ K}$. Equation (6.8) is simply inverted to give

$$T = \frac{K_2}{\ln(K_1/\bar{L}_\lambda + 1)}, \quad (6.9)$$

so that the at-sensor brightness temperature can be determined from the measured spectral radiance. Figure 6.28 shows an example of this.

In principle it is straightforward to determine the SST from a TIR observation, since the emissivity of sea water is well known (0.993 at a wavelength of 10 μm – see Section 3.6.2). The important problem of determining the brightness temperature of the surface from the brightness temperature measured at the sensor is one of atmospheric correction, discussed in the next section. This can be a rather large effect, of the order of 5 to 10 K. There is, however, another potential problem associated with the use of TIR data to determine the SST. The absorption length for TIR radiation in water is very small, typically 0.02 mm or less

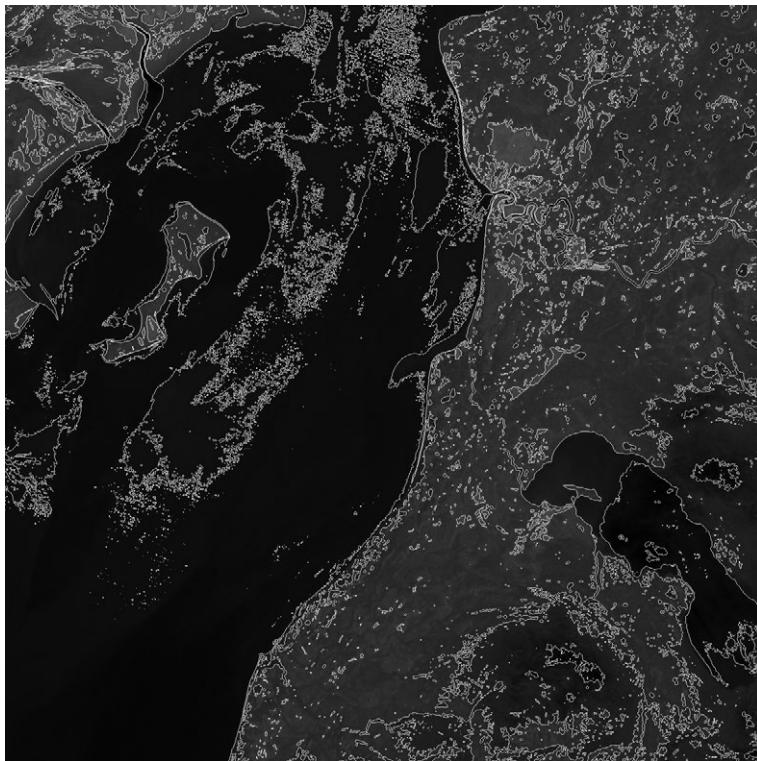


Figure 6.28. Extract of figure 6.25 with contours of at-satellite brightness temperature superimposed. The contours are at 5K intervals from 280 K to 300 K.

(Figure 3.1), which means that the measured brightness temperature is characteristic of only the upper tenth of a millimetre, at most, of the water surface. Oceanographically, the ‘surface’ is a few centimetres deep, and the physical temperature of the upper tenth of a millimetre can differ from the mean temperature of this layer by an amount of the order of 1 K (Donlon *et al.* 2002). This temperature difference, which can be either positive or negative, arises from a combination of evaporative cooling and solar heating.

Measurement of the LST is somewhat more complicated, because of the spatially and temporally variable emissivity. The emissivity can be estimated on the basis of the known land cover, or using the idea of thermal inertia, which is discussed in the following section. The effect of uncertainty in the emissivity of the land surface can also be reduced by constructing maps of surface temperature anomalies, i.e. differences between the at-surface brightness temperature and its long-term average value (Figure 6.29).

6.6.2

Thermal inertia

The Earth’s surface is subjected to a periodically varying input of heat from the Sun, as a result of the alternation of day and night, and of the seasons of the year. In consequence, the surface temperature fluctuates. The amplitude of the fluctuations in surface temperature depends on a combination of physical properties of the material of which the surface

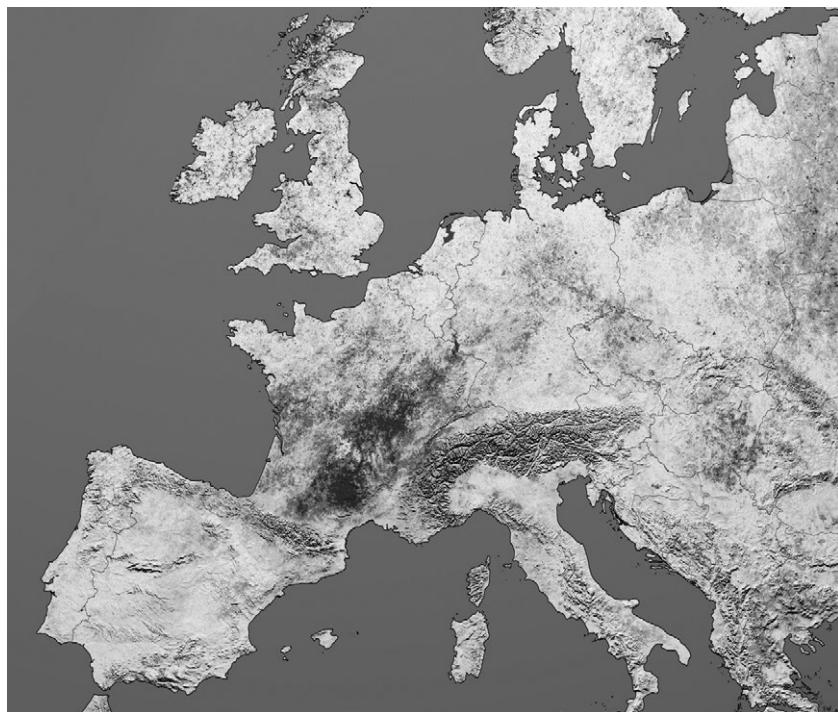


Figure 6.29. Visualisation of the temperature anomaly during the period 20 July 2003 to 20 August 2003, relative to the average for the corresponding periods in 2001, 2002 and 2004. The anomalies were calculated using MODIS TIR imagery. (Image by Reto Stöckli, Robert Simmon and David Herring, NASA Earth Observatory, based on data from the MODIS land team.) See also colour plates section.

is composed, called the *thermal inertia*, and hence measurement of this amplitude can provide some information about the composition of the Earth's surface (Kahle, Gillespie and Goetz 1976, Quattrochi and Luvall 2009). The amplitude also depends on the emissivity of the surface. This is the idea behind thermal inertia mapping, and in this section we will develop a simple model of it.

For our model, we suppose that the material of the Earth's surface is homogeneous and infinitely deep, extending from $z = 0$ at the surface to $z = +\infty$. Heat is propagated by conduction, only in the vertical (z) direction. The thermal behaviour of the material is governed by the heat conduction equation

$$\mathbf{F} = -K\nabla T$$

(in which \mathbf{F} is the vector heat flux, K is the thermal conductivity and T the temperature) and by the heat capacity

$$\nabla \cdot \mathbf{F} = -C\rho \frac{\partial T}{\partial t},$$

where C is the heat capacity per unit mass (often called the specific heat capacity), ρ is the density and t is time. Since in fact we are assuming that heat propagates only in the z -direction, these equations can be simplified to

6.6 Major applications of thermal infrared images

$$F = -K \frac{\partial T}{\partial z} \quad (6.10)$$

and

$$\frac{\partial F}{\partial z} = -C\rho \frac{\partial T}{\partial t}. \quad (6.11)$$

If we also assume that the time-dependence of the flux and the temperature is sinusoidal with angular frequency ω , and that the amplitude of the variation in flux at the surface is F_0 , it can be shown that the flux F is given by

$$F = F_0 \cos(\omega t - z\sqrt{\omega C\rho/2K}) \exp(-z\sqrt{\omega C\rho/2K}) \quad (6.12)$$

and the temperature T is given by

$$T = \frac{F_0}{\sqrt{P\omega}} \cos(\omega t - z\sqrt{\omega C\rho/2K} - \pi/4) \exp(-z\sqrt{\omega C\rho/2K}), \quad (6.13)$$

where

$$P = C\rho K \quad (6.14)$$

is the thermal inertia.

Examining Equations (6.12) and (6.13), we see that both the heat flux F and the temperature T vary as exponentially decaying sinusoidal waves. The speed at which the waves travel down into the ground is

$$v = \sqrt{\frac{2K\omega}{C\rho}} \quad (6.15)$$

and the attenuation depth, over which the amplitude of the waves decreases by a factor of e , is

$$z_0 = \sqrt{\frac{2K}{\omega C\rho}} = \sqrt{\frac{2\Gamma}{\omega}}, \quad (6.16)$$

where

$$\Gamma = \frac{K}{C\rho} \quad (6.17)$$

is the *thermal diffusivity*. We also observe that the ratio of the amplitude of the surface flux variations to the amplitude of the surface temperature variations is $(P\omega)^{1/2}$, and that the temperature variations lag the flux variations by $\pi/4$ radians or one-eighth of a cycle. Thus we would expect, on the basis of this simple model, that the daily surface temperature variations should reach a maximum value at about 3 p.m. and a minimum value at about 3 a.m.

In fact, the simple model that we have developed is not directly applicable, because the incident flux does not usually vary sinusoidally, although this particular limitation can be overcome by using Fourier analysis. The calculation of the flux is a complicated one, which includes contributions from direct solar radiation and from the sky, reflection of

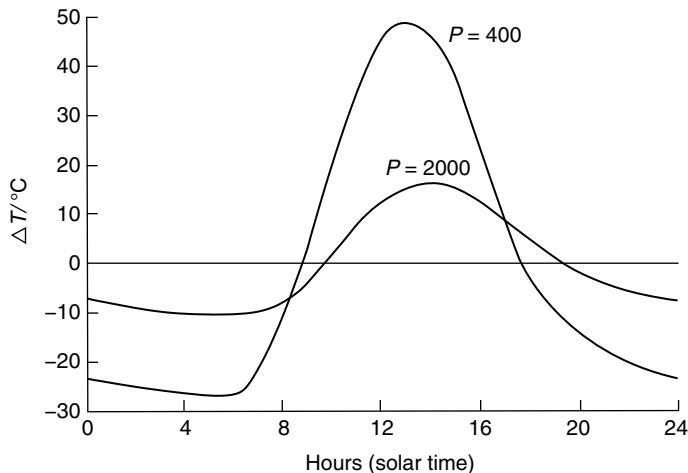


Figure 6.30. Typical diurnal variations of surface temperature. The graphs show ΔT , the surface temperature minus the diurnal mean temperature, plotted against local solar time, for materials with thermal inertias of 400 and $2000 \text{ J m}^{-2} \text{ s}^{-1/2} \text{ K}^{-1}$.

solar radiation from the surface, emission of infrared radiation from the surface, and geothermal heat flux. It thus depends on the geographical location, time of year, cloud cover, orientation, albedo and emissivity of the surface, amongst other factors (Elachi and van Zyl 2006) Nevertheless, our simple model gives a good indication of the general trends to be expected. Figure 6.30 shows typical variations of surface temperature for materials of two different thermal inertias. The material with larger thermal inertia exhibits smaller temperature fluctuations, as expected from our model. However, we can note from the figure that the fluctuations are not sinusoidal, and that the maximum temperatures occur at about 2 p.m. instead of 3 p.m., and the minimum temperatures occur just before dawn instead of at 3 a.m.

Figure 6.31 illustrates the typical values of the thermal inertia P and of the thermal diffusivity Γ for various materials. Geological materials have thermal inertias of a few thousand $\text{J m}^{-2} \text{ s}^{-1/2} \text{ K}^{-1}$. The thermal inertia of metals is typically ten times higher, on account of their higher thermal conductivities and densities, and wood is about ten times lower. There is little contrast between the thermal inertia of water and that of typical minerals. These may nevertheless be distinguished, since a water surface will normally be cooler than a rock surface during the daytime, as a consequence of evaporation, and at night the water surface is correspondingly warmer. A similar effect operates in the case of damp ground, and in this way soil moisture can be assessed, with an accuracy of typically 15%. Dry vegetation may also be distinguished from bare ground at night, since the vegetation and the ground beneath it is physically warmer than the bare ground because of the insulating effect of the vegetation. Again, a converse effect operates in the daytime.

We can extend our simple model of thermal inertia to include the effect of radiation emitted by the surface as follows. The temperature within the ground is assumed to vary with depth z and time t as

6.6 Major applications of thermal infrared images

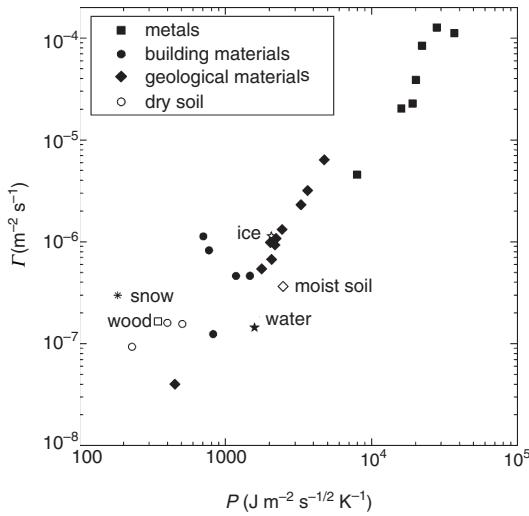


Figure 6.31. Typical values of the thermal diffusivity Γ and the thermal inertia P for various materials.

$$T = \bar{T} + T_0 \exp i(\omega t - kz),$$

where \bar{T} is the mean temperature and T_0 is the amplitude of the surface temperature fluctuations. This expression uses the complex exponential notation that was introduced in Chapter 2; the implication is that the real part of the complex expression represents the temperature. The net upward heat flux from the surface is linearised:

$$F_{up} = \alpha(T - \bar{T}),$$

where α is a constant. By making a Taylor expansion about \bar{T} of the expression for the total black-body radiation from a body of emissivity ϵ at temperature T , we can estimate α as

$$\alpha \approx 4\epsilon\sigma\bar{T}^3, \quad (6.18)$$

where σ is the Stefan–Boltzmann constant. The heat flux into the ground at the surface is

$$-K \left(\frac{\partial T}{\partial z} \right)_{z=0} = ikK(T_s - \bar{T}),$$

where T_s is the surface temperature. Thus, by conservation of energy at the surface, we see that the incident heat flux at the surface must be given by

$$\frac{F_{inc}}{T_s - \bar{T}} = \alpha + ikK$$

From Equation (6.12) we know that $k^2 = \omega c p (1 + i)2K$, so we finally obtain the relationship between variations in the incident flux and in the surface temperature as

$$\frac{F_{\text{inc}}}{T_S - \bar{T}} = \alpha + \left(\frac{\rho c K \omega}{2} \right)^{1/2} (1 + i). \quad (6.19)$$

This shows that the effect of increasing α (i.e. increasing importance of emitted radiation) is to decrease the phase difference between the flux and the surface temperature, as expected.

Thermal inertia mapping was carried out from space in 1978–1980, during the HCMM (Heat Capacity Mapping Mission). This satellite carried a TIR radiometer (10.5–12.5 μm) called the HRIR (High Resolution Infrared Radiometer), which had an accuracy of 0.4 K. The results from this mission indicated that thermal inertia mapping is particularly sensitive to the effects of tectonic disturbance and to lithological boundaries. However, distinguishing one rock type from another is still rather difficult. Since the advent of the NOAA POES satellites from the early 1980s it has been possible to carry out thermal inertia mapping using TIR imagery collected at different times of day (Sobrino and El Kharraz 1999), and a similar approach is used with MODIS imagery to estimate LST and surface emissivity (Figure 6.32).

Thermal inertia mapping also has an application to archaeological surveying. If one material is buried within another, and the materials have different thermal properties, the flow of heat will be distorted and give rise to a *temperature anomaly* at the surface. This will yield information on the nature of the buried object, but the interpretation is complicated by the fact that the simple one-dimensional problem that we have been describing is no longer applicable, and the thermal diffusion equation must be solved in three dimensions.

6.6.3

Cloud detection and monitoring

Contamination of a VNIR image by the presence of cloud is frequently a problem for the analysis of the Earth's surface, and it is highly desirable to be able to remove from the image those pixels that are affected by cloud. Conversely, one might wish to study the distribution of cloud. In either case, identification of cloudy areas is necessary, and this is often accomplished using a combination of VNIR and TIR imagery (although suitable VNIR imagery can also be used on its own). In some cases this can be performed manually, although automated procedures are clearly desirable.

The brightness temperature of a cloud, when viewed from above, is approximately equal to the *cloud-top temperature*, and hence to the atmospheric temperature at the height of the cloud-top. Thus, clouds tend to be colder than the underlying surface, increasingly so for higher altitude clouds. A cloud that is optically thick in the VNIR band will have a high reflectance, as discussed in Section 3.6.1. Thus, the simplest cloud-detection algorithms are designed to look for pixels that are bright (in the VNIR band) and cold (in the TIR band) (Platnick *et al.* 2003). During the daytime, a combination of TIR channels near 4 μm and 10 μm is particularly useful, since the shorter-wavelength channel detects not just thermally emitted radiation from the cloud but also reflected solar radiation, thus increasing the brightness temperature of the detected radiation. The difference between the brightness temperatures in the two channels is thus diagnostic of the presence of clouds.

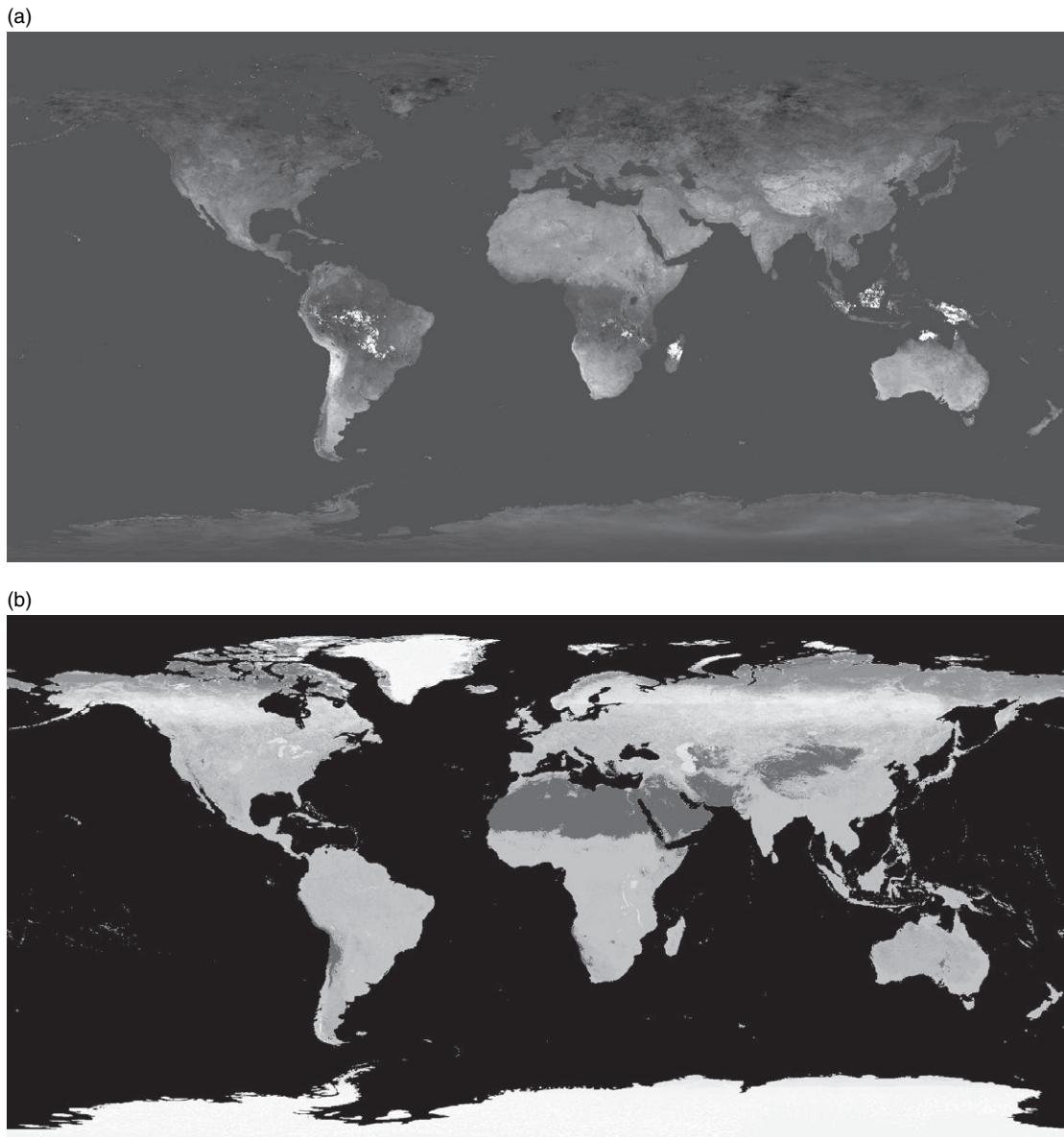


Figure 6.32. (a) average difference between daytime and nighttime LST for the month of January 2007; (b) calculated TIR emissivity of the land surface during the same month. Both figures have been derived from the MODIS MOD11C3 Land Surface Temperature data product. The greyscales cover the ranges from 0 to 40 K and from 0.9 to 1 respectively.

More detailed information on the type of cloud can be obtained by using the brightness–temperature information quantitatively, by examining the texture of the image radiance (see Chapter 11), or by measuring the content of liquid water (or solid ice) integrated vertically through the cloud. Figure 6.33 shows an example of cloud detection in VNIR and TIR imagery.

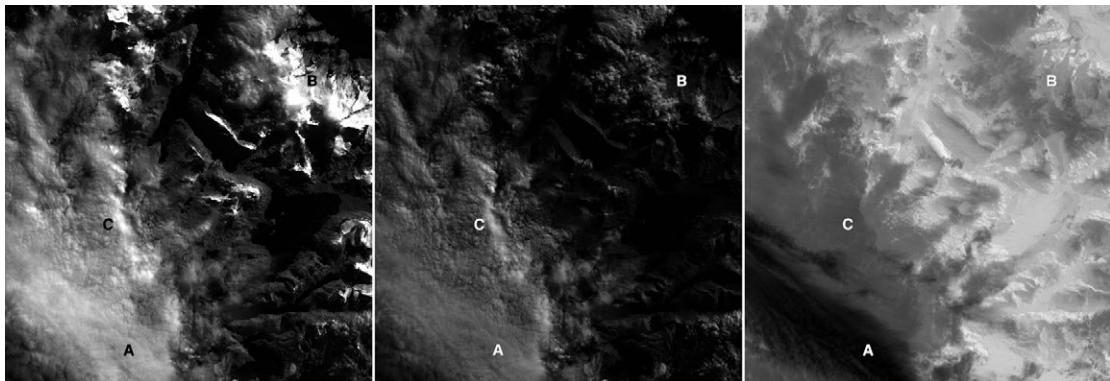


Figure 6.33. Extract of a Landsat image of Tierra del Fuego, Chile, illustrating some principles of cloud detection. The three subimages show bands 4, 5 and 6 respectively. At A, the image shows cold, high cloud with a high reflectance in both band 4 and band 5 and a low brightness temperature (around 248 K). At B, the image shows snow, with a high reflectance in band 4 but a low reflectance in band 5. The brightness temperature is around 270 K. At C, the image shows lower, warmer cloud (brightness temperature around 261 K.)

Summary

The major application of TIR imagery is to determine the Earth's surface temperature. Sea surface temperature (SST) is an important climate variable and is comparatively straightforward to determine from TIR imagery since the emissivity of the surface is uniform and well known. Apart from correction of atmospheric effects, the main difficulty is the fact that the technique measures the temperature of only the upper few tens of micrometres of the sea, and this can be significantly different from the temperature below this surface layer. Over land surfaces, the land surface temperature (LST) is more difficult to determine because of spatial and temporal variations in emissivity. One approach is to use the known land cover to estimate the emissivity, while another is simply to measure departures from the average brightness temperature. The concept of thermal inertia provides the basis for a third approach to measuring LST. Thermal inertia is a physical property of a material that governs the amplitude of diurnal temperature fluctuations at the surface in response to diurnal variation of the heat flux into the material, but the surface emissivity also plays a role. Thus, if measurements of the surface brightness temperature are made at different times of day, it can be possible to estimate both the thermal inertia of the surface material and also its emissivity.

Another important application of TIR imagery is to the detection of clouds, through their low brightness temperature. TIR and VNIR imagery are often combined for this application.

6.7

Atmospheric sounding

Up to this point, this chapter has discussed systems for imaging the Earth's surface (or, where the surface is covered by cloud, for imaging the cloud itself). Thus, the wavebands chosen for such imaging systems are within the atmospheric 'windows' discussed in

6.7 Atmospheric sounding

Chapter 4. However, by choosing to make observations at wavelengths at which the atmospheric attenuation coefficient is significant, information can be obtained about the composition of the atmosphere itself. The type of information that is required is the variation of temperature with altitude, and the variation of the density of atmospheric gases and aerosols with altitude. The most important of the gases are oxygen, because this is a good indicator of the total atmospheric pressure, and water vapour, because of its meteorological significance and also because of its importance for atmospheric correction of TIR measurements (see Section 6.4.5). However, many other gases are also routinely profiled by remote sensing methods, notably ozone and other radiatively active molecules.

Atmospheric sounding techniques exploit all three of the phenomena that were introduced in our discussion of the radiative transfer equation (Section 3.5), namely absorption, scattering and thermal emission. Most observations are made in the thermal infrared or microwave bands (the latter will be discussed in Chapter 7), although the optical and ultraviolet bands are used for some scattering measurements.

6.7.1

Temperature profiling from observations at nadir

In atmospheric temperature profiling observations, the significant terms in the radiative transfer equation are absorption and thermal emission – scattering can be neglected. The principle of nadir (vertical) atmospheric temperature profiling can be stated in words something like this: If a sensor views vertically downwards into the atmosphere at a wavelength at which the atmosphere is optically thick, the brightness temperature of the radiation that is received will be characteristic of the atmosphere at a depth below the sensor that is of the order of the absorption length. Thus the greater the absorption coefficient, the smaller the absorption length and hence the greater the altitude from which the temperature signal is received. Hence, by making observations at a number of wavelengths near a broad absorption line, different altitudes in the atmosphere can be investigated. Thermal infrared temperature profilers normally employ the broad and deep carbon dioxide absorption feature near $15\text{ }\mu\text{m}$.

The foregoing explanation is of course oversimplified. The concept of absorption length is not strictly applicable since the absorption coefficient varies with altitude, and furthermore the brightness temperature received at the sensor is not characteristic of a single altitude but rather of a range of altitudes. Nevertheless the essential principle is valid: the closer the observation is made to the centre wavelength of an absorption line, and hence the larger the optical thickness of the atmosphere, the greater will be the mean altitude from which the signal is obtained. We can deal with the problem that the observed temperature will be derived from a range of altitudes by introducing a weighting function $w(h')$ where h' is the altitude of a layer of the atmosphere and the observed brightness temperature at altitude h is given by

$$T_{\text{obs}} = \int_0^{\infty} T(h')w(h')dh'. \quad (6.23)$$

Our task is to determine this weighting function $w(h')$.

Since only absorption and thermal emission are significant, the appropriate form of the radiative transfer equation is given by Equation (3.91):

$$\frac{dL_v}{dz} = \gamma_a(B_v - L_v).$$

As discussed in Section 3.5, the solution of this equation can be written as

$$L_v(\tau) = L_v(0) \exp(-\tau) + \int_0^\tau B_v(\tau') \exp(\tau' - \tau) d\tau', \quad (3.95)$$

where τ and τ' are optical thicknesses measured from the Earth's surface to some point in the atmosphere. In fact, it is much simpler to recast Equation (6.23) in terms of optical thicknesses, so that

$$T_{\text{obs}} = \int_0^\tau T(\tau') w(\tau') d\tau', \quad (6.24)$$

where τ' is monotonically related to the altitude h' by

$$\tau' = \int_0^{h'} \gamma_a(h') dh'. \quad (6.25)$$

With this substitution, Equation (3.95) is almost in the form we need. We can simplify it greatly by making the same approximation that we used in Section 6.4.5, namely that the physical temperatures in the atmosphere do not differ greatly from the brightness temperature T_{b0} of the radiation leaving the Earth's surface. In this case, we can make a Taylor expansion of Equation (3.95) to obtain

$$T_{\text{obs}} \approx T_{b0} \exp(-\tau) + \int_0^\tau T(\tau') \exp(\tau' - \tau) d\tau', \quad (6.26)$$

where τ is the optical thickness of the entire atmosphere. We should note, first, that this equation has exactly the same form as Equation (3.96), and second, that the approximation that we have used here in deriving Equation (6.26) is not necessary – we have just used it to illustrate the way in which weighting functions are calculated. With this approximation, we see that the appropriate weighting function is just $\exp(\tau' - \tau)$, where the argument of the exponential is the negative of the optical depth below the sensor.

In order to estimate roughly the form of the altitude-based weighting function $w(h')$, we will make two more approximations. The first is that the absorption coefficient is proportional to the atmospheric pressure, and the second is that the pressure varies with height according to the negative exponential form introduced in Equation (4.5). (In fact, the absorption coefficient depends on the temperature as well, for the reasons discussed in Section 4.2, so the procedure for solving the temperature profile is not quite as simple as outlined here. An iterative procedure must be used.) With these assumptions, the relationship between τ' and h' becomes

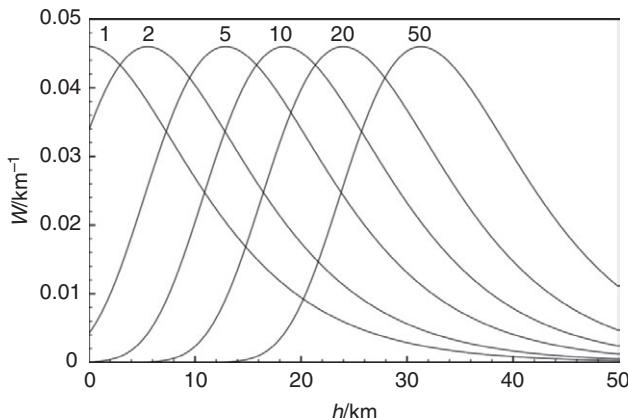


Figure 6.34. Weighting functions $w(h)$ calculated for profiling the atmospheric temperature distribution, using the simplified equation (6.28). The curves are labelled with values of τ , the total optical thickness of the atmosphere.

$$\tau' = \tau \left(1 - \exp(-h'/H) \right), \quad (6.27)$$

where H is the atmospheric scale height, and the altitude-based weighting function is then given by

$$w(h') = \frac{\tau}{H} \exp \left(-h'/H - \tau \exp(-h'/H) \right). \quad (6.28)$$

Figure 6.34 illustrates weighting functions $w(h')$ calculated from Equation (6.28) using a scale height H of 8 km.

The principle of atmospheric temperature sounding is thus to make nadir observations of the brightness temperature at a number of wavelengths corresponding to different values of the total optical thickness τ , giving a family of weighting functions similar to those shown in Figure 6.34. Inversion of the data allows the temperature profile to be retrieved. We can note from Figure 6.34, first, the obvious point that if τ is small, there will be a significant contribution from the Earth's surface as well as from the atmosphere, and second, that the vertical resolution of the technique is rather poor, being of the order of 10 km. The family of curves demonstrates the point that we have already noted, namely that the height in the atmosphere from which the bulk of the signal originates increases with the opacity of the atmosphere.

We have derived our simplified equation (6.28) on the assumption that the distribution of gas pressure with altitude is known. In fact, it is more usual to use the total atmospheric pressure as an altitude-like variable, rather than the altitude itself. The total pressure at altitude h is a measure of the number of molecules above the altitude h , as shown in Equation (4.4), so provided that the absorbing gas (for example carbon dioxide) is well mixed in the atmosphere, the total pressure is a measure of the optical thickness of the atmosphere above h .

Figure 6.34 showed a set of weighting functions calculated using a number of simplifying assumptions. In Figure 6.35, a set of real weighting functions is shown, as functions of the atmospheric pressure rather than the height. The HIRS/3 (High Resolution

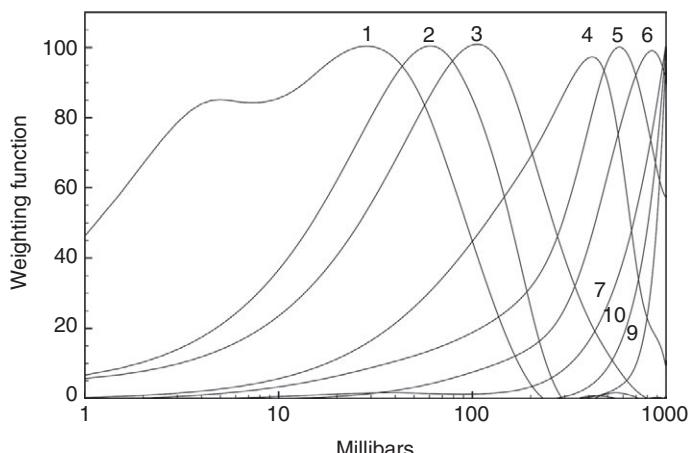


Figure 6.35. Weighting functions for the HIRS/3 instrument. The curves are labelled with the HIRS channel numbers.

Infrared Radiation Sounder) instrument is carried on board the NOAA POES satellites and records TIR radiance in 19 narrow channels (it also measures VNIR radiation). Nine of these make use of the carbon dioxide absorption feature near $15\text{ }\mu\text{m}$. The peak absorption of this feature occurs at a wavelength of around $15.1\text{ }\mu\text{m}$. HIRS/3 channels 1 to 7 are sensitive to radiation with wavelengths decreasing monotonically from $14.95\text{ }\mu\text{m}$ to $13.35\text{ }\mu\text{m}$, so these have weighting functions that peak at progressively lower heights (higher pressures) as seen in Figure 6.35. Channels 7 to 10 are yet further away from the peak absorption, and indeed channel 8 is located at a wavelength ($11.11\text{ }\mu\text{m}$) at which the atmosphere is almost transparent, so there is very little contribution to the detected signal from the atmosphere.

6.7.2 Profiling of gas concentrations at nadir

The essence of atmospheric temperature profiling is the assumption that the vertical distribution of the concentration of some absorbing gas, for example carbon dioxide, is known, and that therefore the vertical distribution of the absorption coefficient is also known. From this known distribution, the radiative transfer equation is solved to determine the temperature distribution. The essence of profiling gas concentrations is the converse of this procedure: if the temperature distribution is known, we can make a number of observations of the brightness temperature at wavelengths near the absorption line of a particular molecular species to determine the profile of the absorption coefficient, and hence concentration, of that species. This application is often referred to as the measurement of *atmospheric chemistry*. As for temperature profiling, it is necessary to use an iterative procedure to solve for the concentration profile, since the absorption coefficient is dependent on temperature. Since most molecular spectral lines are much narrower than the $15\text{ }\mu\text{m}$ carbon dioxide feature commonly used for temperature profiling, high spectral resolutions are generally required for this technique. The methods by which they are achieved are discussed in Section 6.7.5.

6.7.3

Backscatter observations at nadir

Nadir-looking observations can also be used to measure solar radiation that is scattered out of the atmosphere. From our considerations in Section 3.4.1 (Equation 3.85), we expect the scattering coefficient to increase rapidly towards the short-wavelength (blue and ultraviolet) end of the spectrum, so it is in this region that backscatter measurements are made. Thermal emission can be neglected in this part of the electromagnetic spectrum, so it is only necessary to consider the scattering and absorption terms in the radiative transfer equation. Backscatter observations are used mainly for profiling of atmospheric ozone distributions.

6.7.4

Limb-sounding observations

We noted in Section 6.7.1 that a nadir-sounding observation has a comparatively poor vertical resolution. Significantly higher resolutions can be achieved by limb-sounding, in which the sensor views in a direction that is almost tangential to the Earth's surface. Limb-sounding observations can be used to measure both absorption and emission. To illustrate the principle, we will derive a simple model of an absorption measurement.

Figure 6.36 shows the geometry of a limb-sounding absorption measurement. The sensor views a source of radiation (for example, the Sun) in such a direction that the closest distance between the line of sight and the Earth's surface is h_0 . We measure distance x along the line of sight from this point, and write $h(x)$ for the altitude of the line of sight above the Earth's surface at x . Assuming that the sensor is located well outside the atmosphere, the optical thickness traversed by the ray is

$$\tau = \int_{-\infty}^{\infty} \gamma_a(h(x)) dx,$$

where $\gamma_a(h(x))$ is the absorption coefficient at altitude $h(x)$. Provided that $h(x)$ remains small compared with the Earth's radius R , we can put

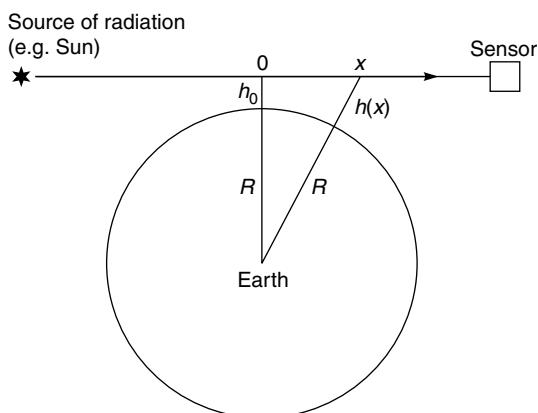


Figure 6.36. Geometry of a limb-sounding observation (schematic).

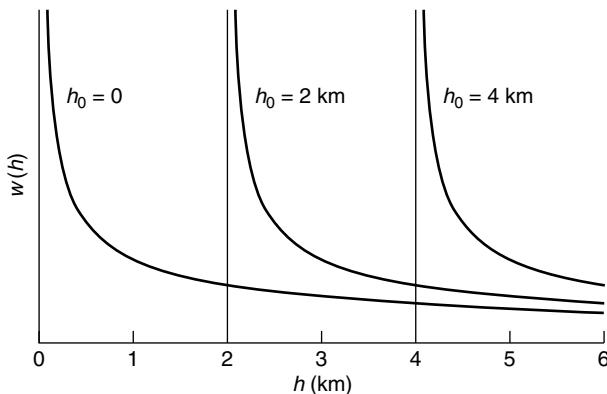


Figure 6.37. Simplified weighting functions for limb-sounding observations. The curves are labelled with values of h_0 , the minimum altitude of the ray.

$$h(x) \approx h_0 + \frac{x^2}{2R} \quad (6.29)$$

and hence rewrite the equation for the optical thickness in terms of h :

$$\tau \approx \sqrt{\frac{R}{2}} \int_{h_0}^{\infty} \frac{\gamma_a(h)}{\sqrt{h-h_0}} dh. \quad (6.30)$$

Thus, the appropriate weighting function in this case (the contribution to the total optical thickness from the altitude h) is proportional to $(h-h_0)^{-1/2}$ for $h > h_0$, and zero for $h < h_0$. Figure 6.37 illustrates this function.

In practice, the weighting function for a limb-sounding observation will not be quite as sharply peaked as implied by Figure 6.37. This is because the finite angular resolution of the sensor will blur and broaden the response somewhat. As an example, we can consider a sensor at an altitude of 800 km arranged to observe the altitude $h_0 = 0$. In this case, the distance from the sensor to the tangent point is approximately 3300 km, so an instrument having an angular resolution of 0.1 milliradian would introduce a vertical blurring of 0.33 km in the vertical response. (For comparison, an angular resolution of 0.1 milliradian is equivalent to a spatial resolution of 80 m in a nadir observation from an altitude of 800 km.)

The improved vertical resolution of a limb-sounding observation, when compared with a nadir observation, is bought at the price of degraded horizontal resolution. This can be seen by considering an observation with a vertical resolution of Δh , so that the weighting function is significantly greater than zero for h between h_0 and $h_0 + \Delta h$. Equation (6.29) thus shows that the range of values of x corresponding to this range of altitudes is from $-(2R\Delta h)^{1/2}$ to $+(2R\Delta h)^{1/2}$. Taking a typical value of 1 km for Δh and setting $R \approx 6400$ km, we see that the sample volume is approximately 230 km long in the direction of the line of sight.

In order to illustrate the principle of limb-sounding observations, we have considered the case of an absorption measurement. As shown in Figure 6.36, this requires that there is

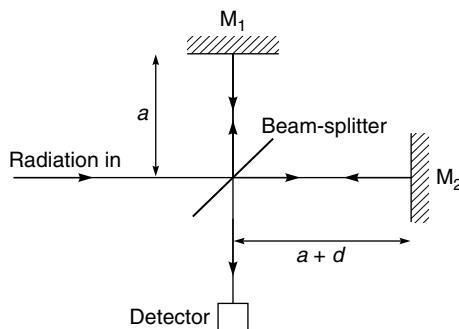


Figure 6.38. Principle of a Michelson interferometer.

a source of radiation external to the atmosphere. Most atmospheric limb-sounders commonly use the Sun as this source of radiation, in which case the observation is often referred to as a *solar occultation* observation. Since a satellite in low Earth orbit makes approximately 14 orbits of the Earth per day (see Chapter 10), there are approximately 28 measurement opportunities per day (14 sunrises and 14 sunsets). The same is true for lunar occultation observations. However, stellar occultations (observations of stars near the Earth's limb) can be made from virtually anywhere on the Earth's night side. Thermal emission measurements can be made from any location.

6.7.5

Spectral resolution for atmospheric sounding observations

We noted in Section 6.7.2 that high spectral resolutions are generally needed for atmospheric sounding measurements (we are attempting to resolve the fine spectral structure that could not adequately be shown in Figure 4.4). At TIR wavelengths, the use of filters can achieve a spectral resolution of the order of $0.1 \mu\text{m}$, and diffraction grating spectrometers can improve this by a factor of 10. However, resolutions of 0.1 nm or better may be required. One common way of achieving this degree of spectral resolution is through *Fourier-transform spectrometry*, in which a Michelson interferometer is used to resolve the fine spectral components present in pre-filtered radiation. Figure 6.38 shows the principle of a Michelson interferometer.

A parallel beam of radiation is incident on a beam-splitter at 45° to the beam. This sends half the amplitude of the radiation towards mirror M_1 and half towards mirror M_2 . The distance from the beam-splitter to M_1 is fixed and equal to a , but the distance to M_2 is equal to $a + d$, where d can be varied. After reflection from the mirrors, the rays are recombined and detected. It is clear that the radiation that has been reflected from M_2 has travelled a distance that is greater by $2d$ than the radiation that has been reflected from M_1 . If the incident radiation consists of a single spectral component with intensity I_0 and wavenumber k , the intensity at the detector will be given by

$$I = \frac{I_0}{2} (1 + \cos(2kd)).$$

Thus, as d is varied, the output of the device will exhibit interference fringes as shown in Figure 6.39.

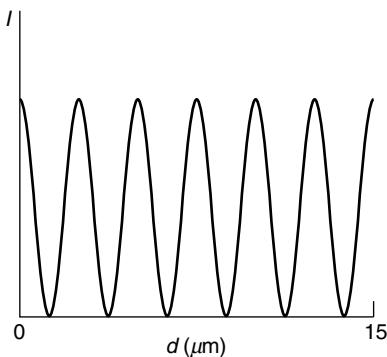


Figure 6.39. Output intensity from a Michelson interferometer when the input radiation consists of a single spectral component (in this example, the wavelength is $5 \mu\text{m}$).

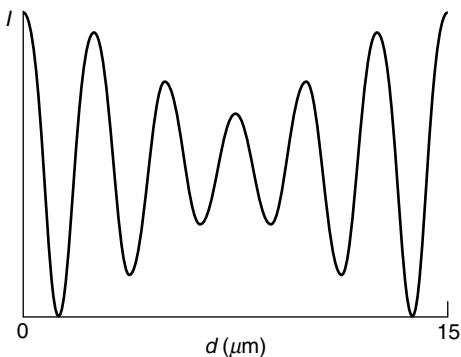


Figure 6.40. Output intensity from a Michelson interferometer when the input consists of two spectral components with different intensities.

If we now add a second component to the incident radiation, with a different intensity and wavenumber, a second fringe system will be added to the output and the result will look something like Figure 6.40. We can see that the fringes have now been modulated so that, in general, the minimum intensity of a fringe is no longer zero.

We define the *visibility* V of a fringe by

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \quad (6.31)$$

where I_{\max} and I_{\min} are respectively the maximum and minimum detected intensities. For the example shown in Figure 6.40, the graph of V against d is as shown in Figure 6.41.

The visibility function $V(d)$ contains all the information about the spectral structure of the input radiation. If the intensity contained in an infinitesimal interval dk of wavenumber is $I(k) dk$, and we set $k' = k - k_0$, where k_0 is the mean wavenumber of the input radiation, it can be shown (e.g. consult any textbook on physical optics) that

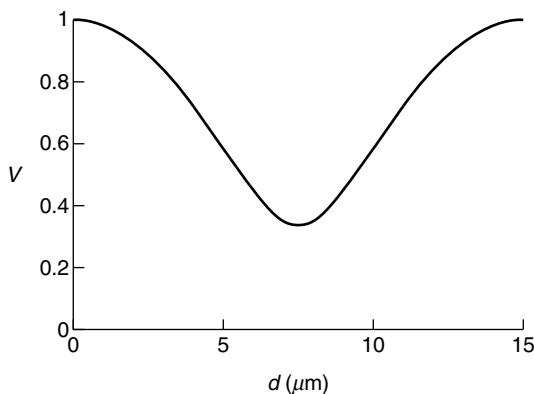


Figure 6.41. Visibility function $V(d)$ for the fringes shown in figure 6.40.

$$V(d) = \frac{\int_{-\infty}^{\infty} I(k') \exp(-2ik'd) dk'}{\int_{-\infty}^{\infty} I(k') dk'}. \quad (6.32)$$

This is a Fourier transform, so the function $I(k')$ can be retrieved from $V(d)$ by performing the inverse Fourier transform (see Section 2.3). (In fact, Equation (6.32) implies that V can be complex. The definition of Equation (6.31) therefore corresponds to the amplitude of the complex visibility. The phase is determined from the phase shift of the fringes.)

Thus, the procedure for determining the spectral structure of the input radiation is to scan the mirror M_2 in d and monitor the variations in output intensity. From these variations the visibility function can be obtained and then inverted to deduce $I(k)$. The maximum spectral resolution that can be obtained by this method depends on the largest value of d that can be achieved. If this value is D , the spectral resolution in terms of wavenumber is of the order of $2\pi/D$. For example, if $D = 0.1$ m, the wavenumber resolution Δk will be about 60 m^{-1} . This corresponds to a wavelength resolution of about 2 nm at a wavelength of 15 μm , and 0.01 nm at 1 μm .

Before leaving the topic of spectral resolution, we briefly describe an alternative technique. This depends on having an absorbing filter whose spectral absorption characteristics exactly match the spectral structure of the feature to be detected. It is clear that, by comparing the output obtained when this filter is present to the output when it is absent, a difference signal will be obtained that corresponds precisely to the intensity of the desired spectral feature. How can such a filter be constructed? The answer is beautifully simple. If we wish to detect radiation corresponding to a particular gas at a particular pressure, the filter consists of a transparent-walled cell containing the same gas at the desired pressure. By varying the pressure in the cell, different gas pressures can be detected. An instrument employing this technique is usually called a *pressure-modulated radiometer*, and it can achieve, in effect, remarkably high spectral resolutions (Δk down to 10^{-4} m^{-1}). Elachi and van Zyl (2006) provide more information on this technique.

6.8 Some profiling instruments

To conclude this chapter, we describe a few spaceborne atmospheric profiling instruments that are representative of the main observing techniques.

As an example of a nadir profiler, we consider the AIRS (Atmospheric Infrared Sounder) instrument that is carried on board the Aqua satellite, launched in 2002 (Maddy and Barnet 2008). The principal waveband of this instrument is in the TIR (it also has six broad VNIR channels), with coverage from 3.74 to 15.4 μm . It uses a grating spectrometer to achieve a spectral resolution that varies from about 3 nm at the short-wavelength end to about 13 nm at the long-wavelength end, giving a total of 2300 spectral channels. In fact the AIRS instrument can be scanned up to 49° either side of nadir, which improves its ability to resolve vertical temperature structure in a manner similar to the discussion of two-look TIR observations presented in Section 6.4.5. AIRS data are used to retrieve atmospheric temperature profiles and also to profile a number of chemical species including carbon monoxide, carbon dioxide, water vapour, ozone, methane and sulphur dioxide. Figures 6.42 and 6.43 show examples of the use of AIRS data to calculate atmospheric temperature profiles.

As an example of a nadir sounder for atmospheric chemistry we consider the MOPITT (Measurements Of Pollution In The Troposphere) instrument. This is carried on board the Terra satellite and has been operational since 2000. It is designed principally to measure profiles of carbon monoxide concentration (Zhang 2011), although it can also measure methane. Radiation is detected by an 8-channel infrared radiometer, defined using pressure-modulation radiometry and a closely related technique called *length-modulated radiometry*. These define spectral responses of $4.62 \pm 0.11 \mu\text{m}$ and $2.33 \pm 0.02 \mu\text{m}$, used for detecting carbon monoxide, and $2.26 \pm 0.07 \mu\text{m}$, used for detecting methane. For each

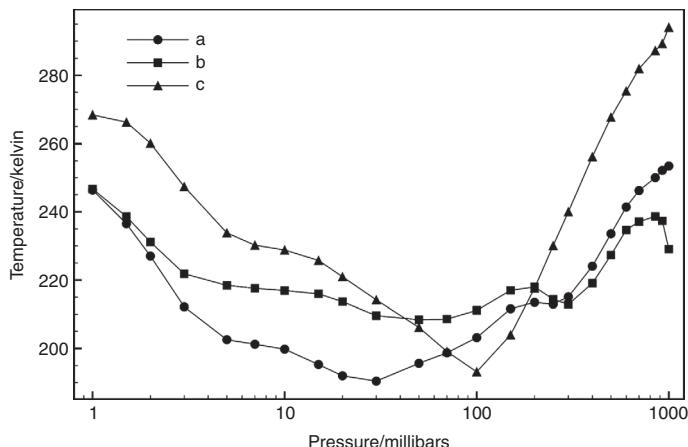


Figure 6.42. Temperature profiles determined from AIRS data collected on 20 December 2010. Each profile is the average for a 1° latitude \times 1° longitude region of the Earth, with north-western corners at (a) 72° N, 16° W (in the Greenland Sea), (b) 68° N, 120° E (in Siberia) and (c) 8° S, 90° W (in the southern Pacific Ocean, off Peru).

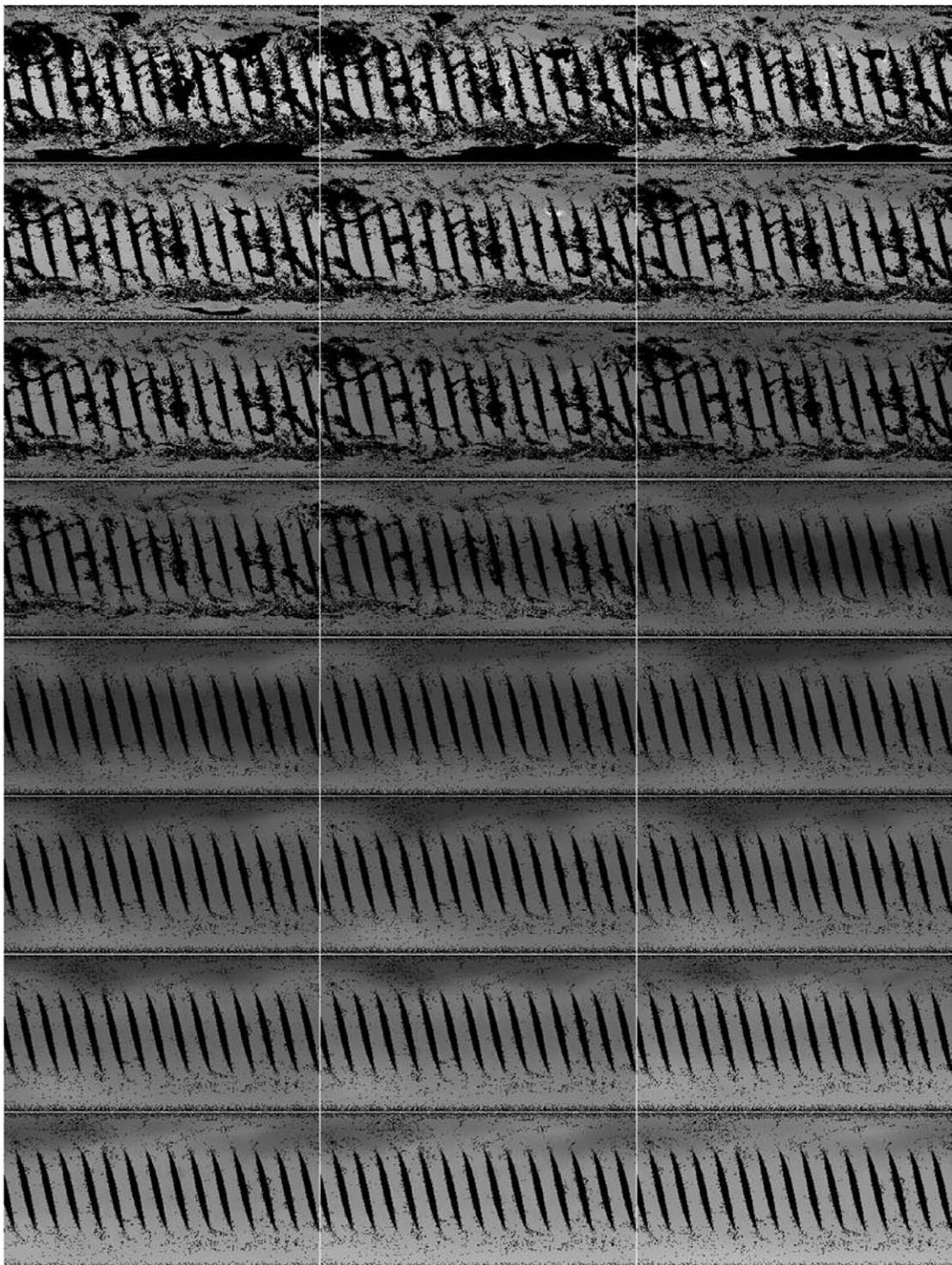


Figure 6.43. Global AIRS temperature data for 20 December 2010, gridded into 1-degree cells. 24 atmospheric pressure levels are shown from left to right, top to bottom, at pressures of 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 15, 10, 7, 5, 3, 2, 1.5 and 1 millibar respectively. The black areas represent cells from which no data were collected, or where the temperature retrieval algorithm failed.

wavelength, at least two different cell pressures are used, corresponding to different heights in the atmosphere. For example, the 2.26 μm wavelength is sampled by two length-modulated radiometers. One of these contains carbon monoxide at a pressure of 200 millibars, the other at 800 millibars; the lengths of both gas cells can be varied between 2 and 10 mm. This same wavelength is also sampled by two pressure-modulated radiometers. Both of these have cells 10 mm in length, but in one the pressure can be varied from 50 to 100 millibars while in the other it can be varied between 25 and 50 millibars. The horizontal resolution of MOPITT is about 22 km. The vertical resolution is nominally 4 km, similar to that achieved by nadir sounding of temperature using TIR radiation. Figure 6.44 shows an example of MOPITT data.

The SBUV/2 (Solar Backscatter Ultraviolet Radiometer) carried on the NOAA POES satellites is an example of a nadir-viewing instrument for profiling ozone concentrations by means of backscattered ultraviolet radiation (Flynn *et al.* 2009). It detects ultraviolet radiation in the wavelength range 160 to 405 nm, using a photomultiplier tube onto which the radiation is directed by a rotating diffraction grating. The spectral resolution is around 1 nm. In the usual mode of operation, radiances are recorded at 12 fixed wavelengths between 252 and 340 nm (Table 6.4). Solar radiation is measured by placing a suitably oriented reflector (the ‘diffuser plate’) in the field of view of the instrument, so that solar radiation is scattered into it, and internal calibration is provided by an on-board mercury vapour lamp. Figure 6.45 visualises ozone concentration data from SBUV/2. It shows a

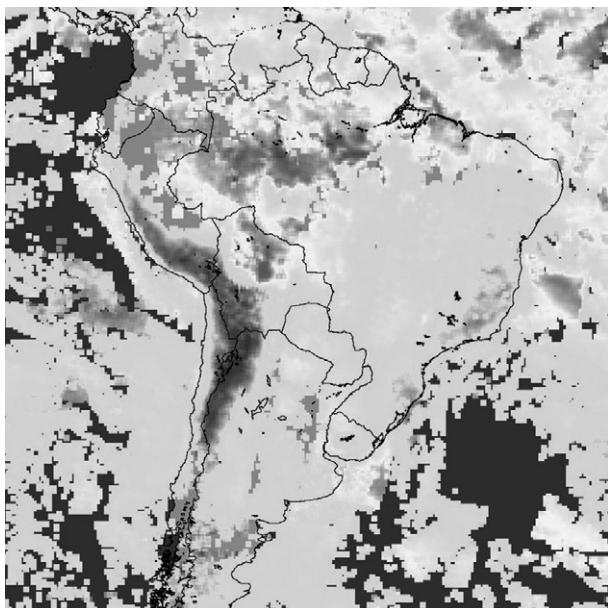


Figure 6.44. Column-integrated total concentration of carbon monoxide over South America, derived from MOPITT data. The image is a one-week gridded composite for the period 19–26 July 2004, and shows higher levels of carbon monoxide (red and yellow colours) in the north of Brazil as a result of fires. (Image downloaded from NASA Visible Earth at http://visibleearth.nasa.gov/view_rec.php?id=19432 and reproduced by courtesy of NASA.) See also colour plates section.

Table 6.4. Wavelengths of SBUV/2 instrument

Band number	Central wavelength (nm)	Band number	Central wavelength (nm)
1	252.0	7	302
2	273.6	8	305.7
3	283.1	9	312.6
4	287.7	10	217.6
5	292.3	11	331.3
6	297.6	12	339.9

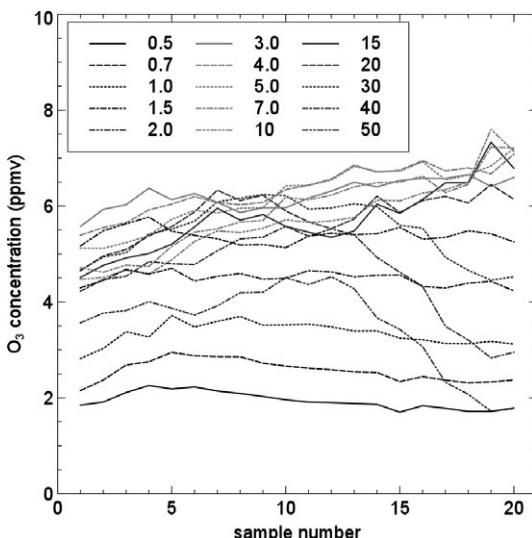


Fig. 6.45. Retrieved ozone concentrations from SBUV/2 data along a transect from 61.27° N, 83.62° W (sample 1) to 26.53° N, 98.04° W (sample 20), representing about 10 minutes of data. The key shows the atmospheric pressure levels in millibars. See also colour plates section.

short transect of data – 20 consecutive samples, collected between 16:40:19 and 16:50:27 Universal Time on 29 December 2007. The subsatellite track at this time extended from Hudson Bay to Texas: sample 1 was collected at 61.27° N, 83.62° W and sample 20 from 26.53° N, 98.04° W. Amongst other features, the figure shows a sharp decline in the lower altitude (pressures above about 20 millibars) ozone concentration from sample 12 (approximately located at Des Moines, Idaho) onwards.

Finally, we consider an example of a limb-sounding instrument. This is GOMOS (Global Ozone Monitoring by Occultation of Stars), which has been carried on the Envisat satellite, operational since 2002. Despite its name, GOMOS is used for profiling other things as well as ozone, notably atmospheric temperature, aerosols, oxygen, oxides of nitrogen, and water vapour (Renard *et al.* 2008). As its name implies, GOMOS observes radiation from stars as it passes through the Earth's limb (it is the first satellite sensor to use limb-sounding of starlight). It is sensitive to radiation with wavelengths between 248 and 956 nm, i.e. ultraviolet and VNIR radiation. Two reflection diffraction gratings are

used to disperse the radiation onto three CCD detectors, giving spectral ranges of 250–675 nm (with a resolution of 1.2 nm), 756–773 nm (0.2 nm resolution) and 926–952 nm (0.2 nm resolution). Photometers also collect radiation over the ranges 470–520 nm and 650–700 nm. The instrument is capable of observing the light from around 40–50 stars per orbit. Since starlight is not particularly intense, this requires high sensitivity from the detectors and also a large telescope (the aperture is 0.2×0.3 m) to collect radiation. The vertical resolution is 1.7 km.

Summary

Regions of the electromagnetic spectrum in which the atmosphere is less transparent than is useful for VNIR or TIR imaging can be exploited for profiling the atmosphere. The aim of temperature profiling is to estimate the variation of the atmospheric temperature with height (or, more commonly, with atmospheric pressure). The aim of chemical profiling is to estimate the variation with height (or pressure) of the concentration of some molecular species, for example water vapour, although the column integral of the concentration, i.e. the total amount of the molecular species in the whole atmosphere, is also useful and is easier to measure.

The signal (e.g. radiance or brightness temperature) measured at a single wavelength is a weighted average of the quantity (e.g. the concentration of some molecular species) over height. A fundamental concept in atmospheric profiling is the *weighting function*, which is a continuous function of height (or pressure) that expresses the relative contribution of a layer of the atmosphere at that height to the signal measured above the atmosphere. By repeating the measurement at a number of different wavelengths, with different weighting functions, the profile can be estimated. There are two main geometries for atmospheric profiling: nadir sounding, in which the instrument looks vertically downwards through the atmosphere, and limb-sounding, in which the line of sight is almost tangent to, though passes just above, the Earth's surface. Nadir sounding gives higher horizontal resolution at the expense of relatively coarse vertical resolution, while limb-sounding has the converse characteristics.

Atmospheric sounding instruments generally require high spectral resolution. This is achieved using diffraction gratings or interferometers (in which case the approach is often referred to as 'Fourier transform spectrometry'). Pressure- and length-modulated radiometers, in which the detected radiation is compared with radiation passed through a cell containing the gas species to be investigated, are also used.

Review questions

What are the main types of detector used for VNIR and TIR radiation?

Why is TIR radiation more difficult to detect than VNIR radiation?

Describe the different ways in which a VNIR imager can build up a two-dimensional image of the Earth's surface.

Problems

What determines the spatial and spectral resolutions of a VNIR imager?

How can a VNIR image obtained from a satellite-borne sensor be corrected for the effects of atmospheric propagation?

Describe the main types of VNIR imager, and suggest applications for each type.

Discuss the main applications of TIR imagery of the Earth's land and sea surfaces. What factors determine the wavelength(s) at which such imagery would be acquired?

Discuss the steps necessary to convert a TIR brightness temperature, measured at the top of the atmosphere, to the physical temperature of the Earth's surface.

Explain what is meant by thermal inertia.

Compare nadir sounding and limb-sounding as means to obtain profiles of atmospheric temperature or chemistry.

Explain what is meant by a 'weighting function' in profiling atmospheric temperature or chemistry from satellite data.

Why is high spectral resolution needed for obtaining atmospheric chemistry profiles? How can this be achieved?

Describe the main types of atmospheric profiling instrument using VNIR or TIR radiation, and give an example of each type.

Problems

1. The cross-section of a glass prism is an equilateral triangle. The refractive index of the glass is 1.601 at a free-space wavelength of 0.4 μm and 1.569 at 0.7 μm . Show that if white light is incident on the prism at an angle of 40° to the surface normal, the spectrum from 0.4 to 0.7 μm will be dispersed over an angular width of 4.9° .
2. Sunlight is incident on a rough surface at an angle of 45° to the normal. Calculate the brightness temperature of the surface at wavelengths of 4 μm and 10 μm , assuming that the surface is a perfect Lambertian scatterer (i.e. one for which the diffuse albedo is 1) at both wavelengths. Ignore atmospheric propagation effects.
3. Derive Equations (6.12) and (6.13).
4. A two-look TIR radiometer views the same point on the Earth's surface at nadir and at 60 degrees from nadir. The brightness temperature of the detected radiation is T_1 and T_2 respectively. By assuming that the surface is a black body of temperature T_b , the atmosphere has a uniform temperature T_1 , and that the Rayleigh–Jeans approximation is valid, show that

$$T_b = \frac{(T_1 - T_0)^2}{T_2 - T_0} + T_0$$

and hence calculate T_b if $T_1 = 274.0$ K, $T_2 = 269.0$ K and $T_0 = 265.0$ K. Ignore any effects arising from the Earth's curvature.

- (i) Calculate the sensitivity of your answer to the assumed value of T_0 and to the measured value of T_1 .

- (ii) Comment on the reasonableness or otherwise of the simplifying assumptions that have been made in deriving your answer.
5. Assume that the atmospheric absorption coefficient γ varies with height h as
- $$\gamma = \gamma_0 \exp(-\beta h),$$
- and use the simple model of limb-sounding geometry derived in Section 6.7.4 to show that the reduction in intensity I is given by
- $$\Delta \ln I = -\gamma_0 \exp(-\beta h_0) \sqrt{\frac{\pi R}{2\beta}}$$
- where h_0 is the height at which the radiation path passes closest to the Earth's surface. For atmospheric aerosols under clear-sky conditions, $1/\beta \approx 3$ km and $1/\gamma_0 \approx 14$ km for visible radiation. Estimate the value of h_0 that will reduce the intensity of a transmitted ray by a factor of 10.
6. A surface shows daily temperature fluctuations with amplitude 23 K, lagging 2.3 hours behind the fluctuations in incoming solar radiation. Use the model of thermal inertia developed in this chapter to estimate
- (i) the value of $\rho c K$,
 - (ii) the amplitude of the variations in the incoming solar flux,
if the value of α is taken as $5.5 \text{ W m}^{-2} \text{ K}^{-1}$. By differentiating Stefan's law with respect to temperature, show that this value of α corresponds to an emissivity of approximately 1 if $\bar{T} = 290$ K.
7. The output of the band 6 sensor of the Landsat-7 ETM+ instrument, as represented by an integer pixel value in an image, is the nearest integer to $16.03(L + 3.2) + 1$, where L is the at-satellite spectral radiance measured in units ($\text{W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$) as the graph. Band 6 is a TIR band, and the relationship between measured spectral radiance and brightness temperature can be assumed to be given by Equation (6.9) with the parameters given in the text. If the band 6 output, when observing an area known to consist entirely of melting snow, is 145, estimate the contribution of the atmosphere to the measured brightness temperature and comment on your answer.

7

Passive microwave systems

In Chapters 5 and 6 we considered passive remote sensing systems in which the diffraction resolution limit λ/D , while important, was not usually a critical parameter of the operation. In this chapter we consider our last major class of passive remote sensing system, the passive microwave radiometer. This is a device that measures thermally generated radiation in the microwave (usually 5–100 GHz) region. (Frequencies much below 1 GHz are unsuitable because of the large signal contributed by the Galaxy, as well as the difficulty of achieving adequate spatial resolution.) As we discussed in Section 2.6, the long ‘tail’ to the Planck distribution at relatively low frequencies means that measurable amounts of radiation are emitted even in this range of frequencies.

Because microwave wavelengths are so much greater than those of visible or even of thermal infrared radiation, the resolution limit plays a much more important role, and we shall need to give more attention to the factors that determine it. More detailed technical treatments of antenna theory are given by Ulaby, Moore and Fung (1981) and by Sharkov (2003), amongst others. Much of the technology and nomenclature of passive microwave radiometry was originally developed in the field of radio astronomy, and further details can also be found in works on that subject.

7.1

Antenna theory

7.1.1

Angular response and spatial resolution

As we have remarked before, electromagnetic radiation is detected through its influence on electrons, which are excited to higher energy states by the incident photons. The energy of a microwave photon is typically only a few μeV , which is too small to excite an electron across an atomic or molecular band-gap. For this reason, electrical conductors (metals) are used. The incident electromagnetic wave induces a fluctuating current in the conductor, and this current can subsequently be amplified and detected. The *antenna* is a structure which serves as a transition between the wave propagating in free space, and the fluctuating voltages in the circuit to which it is connected.

The usual form of a microwave antenna is a paraboloidal dish, although many other designs are possible. Figure 7.1 shows schematically the design of a simple microwave radiometer using such an antenna.

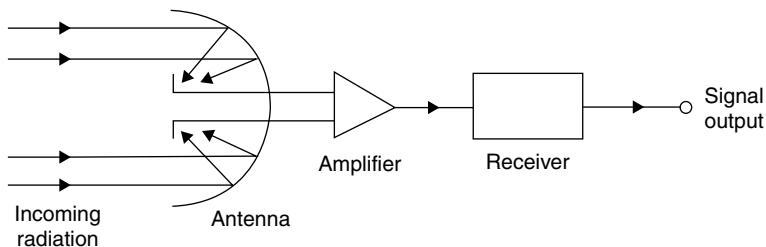


Figure 7.1 Schematic construction of a passive microwave radiometer. The antenna collects the incident radiation and generates a fluctuating voltage, which can then be amplified, detected and processed.

The design of an antenna is dominated by two considerations: (1) the need to achieve a high sensitivity in the desired direction; (2) the need to achieve a high angular resolution (narrow beamwidth). Both of these requirements are met by making the antenna as large as possible.

Let us consider a perfect antenna, in which there are no power losses. It has a characteristic *radiation resistance* R_r , a fictitious quantity that can most easily be understood in the case of a transmitting antenna. In this case, if an alternating current I is fed into the antenna at the appropriate frequency, the antenna will radiate a mean power $\langle I^2 \rangle R_r$, where $\langle I^2 \rangle$ is the mean square current. As far as the part of the circuit that feeds the antenna is concerned, the antenna ‘looks’ like a resistance R_r , in the sense that power is dissipated in it.

If the antenna (now in receiving mode again) is directed at a large, distant region that is emitting microwave energy, a voltage signal will appear at the output of the antenna. If the emission mechanism is thermal, this signal will be noise-like and will have the same characteristics as the *thermal noise* generated in a resistance R , held at a temperature T_A . Thermal noise is in essence caused by the Brownian motion of electrons in the resistance, and is often referred to as Johnson noise or Nyquist noise. It can be shown that the power contained in a frequency interval Δf of Johnson noise is given by

$$P_N = kT_A \Delta f \quad (7.1)$$

where k is the Boltzmann constant. (Note that Equation (7.1) implies that the power per unit frequency interval is constant.) The equivalent temperature T_A of the radiation resistance is then called the *antenna temperature*. If the distant emitting region is large enough, and has the radiation characteristics of a black body at a physical (absolute) temperature T , the antenna temperature is equal to T .

From the preceding discussion, we can provisionally describe the operation of a passive microwave radiometer something like this: the antenna is directed at some ‘target’, as a consequence of which a noise-like voltage signal is generated in it. From Equation (7.1), we can calculate the antenna temperature T_A , which will be some kind of weighted average of the brightness temperature viewed by the antenna. Clearly, our next task is to determine the nature of this weighted average.

The ideal antenna would receive radiation uniformly over a very small range of solid angle. To describe a real antenna, we introduce the concept of the *power pattern* $P(\theta, \phi)$.

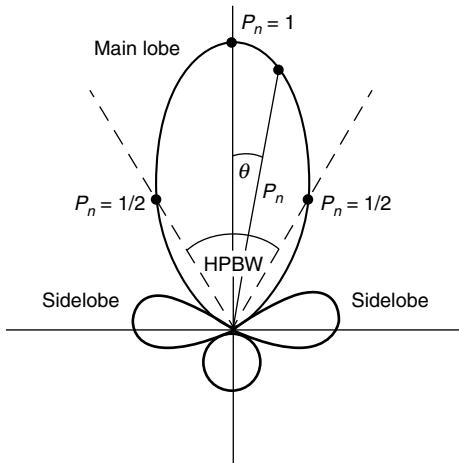


Figure 7.2. Power pattern of a typical antenna, plotted in polar coordinates.

This is the power that would be detected by the antenna from a point transmitter of fixed strength, located in the direction (θ, ϕ) with respect to the antenna axis, at a fixed distance much greater than the Fresnel distance (see Section 2.7). The power pattern is normalised so that neither the strength of the transmitter nor its distance from the antenna appear in its definition:

$$P_n(\theta, \phi) = \frac{P(\theta, \phi)}{P_{\max}(\theta, \phi)}. \quad (7.2)$$

Figure 7.2 illustrates a typical power pattern (or, at least, its θ variation). It has a *main lobe* of maximum sensitivity, usually centred on the direction $(0, 0)$, and a number of undesirable *sidelobes*. The usual measure of the width of the main lobe is its half-power beam width (HPBW), defined as the angular width of the region in which $P_n \geq 1/2$. For antennas with large apertures (in comparison with the wavelength λ of the detected radiation), the power pattern is calculated by Fourier transform methods, since it is effectively the square of the diffraction pattern of the aperture distribution. For antennas that are smaller than a few wavelengths in any direction perpendicular to that of the incident radiation, the calculation is more complicated and requires the application of electrodynamics. Table 7.1 lists some of the properties of the power patterns of some common designs of antenna. The normal measure of the strength of the sidelobes is the peak value of P_n , expressed in decibels. Thus it may be said that a well-designed antenna will have sidelobe levels of -20 dB or lower.

The table shows the HPBW measured in two orthogonal directions ('iso' means isotropic, i.e. no variation of P_n in this plane), the directivity D and the maximum sidelobe level for some of the most important designs of antenna. For the last two types, it is assumed that the physical dimensions a , b and d are much greater than the wavelength λ of the radiation.

Table 7.1. Properties of some common types of microwave antenna

Antenna type	HPBW (degrees)	D (dB)	Sidelobes (dB)	Notes
Monopole	(iso)	(iso)	0	(none)
Short dipole	90	(iso)	1.76	(none)
Half-wave dipole	90	(iso)	2.15	(none)
Six-element Yagi	42	(iso)	10.5	-10 TV-antenna type
Rectangular	$51(\lambda/a)$	$51(\lambda/b)$	$11 + 10 \log_{10} (ab/\lambda^2)$	a, b are sides of rectangle
Circular paraboloid	$72(\lambda/d)$	$72(\lambda/d)$	$\sim 10 + 20 \log_{10} (d/\lambda)$	d is diameter

From our definition of the power pattern, it follows that the antenna temperature will be given by

$$T_A = \frac{\int_{4\pi} T_b(\theta, \phi) P_n(\theta, \phi) d\Omega}{\int_{4\pi} P_n(\theta, \phi) d\Omega},$$

where $T_b(\theta, \phi)$ is the brightness temperature of the ‘target’ in the direction (θ, ϕ) , and the integrals are carried out over all directions (i.e. over a range of Ω of 4π steradians). In fact, it is convenient to use the *beam solid angle*, defined as

$$\Omega_A = \int_{4\pi} P_n(\theta, \phi) d\Omega, \quad (7.3)$$

so the formula for the antenna temperature becomes

$$T_A = \frac{\int_{4\pi} T_b(\theta, \phi) P_n(\theta, \phi) d\Omega}{\Omega_A}. \quad (7.4)$$

Related to the concept of the beam solid angle is the *directivity* D , defined as

$$D = \frac{4\pi}{\Omega_A}. \quad (7.5)$$

This is a measure of the sensitivity of the antenna in its most sensitive direction. For a monopole antenna, which is completely isotropic, the beam solid angle is 4π steradians and the directivity is therefore 1. The directivity, which is often specified in decibels, is therefore a measure of maximum sensitivity relative to an isotropic antenna. Typical values of the directivity are shown in Table 7.1.

Up to this point, we have been assuming that the antenna is loss-free. In a real antenna, resistive losses will reduce the detected power. These losses can be represented by

7.1 Antenna theory

specifying an efficiency η , which is the ratio of the detected power to the received power, or equivalently by specifying the *forward gain* G of the antenna, where

$$G = \eta D. \quad (7.6)$$

It was mentioned earlier that the power pattern $P_n(\theta, \phi)$, and hence the beam solid angle Ω_A , can be determined from the dimensions of the antenna, either by Fourier transform methods or, for small antennas, by electrodynamic calculations. However, there is an alternative method of deriving Ω_A , and this introduces an important new concept – the *effective area* of the antenna.

From our considerations in Section 2.6 we know that, provided the Rayleigh–Jeans approximation is valid, the spectral radiance corresponding to a brightness temperature T_b is

$$L_f = \frac{2kT_b}{\lambda^2}, \quad (7.7)$$

where k is the Boltzmann constant and λ is the wavelength of the radiation. Thus, if we have a uniform source of brightness temperature T_b that subtends a small solid angle $\Delta\Omega$, the spectral flux density reaching the antenna is

$$F_f = \frac{2kT_b\Delta\Omega}{\lambda^2}. \quad (7.8)$$

Next, we define the effective area A_e of the antenna such that, for radiation of spectral flux density F_f incident along its direction of maximum sensitivity, the power collected per unit frequency interval is $F_f A_e/2$. (The factor of 1/2 arises because the radiation is assumed to be randomly polarised, and the antenna is assumed to receive only one polarisation state.) If we now allow for the variation in the antenna's sensitivity that is expressed by the power pattern, and the possibility that the brightness temperature of the incident radiation can vary with direction, the power collected per unit frequency interval must be given by

$$\frac{kA_e}{\lambda^2} \int_{4\pi} T_b(\theta, \phi) P_n(\theta, \phi) d\Omega.$$

From Equation (7.1), it therefore follows that the antenna temperature must be given by

$$T_A = \frac{A_e}{\lambda^2} \int_{4\pi} T_b(\theta, \phi) P_n(\theta, \phi) d\Omega,$$

and we can equate this to Equation (7.4) to derive the following remarkably simple result:

$$\Omega_A A_e = \lambda^2. \quad (7.9)$$

This shows that an antenna with a large effective area will have a small beam solid angle, and conversely. For an antenna that is large compared with the wavelength λ , the effective area A_e will be approximately equal to the geometrical area of the antenna, and in this case Equation (7.9) is obviously plausible. For example, if the antenna is a circular dish of diameter d , we know from the discussion of diffraction in Section 2.7 that the main lobe of the power pattern has an angular radius of $1.22\lambda/d$. The beam solid angle is thus of the order of $(\lambda/d)^2$ steradians, and since the effective area is clearly of the order of d^2 , we can see that Equation (7.9) is at least plausible in this case.

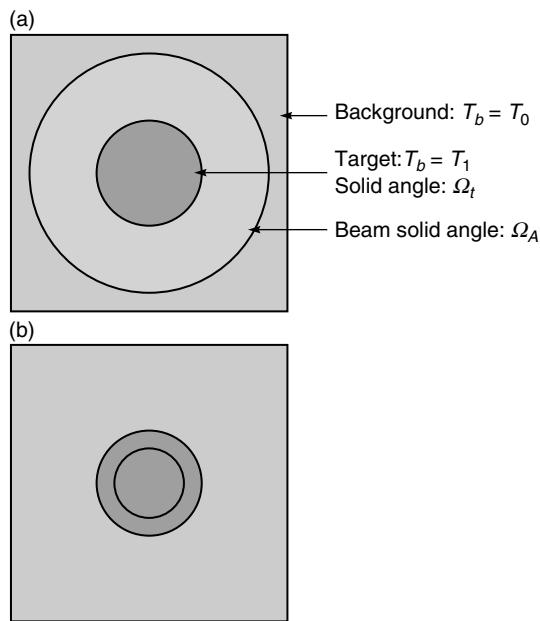


Figure 7.3 A microwave radiometer views of a target of brightness temperature T_1 against a background of brightness temperature T_0 . In (a) the target is not resolved, and the antenna's field of view includes some of the background, whereas in (b) the target is resolved.

We can discuss the implications of the foregoing theoretical development by considering a simple example. We suppose that the antenna views a circular ‘target’ of brightness temperature T_1 against a background of brightness temperature T_0 . To simplify Equation (7.4), we assume that the power pattern of the antenna is 1 for all values of θ up to some maximum value, and zero above this value, so that the antenna responds uniformly to a cone of directions subtending a solid angle of Ω_A , and does not respond at all to radiation arriving outside this cone. Figure 7.3 shows two situations: in Figure 7.3a the target, which we assume subtends a solid angle of Ω_t , is not resolved, whereas in Figure 7.3b it is fully resolved as a result of the antenna’s smaller beam solid angle.

From Equation (7.4), we see that in the first (unresolved) case, the antenna temperature will be given by

$$T_A = \left(1 - \frac{\Omega_t}{\Omega_A}\right)T_0 + \frac{\Omega_t}{\Omega_A}T_1.$$

This is just a linearly weighted average of the brightness temperatures of the components in the antenna’s field of view. If the antenna is made physically larger, thus reducing the size of its beam solid angle, the weighting will shift in favour of T_1 . However, once the target becomes fully resolved (i.e. when $\Omega_A \leq \Omega_t$), the antenna temperature is just

$$T_A = T_1.$$

It perhaps seems surprising that a further increase in the effective area of the antenna does not result in its detecting more power. However, this can be understood by

7.1 Antenna theory

realising that although a larger antenna can indeed collect more radiation, it will do so from a smaller range of directions.

7.1.2 Sensitivity

We have noted that it is usual to express the power per unit frequency interval that is detected by an antenna as a temperature (the antenna temperature). This facilitates calculation of the sensitivity of the system, since the noise power generated by the system itself is also normally expressed as a temperature. The system noise temperature depends on the detailed design of the receiver and other parts of the system, but it cannot be lower than the physical temperature of the receiver and will usually (at the frequencies typical of passive microwave radiometry, and for a well-designed receiver) be a factor of 1.2 to 2 times this value.

In order to improve its signal-to-noise ratio, the output from a radiometer is integrated (averaged) for some time Δt . If the bandwidth (frequency interval) over which the radiation is received is Δf , this is effectively an average over $N = \Delta t \Delta f$ independent samples. We thus expect the signal-to-noise ratio to be improved by a factor of \sqrt{N} , and hence that the sensitivity of the system should be given by

$$\Delta T = \frac{CT_{\text{sys}}}{\sqrt{\Delta t \Delta f}}. \quad (7.10)$$

In this expression, ΔT is the smallest change in antenna temperature that can be detected, T_{sys} is the system noise temperature, and C is a factor of the order of 1 that depends on both the design of the radiometer and the criterion used to define ΔT . C usually has a value of 5 to 10, and the values of Δt and Δf are normally chosen to give a value of ΔT of the order of 1 K.

7.1.3 Scanning radiometers

Let us consider a passive microwave radiometer operating from a satellite at an altitude of 800 km, at a frequency of 10 GHz. The wavelength is therefore 3 cm, so even if the width of the antenna is as large as 1 m the spatial resolution at the Earth's surface will be of the order of 20 km. The region of sensitivity at the Earth's surface is usually termed the *footprint*, although the concept is equivalent to the instantaneous field of view (IFOV) introduced in Section 6.1.2.

Now suppose that we wish to image a strip of the Earth's surface, perhaps of the order of 1000 km wide, with a spatial resolution of 20 km. It is obvious that constructing an array of antennas, pointing in different directions, is an entirely impractical approach since 50 antennas, each occupying an area of the order of 1 m² and each adding significantly to the mass of the instrument, would be needed. However, scanning of the radiometer footprint can be achieved, either by mechanical or electrical means.

The obvious method of scanning a radiometer footprint is by mechanical steering of the antenna. The antenna itself (or part of it – for example a reflector) may rotate or oscillate with respect to the rest of the instrument, or the whole platform can be made to rotate. The latter approach is clearly more appropriate for spaceborne than for airborne platforms. The usual form of mechanical scanning is the *conical scan*, in which the antenna beam is rotated around the nadir direction at a fixed angle to the nadir direction. This is illustrated



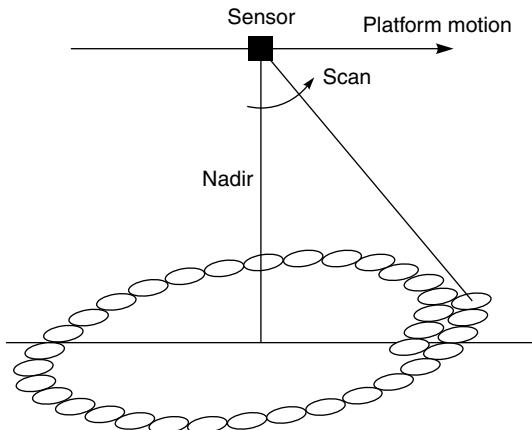


Figure 7.4. Conical scanning of a passive microwave radiometer.

in Figure 7.4. The received radiation makes a constant angle with the nadir direction, which simplifies the task of atmospheric correction.

One potential disadvantage of mechanical scanning is that it may cause some undesirable oscillation or vibration of the instrument, leading to pointing errors and loss of spatial resolution. An alternative approach to scanning the beam that avoids this problem is to use an electrically scanned antenna. This has no moving parts. It consists of a closely spaced array of smaller antennas (e.g. waveguides, horns or dipoles). The signals detected at each of these elements can be advanced or retarded in phase under electronic control (hence the alternative name of *phased array* for this type of system), and in this way beam-steering is achieved.

Figure 7.5 illustrates the principle of electrical steering in one dimension. It shows a simplified array, having only eight detectors at a regular spacing d . The signals from these detectors are shifted in phase and then added together. If no phase shifts are applied, and the elements abut one another so that there are no holes in the array, the device clearly functions as an ordinary antenna of width $8d$, having a power pattern with a main lobe pointing in the direction $\theta = 0$ and a width of approximately $\lambda/8d$ radians.

If, however, we consider radiation arriving at angle θ , it is clear that the phase of the signal received at detector 2 is $kd \sin \theta$ in advance of that at detector 1 (k is the wavenumber $2\pi/\lambda$), the phase at 3 is in advance of that at 2 by the same amount, and so on. In order to add these signals constructively, the phase delays ϕ_1 to ϕ_8 are introduced that will exactly compensate for this. Thus, by introducing a phase gradient across the ray, the direction of the main lobe can be shifted. This system was used by the Electrically Scanned Microwave Radiometer (ESMR), carried on the NIMBUS-5 satellite in the 1970s. More recently it has been used on the MIRAS (Microwave Imaging Radiometer with Aperture Synthesis), carried on the SMOS satellite since 2009. MIRAS is scanned electronically in two dimensions.

Electronic steering of an array has some disadvantages. One is the increased complexity of the system, since in practice it will have many more than eight elements, probably steerable in two dimensions instead of just one. A second disadvantage is the decrease in

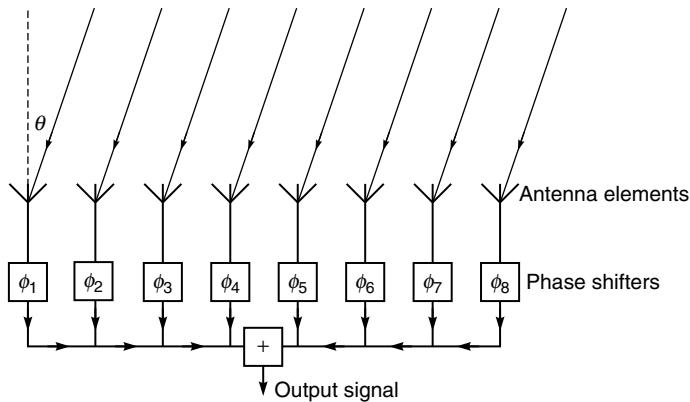


Figure 7.5 An electrically-steered (phased) array. The signals from the antenna elements are given phase shifts ϕ_1 to ϕ_8 so that radiation from the direction θ is combined in phase.

performance caused by errors in setting the phases and gains of the phase shifters. However, a problem that is fundamental rather than technological is caused by the fact that the *projected* length of the antenna, measured perpendicularly to the direction of the incoming radiation, is proportional to $\cos \theta$, so that as the scan angle is increased the HPBW will widen accordingly. This can be an important consideration if the instrument is required to have a very wide imaging swath.

Summary

The angular response of a microwave antenna is controlled by the physical dimensions of the antenna and can be specified by the antenna's *power pattern* or more simply by the *solid angle* of the beam, Ω_A . The *directivity* of the antenna is inversely proportional to the beam solid angle. For an antenna whose width w is large relative to the wavelength λ of the radiation, the angular beamwidth is roughly λ/w , set by diffraction. The effective area of the antenna is given by λ^2/Ω_A .

The response of an antenna to a microwave signal is represented by its *antenna temperature*, which is a measure of the power extracted from the incident radiation. If the antenna receives radiation with a uniform brightness temperature, the antenna temperature is simply equal to the brightness temperature of the radiation (apart from any resistive losses). If the antenna views a non-uniform distribution of brightness temperature, the antenna temperature is an average of the brightness temperature, weighted according to the power pattern. The radiometric sensitivity of a passive microwave radiometer is also expressed in terms of temperature, as the smallest detectable change in the antenna temperature.

Passive microwave radiometers are normally scanned mechanically, by arranging for the direction of the beam to be swung around a conical arc while the platform (e.g. satellite) carries the instrument forward. However, electrical scanning is also possible in the case of a *phased array*, consisting of a number of detectors.

7.2

Applications of passive microwave radiometry

Passive microwave radiometry is applied to both observation of the Earth's surface and to atmospheric sounding. In this section we discuss surface observations; atmospheric applications are described in Section 7.5.

7.2.1

Oceanographic applications

The main applications of surface imaging by passive microwave radiometers are oceanographic (Barale, Gower and Alberotanza 2010), particularly the determination of sea surface temperature (SST). This can be determined with a relative accuracy of about 0.2 K, and an absolute accuracy of about 1 K if careful atmospheric correction is performed. Since the brightness temperature of the ocean surface is influenced not only by the physical temperature (the SST) but also by the observation frequency and polarisation, the salinity, surface roughness, and the presence of any surface materials such as slicks or foam, an accurate measurement of the SST requires a multifrequency observation, usually in two polarisations. The absorption length for microwaves in sea water is of the order of 1 cm (see Figure 3.1) – considerably greater than for thermal infrared radiation. A comparison of the SST measured using the two techniques thus has potential for investigating the surface temperature anomaly discussed in Section 6.6.1, although at present neither technique has sufficient accuracy for this to be a practical proposition. Figure 7.6 shows an example of SST data derived from passive microwave radiometry.

At low frequencies, passive microwave radiometry can be used to determine ocean salinity. Below about 5 GHz, the ionic conductivity of sea water increases the imaginary part of the dielectric constant significantly with respect to the value for pure water, thus increasing the Fresnel reflection coefficients and decreasing the emissivity (Figure 7.7). Until recently this has not really been a practicable technique for spaceborne use; a 1-m diameter antenna operated from an altitude of 800 km at a frequency of 1 GHz, low enough for the effect of salinity to be appreciable, would have a footprint of over 200 km. However, salinity measurement from airborne systems is practicable. More recently the L-band (1.4 GHz) MIRAS radiometer carried on the SMOS satellite has an effective diameter of about 10 m, giving it a useful spatial resolution at nadir of around 30 km.

Passive microwave observations of ocean surfaces can also be used to deduce the surface roughness of the ocean. In this context 'surface roughness' means surface waves, and these are generated by the action of wind on the ocean surface. Thus, indirectly, wind speed can be inferred from passive microwave measurements. In fact, the effect of wind speed appears to be mediated by two mechanisms: an increase in surface roughness, and an increase in the proportion of the surface covered by foam. Modelling the surface roughness effect is rather complicated because the ocean surface contains a whole spectrum of waves rather than a single dominant component (e.g. Stewart 1992). It is observed that while the horizontally polarised component of the emitted radiation is sensitive to the wind speed, the vertically polarised component is independent of surface roughness for viewing angles near 50° from nadir (e.g. Robinson 2010). Thus, a viewing angle of about 50°, and a dual-polarisation observation, offers scope for discriminating the effect of surface roughness from other influences on the observed brightness temperature. Rainfall in particular presents a difficulty

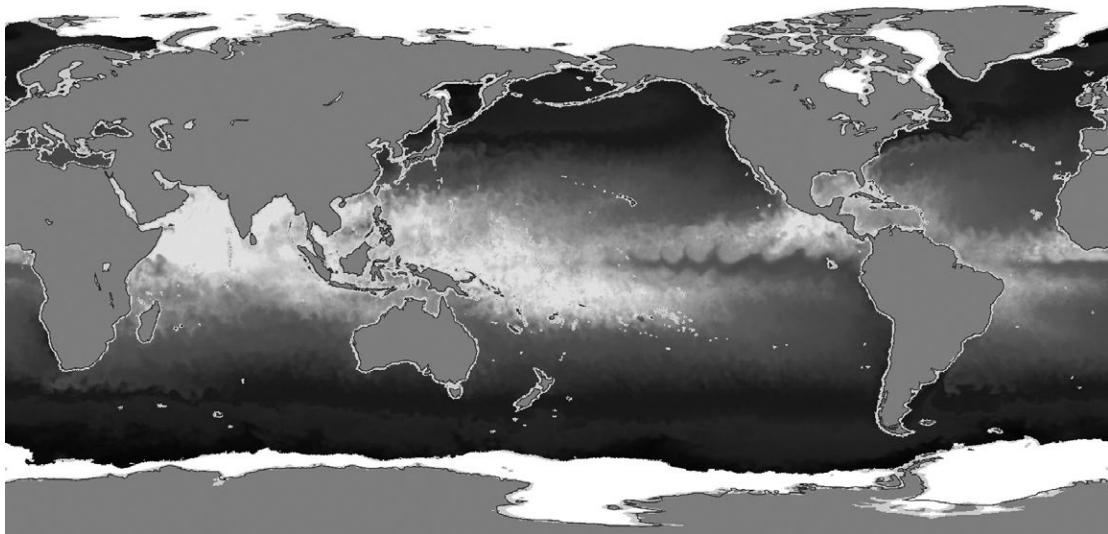


Figure 7.6. Global SST on 1 June 2003, determined using data from the AMSR-E microwave radiometer. The colour scale ranges from -2°C (dark green) to $+35^{\circ}\text{C}$ (bright yellow). Areas of sea ice are shown in white. Areas close to land are masked, reflecting the coarse spatial resolution of the data. Image downloaded from <http://aqua.nasa.gov/highlight.php?id=15> and reproduced by courtesy of NASA. See also colour plates section.

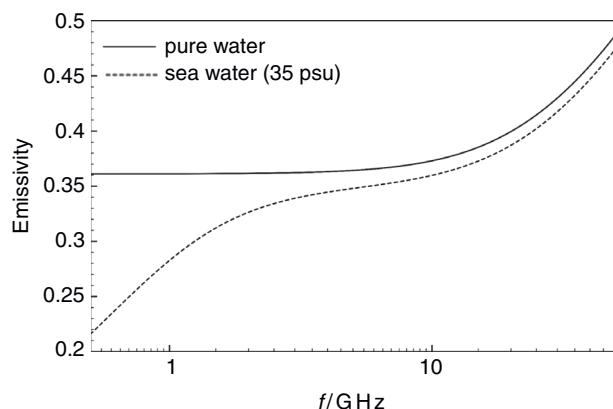


Figure 7.7. Normal emissivities at 20°C of pure water and sea water (35 practical salinity units) as functions of frequency.

in retrieving wind speeds, since it alters the roughness properties of the surface. In practice, estimation of wind speed from passive microwave radiometry over ocean surfaces is accurate to about $\pm 2 \text{ m s}^{-1}$ (Figure 7.8).

The last oceanographic application of passive microwave radiometry that we mention is in the identification of *sea ice* (Carsey 1992, Comiso 2009). At frequencies below about

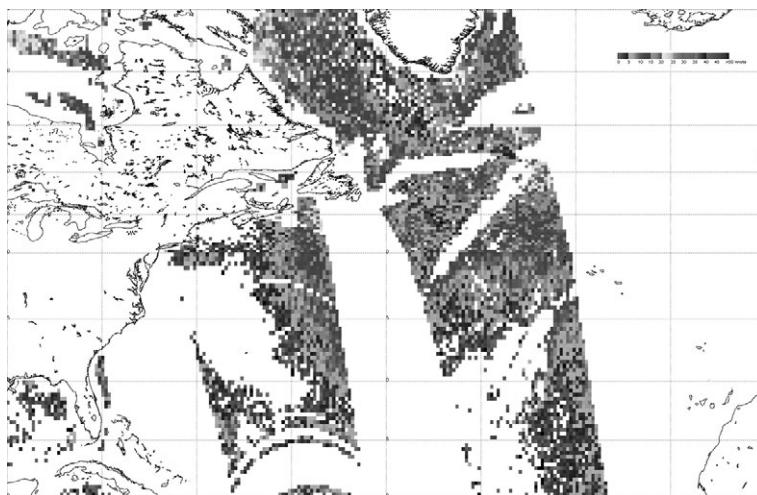


Figure 7.8. Surface wind speeds over the ocean determined from SSM/I passive microwave data. The data were recorded on 12 October 2011. Figure compiled from data supplied by NOAA NESDIS Center for Satellite Applications Research at <http://manati.orbit.nesdis.noaa.gov/datasets/SSMIData.php>. See also colour plates section.

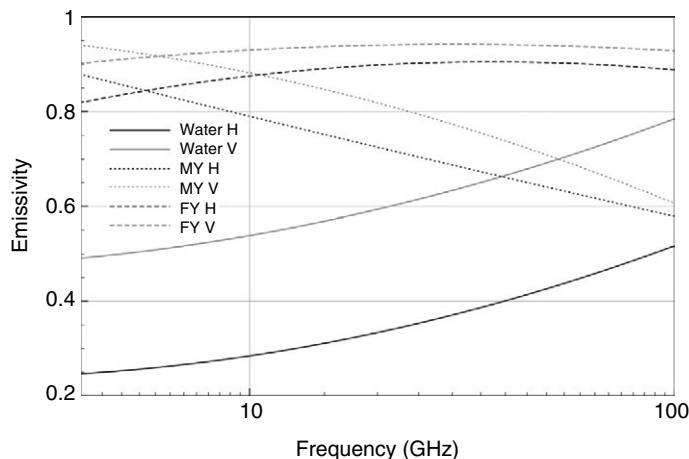


Figure 7.9. Emissivities of water, first-year sea ice (FY) and multi-year sea ice (MY) at 50° to nadir. The black curves show horizontal polarisation, the grey curves vertical polarisation. (The data for this figure have been adapted from Eppler *et al.* (1992) and from Rees (2006).)

30 GHz the emissivity of sea ice is significantly greater than that of sea water (see Figure 7.9), so the proportion of the ocean surface that is covered by ice can be determined rather easily. In fact, as shown in Figure 7.9, different types of sea ice generally have different emissivities (as sea ice ages, its internal structure, salt content and surface roughness all change), so a multifrequency observation can be used to

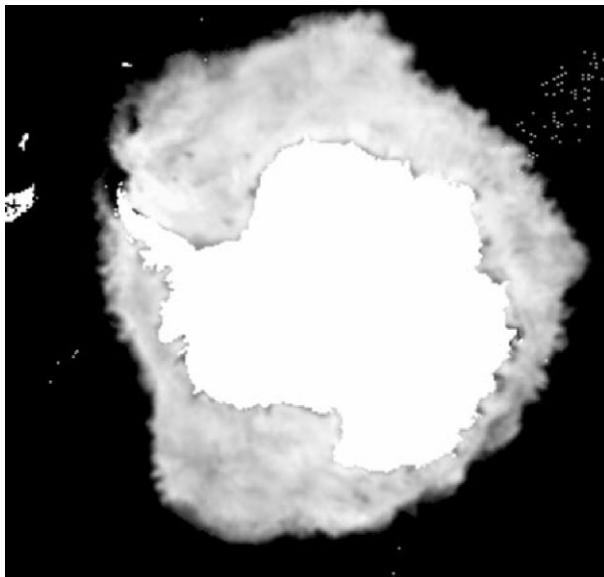


Figure 7.10. Sea-ice concentration in the southern hemisphere on 12 October 2011, calculated from SSM/I passive microwave data. (The image was derived from data provided by J. Maslanik, and J. Stroeve 1999, updated daily. *Near-Real-Time DMSP SSM/I-SSMIS Daily Polar Gridded Sea Ice Concentrations*. Boulder, Colorado, USA: National Snow and Ice Data Center. Digital media.)

estimate the proportions of different ice types and of open water present in the antenna's footprint. Figure 7.10 shows an example of this application.

7.2.2 Land surface applications

The relatively coarse angular resolution of passive microwave radiometry implies that, except from low altitude airborne platforms, the technique is not always particularly useful over land surfaces. As we have already remarked, a spaceborne system will have a spatial resolution of the order of 20 km or so, and while this is adequate for characterising many ocean features, the scale of spatial variability on land surfaces is generally finer than this so that the technique will often fail to resolve interesting features. Most obviously, passive microwave observations, even from space, can be used to measure the surface temperature of large, physically homogeneous regions (Figure 7.11) (Yang and Weng 2011). Vegetation parameters (primarily optical thickness, which can be related to above-ground biomass) can be estimated (Jones *et al.* 2011). Soil moisture content can also be estimated from low frequency measurements, since the presence of liquid water raises the dielectric constant, and hence lowers the emissivity, of soil (Njoku *et al.* 2003, 2010; Figure 7.12). The superficial extent and water equivalent of snowfields can also be estimated (Figure 7.13). Passive microwave imagery has also found a significant application in the identification of melt areas on the Greenland ice sheet (Abdalati and Steffen 1995). In this case, the algorithm exploits the difference in the brightness temperatures in two channels, calculating the 'cross-polarised gradient ratio' (XPGR):

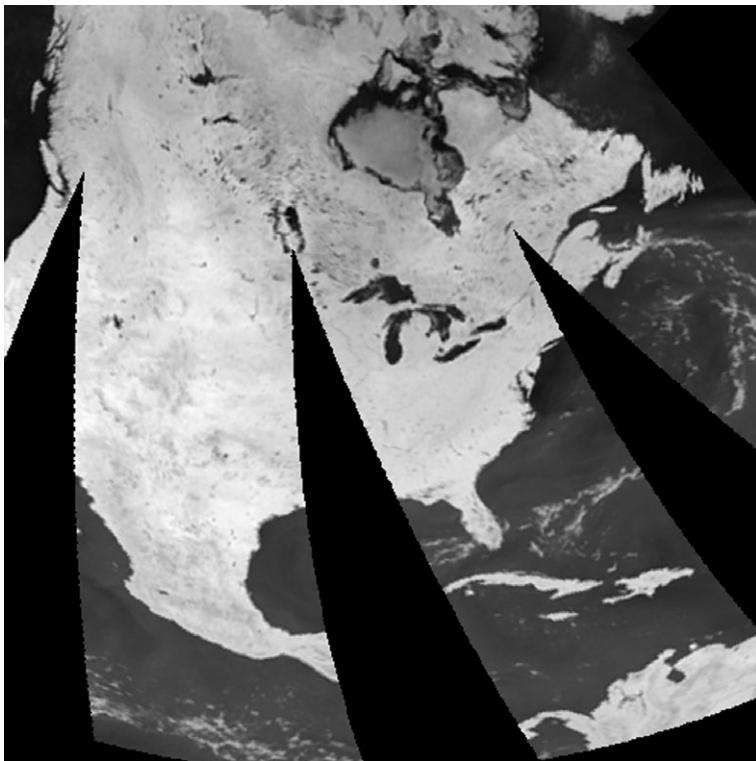


Figure 7.11. 18 GHz vertically polarised brightness temperature over North America on 30 June 2009, recorded by the AMSR-E passive microwave radiometer. (Data downloaded from ftp://sidads.colorado.edu/pub/DATASETS/nsidc0464_enhanced_ssmi_amsre_tbs/amsre/2009. (Long, D. and J. Stroeve and J. 2011. *Enhanced-Resolution SSM/I and AMSR-E Daily Polar Brightness Temperatures*. Boulder, Colorado, USA: National Snow and Ice Data Center. Digital media.)

$$\text{XPGR} = \frac{T_{19H} - T_{37V}}{T_{19H} + T_{37V}} \quad (7.11)$$

where T_{19H} is the brightness temperature of horizontally polarised radiation at 18 or 19 GHz and T_{37V} is the brightness temperature of vertically polarised radiation at 37 GHz. When the algorithm is applied to data from the SMMR instrument, which has a channel at 18 GHz, the threshold value is set to -0.027: values of XPGR above this threshold are assumed to correspond to melting snow, while values below this correspond to frozen snow. For the SSM/I instrument, which has a channel at 19 GHz, the corresponding threshold is -0.016. Figure 7.14 shows some results from this application.

However, applications that depend principally on the ability to estimate the dielectric constant of the material will also be influenced by the surface roughness. In such cases, more information can usually be obtained by the use of an *imaging radar* technique, discussed in Chapter 9.

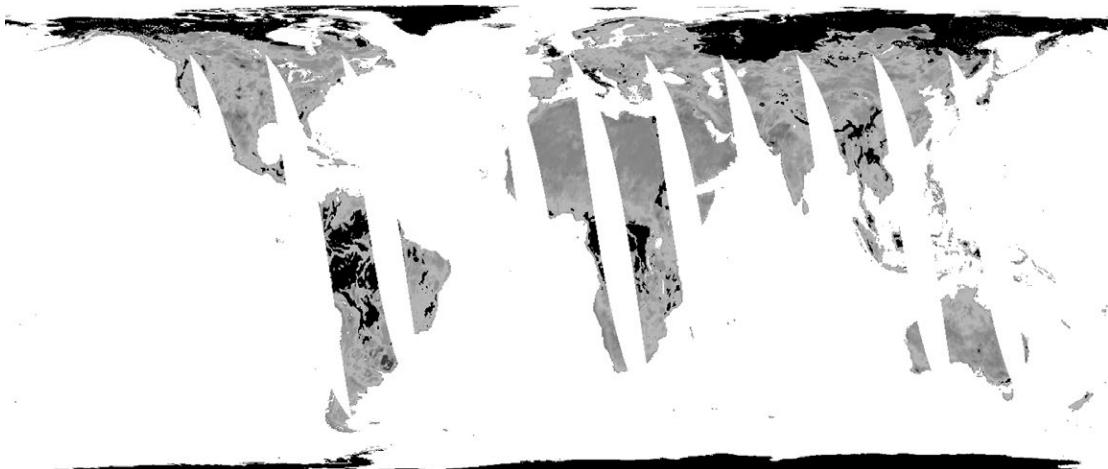


Figure 7.12. Global distribution of soil moisture on 1 April 2005, deduced from AMSR-E passive microwave data. The retrieval algorithm failed in the areas shown as black. (E. Njoku, 2004, updated daily. *AMSR-E/Aqua Daily L3 Surface Soil Moisture, Interpretive Parameters, & QC EASE-Grids V002*, [1 April 2005]. Boulder, Colorado, USA: National Snow and Ice Data Center. Digital media.) See also colour plates section.

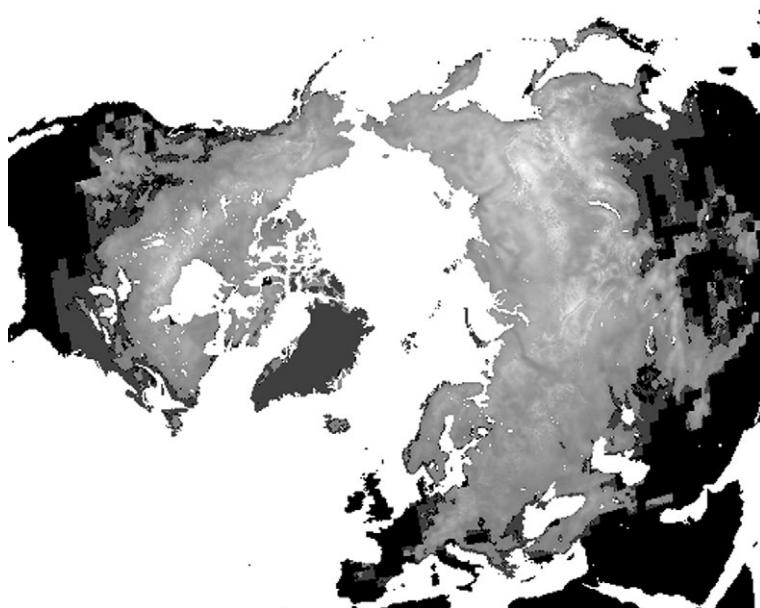


Figure 7.13. Northern Hemisphere snow water equivalent (SWE) in February 2005, calculated from SSM/I passive microwave data. Black areas had no snow; dark grey shows areas that had snow but for which the SWE could not be estimated from the microwave data; in other areas the greyscale corresponds to the SWE, ranging from 0 to 384 mm. Data from R. L. Armstrong, M. J. Brodzik, K. Knowles, and M. Savoie. 2007. *Global Monthly EASE-Grid Snow Water Equivalent Climatology*. Boulder, Colorado USA: National Snow and Ice Data Center. Digital media.

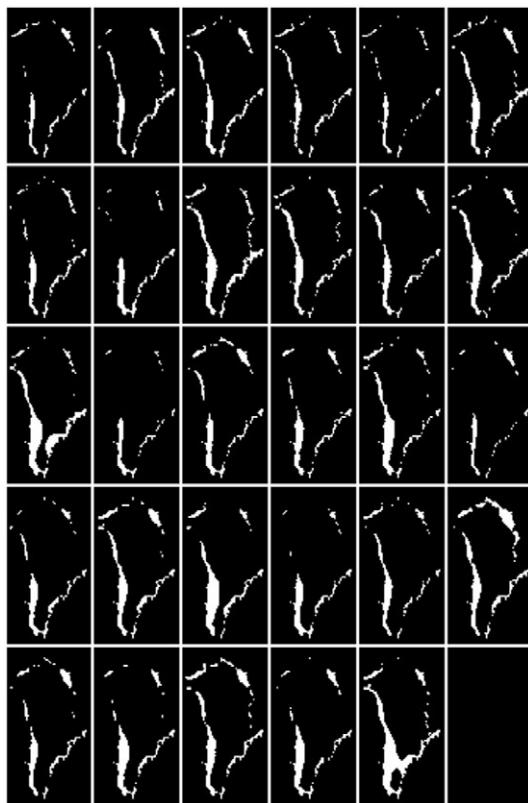


Figure 7.14. The figure shows (in white) areas of the Greenland ice sheet experiencing at least one month of surface melting per year, for consecutive years from 1979 to 2007. The figure has been derived from passive microwave data from the SSM/I and AMSR/E instruments. Data downloaded from ftp://sidads.colorado.edu/pub/DATASETS/parca/nsidc0218_melt/. Abdalati, W. 2007. *Greenland Ice Sheet Melt Characteristics Derived from Passive Microwave Data, 1979–2007*. Boulder, Colorado, USA: National Snow and Ice Data Center. Digital media.)

Summary

Passive microwave observations are used to determine the temperature of the Earth's surface, especially the sea surface temperature (SST). Multichannel observations are needed in order to identify and correct for the effect of factors such as salinity and surface roughness on the emissivity of the sea surface; these effects are valuable since they can also be used to infer surface roughness (and hence wind speed) and salinity, although sensitivity to salinity is confined to low frequencies. A major application of passive microwave radiometry over oceans is to the identification of sea ice, which has emissivity characteristics very different from those of open water.

7.3 Atmospheric correction of passive microwave imagery

Passive microwave imagery can also be used to measure land surface temperatures, although the comparatively coarse spatial resolution of spaceborne data (typically tens of kilometres) may fail to resolve important spatial variations. At low frequencies, soil moisture content can be estimated. Multichannel passive microwave imagery can be used to identify snow-covered areas and to estimate the amount of water stored in a snowpack, and to identify areas of surface melting on the Greenland ice sheet.

7.3

Atmospheric correction of passive microwave imagery

As we have seen in Chapter 4, the Earth's atmosphere is not entirely transparent in the microwave region of the electromagnetic spectrum, which means that the brightness temperature recorded at a spaceborne sensor will not correspond exactly to the quantity εT , ε being the emissivity of the surface and T its physical temperature. The measured brightness temperature can be considered to have three components: (1) the emitted component εT , which will be reduced by atmospheric attenuation as it propagates upwards to the detector; (2) downwelling atmospheric radiation that has been reflected from the surface (and is of course also attenuated on its upward journey through the atmosphere); (3) upwelling atmospheric emission. It is necessary to correct for components (2) and (3) if the first component is to be retrieved from the measured brightness temperature.

These effects can be modelled using the radiative transfer equation (Section 3.5). Since we are considering microwave radiation and ordinary terrestrial temperatures, the Rayleigh–Jeans approximation provides a considerable simplification. Let us first consider component (3), the contribution from upwelling atmospheric radiation, in the case of a clear atmosphere (one without any cloud, rain, snow etc. in it). From Equation (3.96) we can write this upwelling component as

$$T_b = \int_0^\tau T(\tau') \exp(-\tau') d\tau',$$

where τ is the total optical thickness of the path and τ' is the optical thickness between the surface and some point on the path where the physical temperature is $T(\tau')$. To obtain a rough idea of the magnitude of this contribution, we will assume that $T(\tau')$ is constant, so that the contribution is just

$$T_b = T(1 - \exp(-\tau)). \quad (7.12)$$

Figure 7.15 shows the result of applying equation (7.12) to the atmospheric absorption data of Figure 4.4, assuming an atmospheric temperature T of 250 K. Two curves are shown in the figure, one for radiation propagating vertically upwards (towards the zenith), and one for radiation propagating at an angle of 50° to the zenith. (We noted in Section 4.2 that the total optical thickness for a path that makes an angle of θ to the vertical is approximately $\tau/\cos \theta$, where τ is the optical thickness of the vertical path, provided that θ

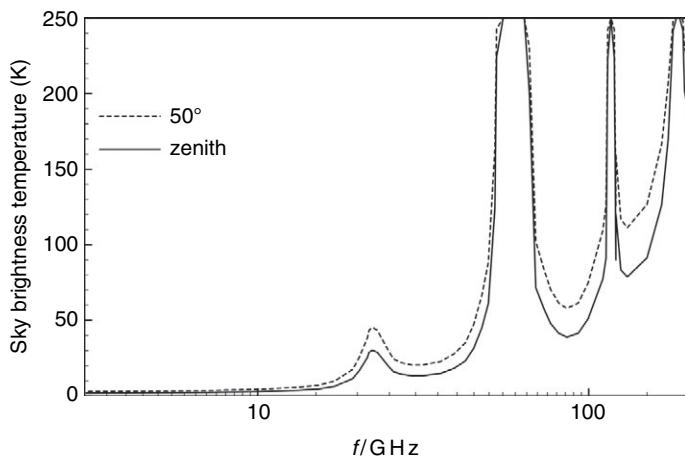


Figure 7.15. Typical upwelling brightness temperature of the atmosphere for radiation propagating vertically and at 50° to the zenith.

is less than about 75° .) We see that for frequencies below about 15 GHz the contribution is only a few kelvin, although it is significantly increased near the water vapour absorption lines at 22 GHz and 183 GHz and the oxygen lines at 60 GHz and 119 GHz. Thus, for frequencies below about 15 GHz, a fairly crude knowledge of the atmospheric temperature profile is sufficient to correct for this term.

Next, we consider the downwelling atmospheric radiation. Provided that no radiation enters the atmosphere from above, it is clear that this is given by the same calculation as for the upwelling radiation, so the data in Figure 7.15 also show the downwelling component. Again, for frequencies below about 15 GHz, this is just a few kelvin. If in addition the surface reflectance is small (i.e. the emissivity is large), only a small proportion of this already small component will be reflected. We should, however, make two remarks about the reflection of downwelling radiation. The first is that downwelling radiation will reach the surface from all directions, so it is necessary to integrate the contribution over all incidence directions. The radiation will in fact be most intense at large angles to the vertical, where the optical thickness of the path through the atmosphere is greatest. The second remark is that our assumption that no radiation enters the atmosphere from above is only valid for frequencies above about 3 GHz. Below this frequency, background radiation from our galaxy is also significant.

We have discussed corrections due to the clear atmosphere. We ought now also to consider the effects of particulate material. At microwave frequencies the effects of aerosols are negligible. The larger meteorological particles were discussed in Section 4.4, where it was shown that *fog* and *cloud* also have negligible effects for frequencies below about 15–20 GHz. Heavy *rain*, on the other hand, can introduce significant absorption and scattering effects at frequencies as low as 10 GHz. Since these are very strongly dependent on frequency (Figure 4.13), they can be estimated and corrected for by a suitable multifrequency observation.

Returning to Figure 7.15, or equivalently to Figure 4.4, we can say that at frequencies between about 3 and 15 GHz the signal detected by a passive microwave radiometer will

7.4 Examples: the SSMIS and the MSMR

normally be dominated by surface emission, with a correction of a few kelvin due to atmospheric water vapour. Between about 15 and 35 GHz, surface emissions are still generally dominant although the water vapour signal is significantly larger. Above 35 GHz, the effects of molecular absorption become dominant, and observations at these frequencies will be more useful for atmospheric sounding than for surface sensing.

Summary

The effect of the clear atmosphere on passive microwave radiometer measurements can be understood in terms of the radiative transfer equation in which only absorption and emission play a significant part. The total optical thickness τ of the atmosphere thus plays an essential role in determining the size of the effect. At frequencies well below the water vapour absorption peak at 22 GHz τ is normally below about 0.02 for vertically propagating radiation, so the correction is only a few kelvin. Between about 15 and 35 GHz the effect of the water vapour feature is appreciable (τ can be as large as 0.1 or 0.2). Above 35 GHz the background optical thickness increases steadily, and there are also deep absorption features due to oxygen at around 60 GHz and 120 GHz and to water at around 180 GHz. Imaging of the Earth's surface is not possible at frequencies close to these absorption features (they are valuable for atmospheric sounding, however), but can be performed between them. For example, the optical thickness of the atmosphere at around 90 GHz can be as low as around 0.2. For oblique viewing through the atmosphere, the optical thickness is increased so atmospheric effects on the signal are larger.

7.4

Examples: the SSMIS and the MSMR

In this section we describe a typical spaceborne passive microwave imaging radiometer. This is the Special Sensor Microwave Imager Sounder (SSMIS) carried on the DMSP series of satellites of the United States Air Force. The SSMIS is, as its name implies, a combined imager and sounding instrument. It combines the functions of a number of instruments carried on earlier DMSP missions, including the SSM/T (a microwave temperature profiling instrument) and the SSM/I (Special Sensor Microwave Imager). The SSM/I has been flown on DMSP missions since 1987 and is still operational, while SSMIS instruments have been operational since 2003.

The SSMIS is a conically scanned radiometer. It operates from an altitude of 833 km and the antenna axis makes an angle of 45° to the nadir. This gives an incidence angle at the Earth's surface of 53° and a range to the surface of 1270 km. The 'active scan', during which data are collected, consists of 144° of the conical path giving a swath width of about 1700 km. One revolution of the conical scan takes 1.9 seconds, during which time the footprint moves forward by 12.5 km. The sampling interval is 4.2, 8.4, 12.6 or 25.2 milliseconds, giving a spatial sampling interval of 12.5, 25, 37.5 or 75 km. The antenna diameter is 1 m, giving a beamwidth of approximately 0.035 radians (2°) at 19.4 GHz, so at a distance of 1270 km and at an incidence angle of 53° this results in an elliptical

Table 7.2. Characteristics of the SSMIS instrument. The first five channels, italicised, are used primarily for imaging the Earth's surface and are shared by the SSM/I instrument (which also has H and V channels at 85 GHz). The remaining channels are used primarily for atmospheric sounding. In the 'polarisation' column, H and V denote horizontal and vertical linear polarisations respectively, while R denotes right circular polarisation. The next column shows the radiometric resolution expressed as a change ΔT in brightness temperature.

Frequency (GHz)	Bandwidth (MHz)	Polarisation	ΔT (K)	Resolution (km)
19.350	355	H	0.35	<i>73 × 47</i>
19.350	357	V	0.35	<i>73 × 47</i>
22.235	401	V	0.45	<i>73 × 47</i>
37.000	<i>1616</i>	H	0.22	<i>41 × 31</i>
37.000	<i>1545</i>	V	0.22	<i>41 × 31</i>
50.300	380	H	0.34	<i>27 × 18</i>
52.800	389	H	0.32	<i>27 × 18</i>
53.596	380	H	0.33	<i>27 × 18</i>
54.400	383	H	0.33	<i>27 × 18</i>
55.500	391	H	0.34	<i>27 × 18</i>
57.290	330	R	0.41	<i>27 × 18</i>
59.400	239	R	0.4	<i>27 × 18</i>
62.998–63.569	1.35	R	2.7	<i>27 × 18</i>
60.435–61.151	1.35	R	2.7	<i>27 × 18</i>
60.435–61.151	1.3	R	1.9	<i>27 × 18</i>
60.435–61.151	2.6	R	1.3	<i>27 × 18</i>
60.435–61.151	7.35	R	0.8	<i>27 × 18</i>
60.435–61.151	26.5	R	0.9	<i>27 × 18</i>
91.665	1411	H	0.19	<i>14 × 13</i>
91.665	1418	V	0.19	<i>14 × 13</i>
150.000	1642	H	0.53	<i>14 × 13</i>
182.311–184.311	513	H	0.38	<i>14 × 13</i>
180.311–186.311	1019	H	0.39	<i>14 × 13</i>
176.711–189.911	1526	H	0.56	<i>14 × 13</i>

footprint with dimensions of approximately 70×45 km. The details of the various channels used by both the SSM/I and the SSMIS instruments are given in Table 7.2.

The SSMIS operates entirely above 15 GHz, where the effects of atmospheric propagation are at least appreciable. Other currently operational imaging microwave radiometers also include lower frequency bands, for example the MIRAS instrument that we have already noted, and which operates only at L-band (1.4 GHz) and the MSMR (Multifrequency Scanning Microwave Radiometer).

The MSMR is carried on the Indian Space Research Organisation's Oceansat satellite, operational since 1999. It is a conically scanned radiometer with a swath width of 1360 km and V- and H-polarised channels at 6.6, 10.6, 18 and 21 GHz. The antenna diameter is 0.8 m. The characteristics of the channels are shown in Table 7.3.

Table 7.3. Characteristics of the MSMR instrument

Frequency (GHz)	Bandwidth (MHz)	Polarisation	ΔT (K)	Resolution (km)
6.6	350	V, H	1	105 × 68
10.65	100	V, H	1	66 × 43
18	200	V, H	1	40 × 26
21	400	V, H	1	34 × 22

Each scan of the MSMR takes 5.38 seconds, during which the satellite advances by 36 km with respect to the Earth's surface. The along-track sampling interval is 50 km for the two lower frequencies and 25 km for the two higher frequencies.

Summary

Most passive microwave radiometers designed to image the Earth's surface from space are multichannel instruments, operating at a range of frequencies and polarisations, and are enabled to scan a wide swath of the surface (typically 1000 km or more) by mechanical scanning of the beam direction around a cone. The normal arrangement is that the line of sight from the surface to the radiometer makes an angle of about 50° to the vertical. Since the beamwidth is controlled by diffraction, the spatial resolution is frequency-dependent. Typical figures are 50 km at a frequency of 10 GHz and 10 km at 100 GHz. Temperature sensitivities are typically 0.2–1 K unless the bandwidth of the channel is particularly narrow, in which case the sensitivity will be poorer.

7.5

Atmospheric sounding using passive microwave observations

Passive microwave radiometry has important applications in atmospheric sounding as well as in surface imaging, for the obvious reason that the microwave region contains a number of important absorption lines. The principles involved are not essentially different from those discussed in Section 6.7 for the optical and infrared regions of the electromagnetic spectrum.

Temperature profiling from nadir-viewing observations is normally performed by using one of the deep absorption lines of oxygen, for example at 60 or 118 GHz. The weighting functions have typical widths of 10 km in altitude at 60 GHz, and somewhat less than this at 118 GHz. As we saw in Section 6.7, the requirement for this type of profiling is an instrument having a number of frequency bands near the absorption line. For example, we noted in Table 7.2 that the SSMIS instrument has 13 channels between 50 and 61 GHz. Figure 7.15 shows that these correspond to progressively higher atmospheric optical thicknesses, so the peaks of the weighting functions occur progressively higher in the atmosphere. As an example of this type of instrument we can consider the Advanced

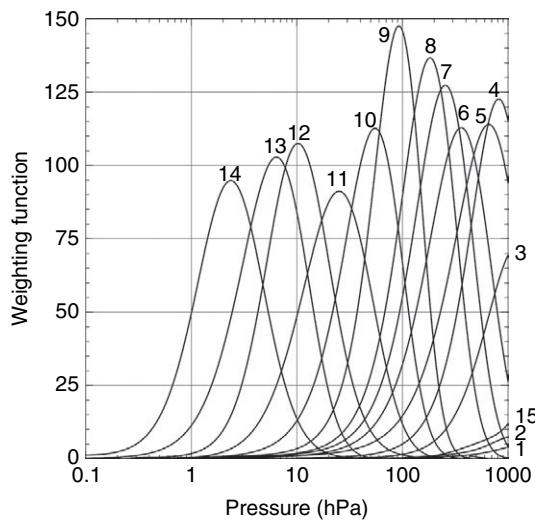


Figure 7.16. Weighting functions for AMSU-A. The curves are labelled by channel number.

Table 7.4. Characteristics of the AMSU-A microwave sounder

Channel	Frequency (GHz)	Bandwidth (MHz)	Polarisation	Bands	ΔT (K)
1	23.8	270	V	1	0.3
2	31.4	180	V	1	0.3
3	50.3	180	V	1	0.4
4	52.8	400	V	1	0.25
5	53.481–53.711	170	H	2	0.25
6	54.4	400	H	1	0.25
7	54.94	400	V	1	0.25
8	55.5	330	H	1	0.25
9	57.29	330	H	1	0.25
10	57.073–57.507	78	H	2	0.4
11	56.920–57.660	36	H	4	0.4
12	56.946–57.634	16	H	4	0.6
13	56.958–57.622	8	H	4	0.8
14	56.963–57.617	3	H	4	1.2
15	89	6000	V	1	0.5

Microwave Sounding Unit (AMSU)-A, which has been carried on the NOAA POES satellites since 1998, the Aqua satellite (since 2002) and the MetOp satellite (since 2006). AMSU-A is a 15-channel microwave radiometer; the channel characteristics are shown in Table 7.4, and the corresponding weighting functions in Figure 7.16. It is nominally a nadir-sounding instrument although it scans to 48 degrees either side of the nadir in 30 steps. The nominal horizontal spatial resolution is 48 km.

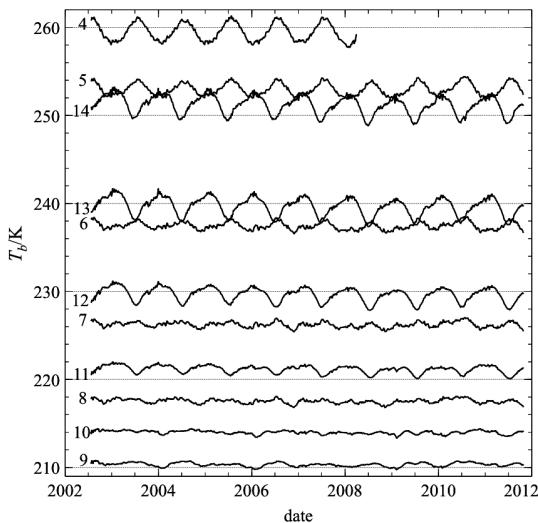


Figure 7.17. Globally averaged brightness temperatures measured by the AMSU-A radiometer carried on the NOAA-15 satellite. The traces are labelled with the corresponding AMSU-A channel numbers (the channel 4 sensor failed in 2008). Data downloaded from <http://discover.itsc.uah.edu/amsutemps/>.)

As can be seen from Figure 7.16, channels 1–3 and 15 correspond to relatively transparent regions of the atmosphere and are not useful for temperature sounding (channels 1–3 are used mainly for assessing the total amount of water vapour in the atmosphere and channel 15 for measuring cloud top and snow surface temperatures). Channels 4 to 9 have centre frequencies from 53 to 57 GHz, corresponding to progressively higher total atmospheric optical thickness and hence to weighting functions with peaks at progressively lower atmospheric pressures. Channels 10 to 14 are centred on a resonant absorption feature at 57.29 GHz; by using progressively smaller bandwidths they become progressively more heavily weighted towards the most opaque part of the spectrum and hence to lower pressures. One disadvantage of the use of very small bandwidths is the increase in noise temperature, as implied by Equation (7.10). Figure 7.17 illustrates the use of data from the AMSU-A radiometer.

At frequencies below about 200 GHz, profiling of atmospheric molecular species is limited to oxygen and water vapour, since no other absorption lines occur in this region (see Figure 4.4). Since oxygen is well mixed in the atmosphere, profiling of this molecule is equivalent to establishing the atmospheric density profile, and determination of the total (column-integrated) oxygen concentration is equivalent to establishing the surface atmospheric pressure. As an example of an instrument for profiling water vapour, we can consider the MHS (Microwave Humidity Sounder) that has been carried on the NOAA POES and MetOp satellite programmes since 2005 and 2006 respectively. This has five frequency channels: 89 GHz, 157 GHz and three at 183.3 GHz. The characteristics of these channels are summarised in Table 7.5. Channels 3–5 exploit the water vapour absorption feature centred at around 183 GHz. The horizontal resolution of the instrument

Table 7.5. Characteristics of the MHS microwave sounder

Channel	Frequency (GHz)	Bandwidth (MHz)	Polarisation	Bands	ΔT (K)
1	89	2800	V	1	0.22
2	157	2800	V	1	0.34
3	182.311–184.311	500	H	2	0.51
4	180.311–186.311	1000	H	2	0.4
5	190.31	2200	V	1	0.46

is 16 km and the field of view is scanned to 49.4° either side of nadir in 90 steps, giving a swath width of around 2200 km.

Limb-sounding observations can also be made using passive microwave radiometry. In general, the principles are the same as those discussed in Section 6.7.4. The vertical resolution of the technique is substantially better than for nadir-viewing observations, but this is at the expense of the horizontal resolution (the long path through the atmosphere does, however, mean that constituents with very low concentrations can be detected) and is also dependent upon the sensor having sufficient angular resolution. Since the distance from a spaceborne sensor to the point in the atmosphere from which most of the signal is derived is about 3300 km, a beamwidth of 0.3 milliradians is needed to achieve a vertical resolution of 1 km, and this of course is much more difficult to arrange for a microwave system than for one operating in the infrared. As an example, we can consider the MLS (Microwave Limb Sounder) instrument carried on the Aura satellite since 2004. This has channels at around 118, 190, 240 and 640 GHz and at 2.5 THz. The dimensions of the primary mirror are 0.8×1.6 m. Temperature profiles are retrieved using the 118 and 190 GHz channels, achieving an effective vertical resolution of about 5 km in the lower atmosphere (pressures between 100 and 260 hPa), rising to around 10 km in the upper atmosphere (0.001–0.1 hPa). Water vapour concentrations are retrieved using the 190 GHz channel, achieving effective vertical resolution of around 2.5–3.5 km.

In Figure 4.4, we saw that the part of the microwave spectrum below about 200 GHz is sparsely populated by molecular absorption lines. Above this frequency, and especially above 300 GHz (submillimetre wavelengths), the situation changes dramatically and the region is very densely populated with molecular rotational transitions. These include H_2O , O_2 , CO , SO_2 , N_2O and NO_2 , as well as a large number of less abundant but nevertheless significant molecules and ions such as ClO and HCO^+ . Limb sounders have been designed to exploit this abundance of absorption lines, and indeed this is the reason why the MLS includes channels above 240 GHz. Another example is provided by the Submillimetre Radiometer (SMR) carried on board the Odin satellite (launched in 2001). This is also a limb-sounding instrument, with a channel at 118.7 GHz and four between 480 and 580 GHz. The channels are tuneable over a few GHz. The SMR is particularly useful for profiling water vapour and ozone between altitudes of around 5 and 100 km, but is also used to profile some other gases important in atmospheric ozone chemistry. It achieves a vertical resolution of 1.5–3 km and a horizontal resolution of typically 600 km. The design of instruments that can operate in the millimetre and submillimetre bands presents considerable technical challenges. This is especially true in the submillimetre

range, where the instruments will typically use a hybrid of radio and optical techniques. A discussion of these is beyond the scope of this book, although Elachi and van Zyl (2006) provides some technical details.

Summary

The principles of atmospheric sounding using passive microwave radiometry are very similar to those discussed in Chapter 6. Below around 200 GHz there are very few atmospheric absorption lines to choose from, due only to oxygen and to water vapour. Temperature sounding at nadir usually exploits the oxygen absorption line at 60 GHz or the one at 118 GHz: by observing at a number of frequencies progressively further way from the absorption maximum, the sounding instrument can ‘see’ progressively deeper into the atmosphere (i.e. the peak of the weighting function occurs at progressively lower altitudes or higher pressures). Nadir sounding achieves a vertical resolution of typically around 10 km; limb-sounding gives higher vertical resolutions, down to a few kilometres, at the expense of poorer horizontal resolution. At frequencies above 300 GHz (submillimetre wavelengths) there are many absorption lines that can be used to profile atmospheric chemistry.

Review questions

Give an outline of the principles of operation of passive microwave radiometry, and its main applications.

Explain why a passive microwave radiometer can *detect* a smaller object than it can *resolve*.

Compare the different methods available to scan the beam of a passive microwave radiometer.

Describe the main applications of spaceborne passive microwave radiometry to studies of the ocean and land surface.

Discuss the concept of atmospheric windows in relation to passive microwave radiometry.

Compare nadir sounding and limb-sounding as techniques for atmospheric profiling using microwave radiometry.

Explain why microwave sounding instruments that operate below 200 GHz can be used only to measure temperature and water vapour profiles, whereas at higher frequencies many other molecular species can be measured.

Problems

1. Use the data of Table 7.1 to show that the effective area of a rectangular or circular paraboloidal antenna is roughly equal to its geometrical area.
2. A phased array is designed to operate at a wavelength of 6 cm. The antenna elements are spaced at intervals of 4 cm. (i) Show that if the beam is steered more than 30° away from the normal ($\theta = 0$) direction it will respond to radiation from *two* directions instead of just one. (ii) Show that this phenomenon of multiple responses can be eliminated if the antenna elements are less than half a wavelength apart.
3. At 10 GHz and a certain incidence angle, the apparent brightness temperatures of sea water, first-year ice and multi-year ice are 80 K, 252 K and 200 K respectively. At 37 GHz these figures become 119 K, 253 K and 168 K. If a microwave radiometer measures a brightness temperature of 180 K at both frequencies, what are the fractions of open water and multi-year ice present in the IFOV?
4. A passive microwave radiometer operating at 37 GHz has an effective area of 0.3 m^2 , and can detect a change of 0.9 K in its antenna temperature. It is operated from an altitude of 800 km to observe a single ice floe (whose brightness temperature is 253 K) surrounded by water (brightness temperature 119 K). Calculate the area of the smallest *detectable* floe.
5. If the atmospheric microwave absorption coefficient γ varies with height z as

$$\gamma = \tau \beta \exp(-\beta z),$$

τ being the optical thickness of the entire atmosphere, show that the brightness temperature measured by a passive microwave radiometer looking vertically down through the atmosphere is given by

$$T(0)\exp(-\tau) + \int_0^\infty a(z)T(z)dz,$$

where the weighting function $a(z)$ is given by

$$a(z) = \tau \beta \exp(-\beta z) \exp\left(-\tau \exp(-\beta z)\right).$$

Show that $a(z)$ takes its maximum value at $(\ln \tau)/\beta$, and that, for a fixed value of β , the only effect of changing τ is to shift $a(z)$ along the z -axis without change of scale. Interpret this result.

6. The NASA Team algorithm for calculating sea-ice concentration from SSM/I data defines the *polarisation ratio* PR as

$$PR = \frac{T_{19V} - T_{19H}}{T_{19V} + T_{19H}}$$

and the *gradient ratio* GR as

$$GR = \frac{T_{37V} - T_{19V}}{T_{37V} + T_{19V}}$$

where, in both of these expressions, T_{fP} represents the at-satellite brightness temperature of P -polarised radiation at a frequency of f GHz. It is known from experimental measurements that open water shows a PR of typically 0.22 and a GR of 0.08, while multi-year ice shows a PR of 0.02 and a GR of 0, and first-year ice a PR of 0.04 and a GR of -0.09. Plot a graph of GR against PR and identify the points corresponding to open water, 100% first-year ice and 100% multi-year ice. Use your graph to suggest, with reasons, interpretations of SSM/I data giving the following values:

- (i) PR = 0.06, GR = 0.01,
- (ii) PR = 0.12, GR = 0.01.

Interpret the low values of PR for both ice types, and the low value of GR for multi-year ice, in terms of the respective emissivities.

8

Ranging systems

Chapters 5 to 7 considered *passive* sensors, detecting naturally occurring radiation. In this chapter and the next we shall discuss *active* sensors, which emit radiation and analyse the signal that is returned by the Earth's surface or atmosphere. We have already identified three possible classifications of remote sensing systems, distinguishing between passive and active and between imaging and non-imaging, as well as classifying them according to the wavelength of radiation employed. We can also classify active systems according to the use that is made of the returned signal. If we are principally concerned with the time delay between transmission and reception of the signal we shall call the method a *ranging technique*, whereas if we are also (or mainly) interested in the strength of the returned signal we shall call it a *scattering technique*. The distinction between the two cannot be made entirely rigorous, but it provides a useful way of thinking about active remote sensing systems. It is clear that ranging systems are simpler both to visualise and, because of their less stringent technical demands, to construct, and we shall therefore consider them first. In Chapter 9 we shall discuss the scattering techniques.

8.1

Laser profiling

Laser profiling (or *laser altimetry*) is the simplest application of the LiDAR (Light Detection and Ranging) technique. Conceptually it is extremely straightforward (Baltsavias 1999, Flood 2001). A short pulse of 'light' (visible or near-infrared radiation) is emitted towards the Earth's surface by the instrument, and its 'echo' is detected some time later. By measuring the time delay and knowing the speed of propagation of the pulse, the range (distance) from the instrument to the surface can be determined. By transmitting a continuous stream of pulses, a profile of the range can be built up, and if the position of the platform as a function of time is accurately known the surface profile may then be deduced.

The operation and construction of a typical laser profiler are shown schematically in Figures 8.1 and 8.2. The transmitter is a semiconductor laser, usually Nd:YAG (neodymium:yttrium-aluminium-garnet) operating at 0.53 μm or 1.06 μm , or GaAs (gallium arsenide) operating at 0.9 μm . This is capable of producing a short (of the order of 1 ns), intense pulse with a small angular width. The receiver is a photodiode (see Chapter 6). An interval timer with a resolution of the order of 1 ns is started by the signal that

8.1 Laser profiling

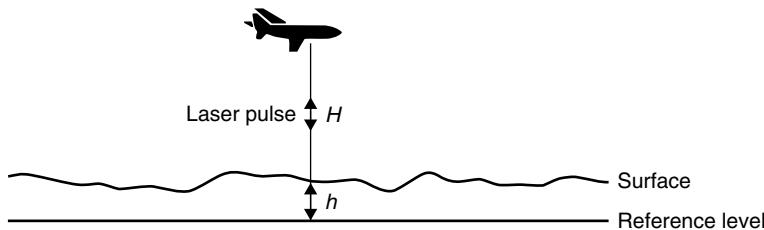


Figure 8.1. Principle of operation of a laser profiler.

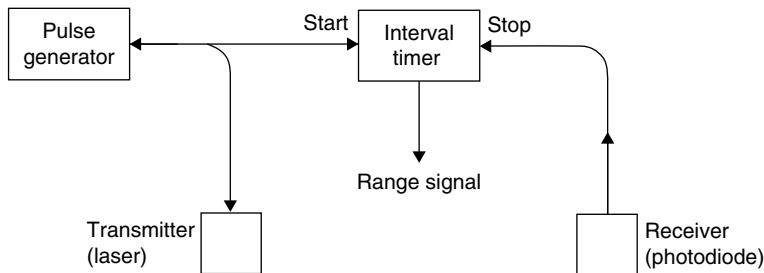


Figure 8.2. Construction of a laser profiler (schematic).

generates the transmitted pulse, and stopped by the received pulse. The travel time of the pulse, T_t , is given by

$$T_t = \frac{2H}{v_g}, \quad (8.1)$$

where H is the range and v_g is the group velocity (Section 3.1.3) of the pulse. As we saw in Figure 3.5, the group velocity for optical and near-infrared radiation propagating in dry air differs from c , the speed of light *in vacuo*, by at most 0.03%, so the error caused by setting $v_g \approx c$ in Equation (8.1) is small. Atmospheric correction of laser profile data is discussed in Section 8.1.3.

The desirable features of such a system are that it should achieve a high spatial resolution at the surface (i.e. the sampled points should be close together) and a high range resolution, and that the sensitivity should be great enough to detect signals returned from weakly reflecting surfaces.

The accuracy ΔT_t with which the travel time can be determined is normally governed by the *rise time* t_r of the received pulse, and its signal-to-noise ratio S . This can be understood from Figure 8.3. If V_s is the voltage amplitude of the received pulse and V_n is the amplitude of its variation due to noise, the (voltage) signal-to-noise ratio is defined as $S = V_s/V_n$. It is evident from the figure that the greatest accuracy with which the timing of the received pulse can be determined is given by

$$\Delta T_t = \frac{t_r}{S}, \quad (8.2)$$

although the precision of the system may in fact be limited by that of the interval timer.

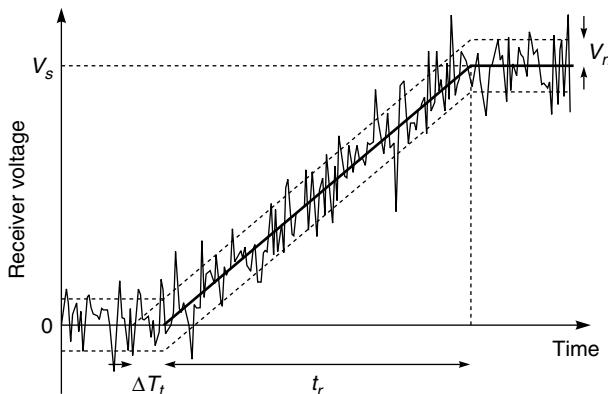


Figure 8.3. Determination of the arrival time of a noisy pulse. The accuracy ΔT_t with which the pulse can be located is $t_r V_n / V_s$.

A typical pulse transmitted by a laser profiler will have a rise time t_r of a few nanoseconds, although the received pulse may be somewhat longer if the surface that is being profiled is particularly rough. The signal-to-noise ratio S of the received pulse will depend on the reflectivity of the surface and the range H , as well as on system parameters such as the transmitted power, and less easily calculated influences such as the amount of incident sunlight, the weather, and atmospheric attenuation.

If pulses are transmitted at a frequency p (called the *pulse repetition frequency* or PRF) and the platform velocity is v , the linear sampling interval is v/p . If the angular beamwidth of the system is $\Delta\theta$, the linear dimension of the footprint is $H\Delta\theta$. It is obviously desirable that $H\Delta\theta$ should be no larger than some maximum value set by the nature and type of the surface under investigation. It might also be imagined that there would be no point in reducing the value of v/p below $H\Delta\theta$, but this is not so. Any decrease below this value means that, in effect, a number of independent measurements of the range are being made over a single footprint, and averaging over these measurements will therefore improve the range accuracy. Since the number of independent measurements is

$$N = \frac{H\Delta\theta}{v/p}$$

and the improvement in range accuracy is proportional to \sqrt{N} , we may (provided that $N > 1$) write the range accuracy as

$$\Delta H = \frac{v_g t_r}{2S} \left(\frac{v}{pH\Delta\theta} \right)^{1/2}. \quad (8.3)$$

For example, an airborne system might have $t_r = 5$ ns, $S = 1$, $v = 50$ m s⁻¹, $H = 200$ m and $\Delta\theta = 0.001$ radian. At a PRF $p = 1000$ s⁻¹, this would give $N = 4$ and $\Delta H = 0.38$ m. Of course, a further increase in range resolution can be obtained by averaging over more than four pulses (i.e. over more than one footprint, and hence at the expense of the horizontal resolution). The system we have just described has a horizontal resolution $H\Delta\theta$ of 0.2 m. By averaging over a horizontal distance of, say, 1 m, the range resolution can be improved to 0.17 m.

8.1 Laser profiling

Equation (8.3) shows that the measurement accuracy is increased by increasing the PRF. However, if p is increased beyond a certain point the measured range will become ambiguous, for if the travel time T , exceeds the interpulse period $1/p$ it will not be certain which echo belongs to which transmitted pulse. In this case, the calculated range will suffer from a *range ambiguity* of

$$H_{\text{amb}} = \frac{v_g}{2p} \quad (8.4)$$

in the sense that the calculated range H may be increased or decreased by integer multiples of H_{amb} without changing the apparent travel time (this is an example of the phenomenon of aliasing, discussed in more detail in Section 10.3.2.4). For this reason it is desirable to operate the system in such a way that $H_{\text{amb}} > H$, i.e. that

$$p < \frac{v_g}{2H}. \quad (8.5)$$

For an airborne system this imposes an upper limit of tens or hundreds of thousands of pulses per second, whereas for a spaceborne system it is a few hundred pulses per second.

8.1.1 Scanning laser profilers

Most airborne laser profilers are scanning instruments. One dimension of the scanning pattern is provided by the forward motion of the platform, while the perpendicular dimension is provided by an oscillating mirror or similar arrangement (e.g. a rotating polygonal mirror), analogous to the whiskbroom scanning described in Chapter 6. A common arrangement uses a plane mirror oscillating at some frequency f through an angle φ either side of nadir. To illustrate the principles of this kind of scanning, while minimising the mathematical complexity, we will suppose that the angular speed of the mirror is constant and that the angle φ is small enough for the small-angle-approximation to be valid. If the instrument is flown at a height H above a horizontal surface, the pattern of points on the ground from which data are collected will form a zig-zag arrangement, illustrated schematically in Figure 8.4.

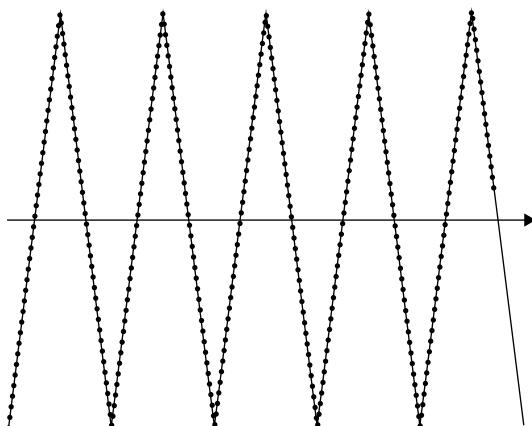


Figure 8.4. Sampling pattern of a scanning laser profiler (schematic). The arrow represents the direction of motion of the platform.

The swath width in this case is given by $2H\varphi$. The time taken for a one-directional scan by the mirror is $1/2f$, so the number of samples acquired during this scan is $p/2f$. The across-track spacing of the sampled points is thus given by

$$S_{\text{cross-track}} = \frac{4\phi f H}{p}. \quad (8.6)$$

During a one-directional scan the platform advances through a distance $v/2f$, and this determines the average sampling interval in the along-track direction:

$$S_{\text{along-track}} = \frac{v}{2f}. \quad (8.7)$$

The laser profiler would normally be operated in such a way as to keep these two sampling intervals roughly equal. For example, a scanning laser profiler might be operated with the following parameters: $v = 100 \text{ m s}^{-1}$, $H = 2000 \text{ m}$, $p = 33 \text{ kHz}$, $f = 19 \text{ Hz}$, $\varphi = 20^\circ = 0.35 \text{ radians}$. This would give values of 1.6 m and 2.6 m for the cross-track and along-track intervals respectively (see Figure 8.5).



Figure 8.5. Location of sampled points at the edge of the swath of an Optech ALTM3033 scanning laser profiler. The diagram shows an area of $400 \times 400 \text{ m}$ with a 100-m grid. (The study area is located near the village of Porsangmoen in the north of Norway.) Contours are labelled at 5-m intervals. The platform direction was from left to right. The observation parameters were as described in the text, giving average sampling intervals of 1.6 m in the cross-track direction and 2.3 m in the along-track direction.

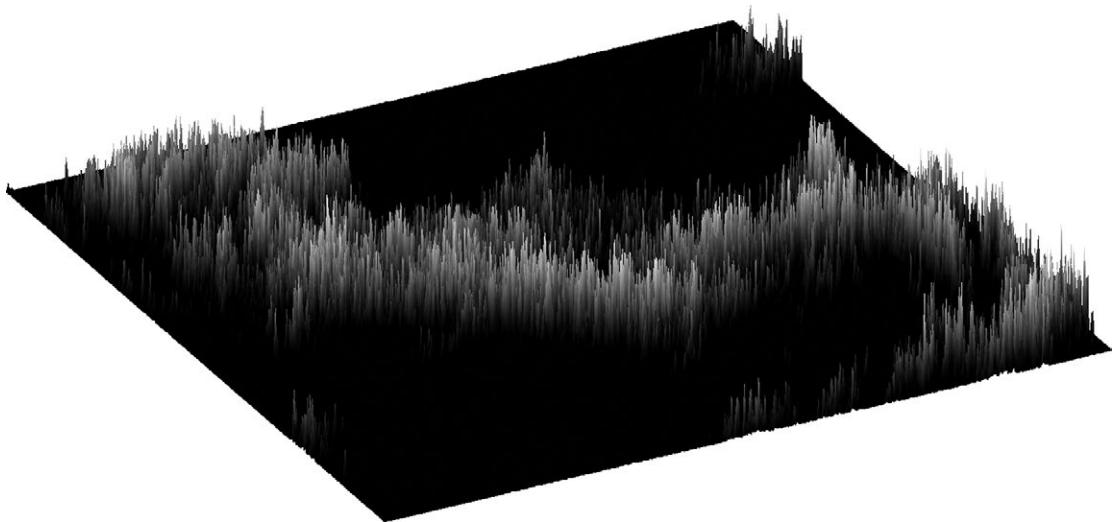


Figure 8.6. Difference between first and last return pulses over a $3000 \times 3000 \text{ m}^2$ area, visualised using both perspective view and greyscale shading. The study site is a partly forested area near Lakselv in northern Norway, so the difference in pulse return times represents the height of the forest canopy.

8.1.2

Waveform-resolving laser profiling

The essential variable recorded by a laser profiler is, as we have noted, the travel time of the pulse, or equivalently its arrival time at the instrument. However, more information is potentially available from the temporal structure of the return signal and some laser profilers are designed to exploit this. At the simplest level, a laser profiler may be able to detect more than one pulse returned for each pulse transmitted. This situation could arise, for example, if the footprint of the instrument illuminates a fairly open vegetation canopy, so that some parts of the pulse are reflected from the top of the canopy, others from understorey vegetation, and yet others from the ground surface. Recording the time of arrival of the first return pulse and the last return pulse would, in this case, provide information on the height of the vegetation canopy (Figure 8.6). More sophisticated systems are capable of identifying more than two discrete return pulses, or even of digitising the time structure of the whole return signal (i.e. without the need to identify discrete pulses within it) (Mallet and Bretar 2009).

8.1.3

Atmospheric correction of laser profiler data

As we have already noted, the propagation speed of the pulses is given by the group velocity, and this is different from (slightly less than) the speed of light c in *vacuo*. Under normal atmospheric conditions the group velocity differs from c by at most 0.03%, so for an observation made over a range H of 1000 m the error incurred by assuming $v_g = c$ will be at most 0.3 m. For very precise measurements, the correct value of v_g should be used in

Equation (8.1). This depends on the wavelength, the atmospheric pressure, temperature and water vapour content.

For spaceborne or high altitude airborne measurements, the atmospheric properties (pressure, temperature and water vapour content) are not constant along the path of the laser and it is necessary to integrate to find the travel time. Specifically, the *one-way* travel time for a path of length z is given by

$$T_t = \int_0^z v_g^{-1} dz'$$

where v_g is the group velocity as a function of the distance z' along the path. It is often convenient to use instead the quantity P , defined by

$$P = \int_0^z \left(\frac{c}{v_g} - 1 \right) dz' \quad (8.8)$$

so that

$$P = cT_t - z. \quad (8.9)$$

The quantity P thus has the dimensions of a length, and is in fact equal to the one-way range error that would be incurred if we assumed that the pulse had travelled at the speed c rather than at v_g .

The quantity P is useful because, at a given wavelength, it is proportional to the integral of the number density of molecules along the path. For practical purposes we can assume that the atmosphere consists of two components: the *dry atmosphere* (mainly nitrogen, oxygen and carbon dioxide) and the *water vapour* component. To determine the integrated number density of molecules in the dry atmosphere for a vertical path, all we need to know is the atmospheric pressure difference (specifically the difference in the partial pressures of dry air) at the top and bottom of the path. Figure 8.7 shows the value of P as a function of wavelength when this pressure difference is equal to the standard atmospheric pressure of 101 325 Pa, i.e. when the path traverses the entire atmosphere. For a path that does not traverse the whole atmosphere, or that makes an angle θ to the vertical, the values shown in Figure 8.7 should be multiplied by $\Delta p \sec \theta$ where Δp is the pressure difference between the two ends of the path, expressed in atmospheres. This expression is only valid for values of θ up to about 75° , as noted in Section 4.2.

For the water vapour component, the integrated number density of molecules for a vertical path is usually expressed as the thickness of the layer of water that would result if the water vapour were precipitated (condensed). Figure 8.8 shows the value of P as a function of wavelength for a vertical path through one metre of precipitable water.

In Table 4.1 we noted that the total mass of water vapour in the atmosphere varies typically between 6.5 and 180 kg m⁻². This corresponds to a range of 6.5 to 180 mm of precipitable water, so we see that a typical value of the correction due to water vapour will be of the order of 0.05 m.

8.1 Laser profiling

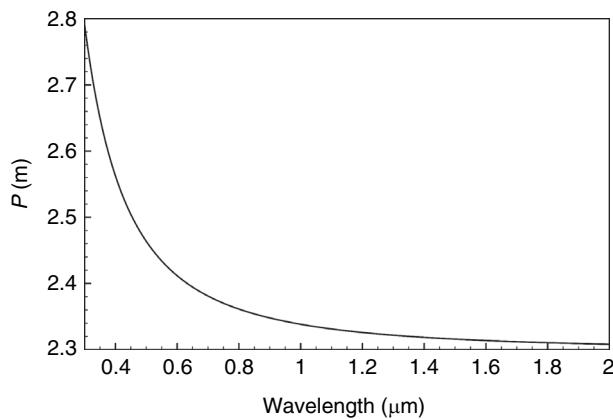


Figure 8.7. Dry-atmosphere propagation delay for one standard atmosphere.

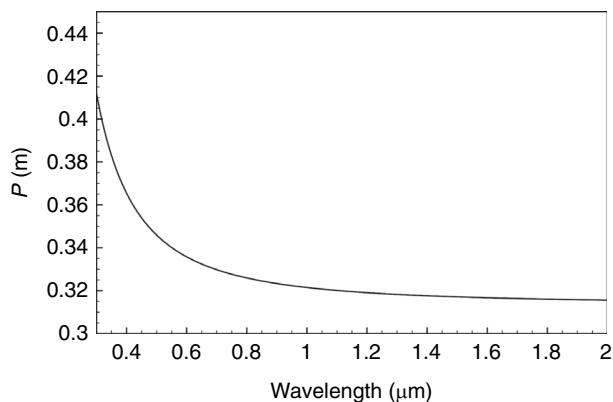


Figure 8.8. Propagation delay due to water vapour for one metre of precipitable water.

8.1.4 Applications of laser profiling

The first laser profiler to be operated from space was the Balkan-1 instrument, carried on the Mir space station ($H \approx 350$ km) from 1995. This operated at a wavelength of $0.532 \mu\text{m}$ and at a PRF of 0.18 s^{-1} , giving it an along-track sampling interval of about 45 km. The footprint width was 150 m and the range resolution about 4 m. It was mainly used for measuring cloud-top altitudes. The first continuously operating spaceborne laser profiler was the GLAS (Geoscience Laser Altimeter System), carried on the ICESat (Ice, Cloud and land Elevation Satellite). This satellite had an orbital altitude of about 590 km and was operational from 2003 to 2010. GLAS operated at wavelengths of $0.532 \mu\text{m}$ and $1.064 \mu\text{m}$, emitting short (4 ns) pulses that were detected using a 1-metre diameter telescope. The PRF of 40 s^{-1} gave an along-track sampling interval of 170 m and the beamwidth of 0.1 milliradian gave a footprint width of 70 m. In fact, the GLAS instrument consisted of three lasers because of the expected (and realised) difficulties of operating high power lasers in space. It did not

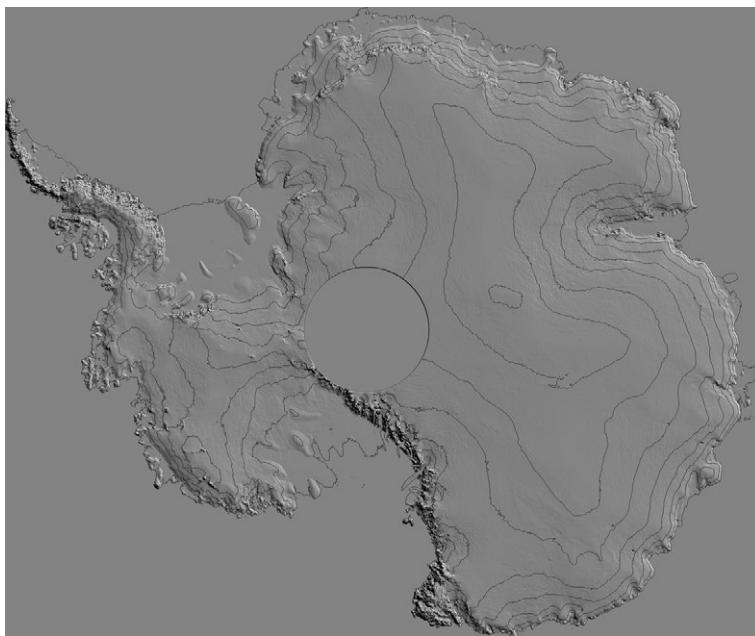


Figure 8.9. Shaded-relief topography of Antarctica constructed from GLAS data. The superimposed contours have an interval of 500 m. The hole centred on the South Pole is caused by the fact that the orbit of ICESat reached a maximum south latitude of 86°. (Source of data: Zwally, H.J., Schutz, R., Bentley, C., Bufton, J., Herring, T., Minster, J., Spinhirne, J. and Thomas, R. 2003, updated 2011. *GLAS/ICESat LIB Global Elevation Data V018*, 15 October to 18 November 2003. Boulder, Colorado: National Snow and Ice Data Center. Digital media.)

prove possible to maintain continuous operation of the lasers even so. Nevertheless, GLAS produced a huge quantity of data.

GLAS data were collected in order to measure land and sea surface topography, cloud-top altitudes, vegetation canopy structure, snow cover, and some atmospheric profiling tasks. However, the principal application of GLAS data was to measuring the surface topography and change of the large polar ice sheets (Figure 8.9).

Airborne laser profiling has found many applications in topographic mapping (Figure 8.10), and its high range accuracy (of the order of 0.1 m) has made it suitable for urban mapping (Balsavias and Gruen 2003) (Figure 8.11) and surveying applications. Waveform-measuring instruments, or those that simply record first and last return pulse, have found applications to the study of vegetation canopies (Naesset 1997) and to measuring the topography of the terrain beneath the canopy. This has been valuable for archaeological surveying (Chase *et al.* 2011). Profiling has also proved particularly useful in reconnaissance of sea ice, where a knowledge of the freeboard (the height of the ice surface above the water level) allows the extent of the submerged portion to be estimated, and in terrestrial glaciology. Repeat acquisitions of LiDAR data allow changes in topography to be measured, for example over a glacier (e.g. Figure 8.12) or a volcano (Höfle and Rutzinger 2011), and to measure the thickness of snow cover.

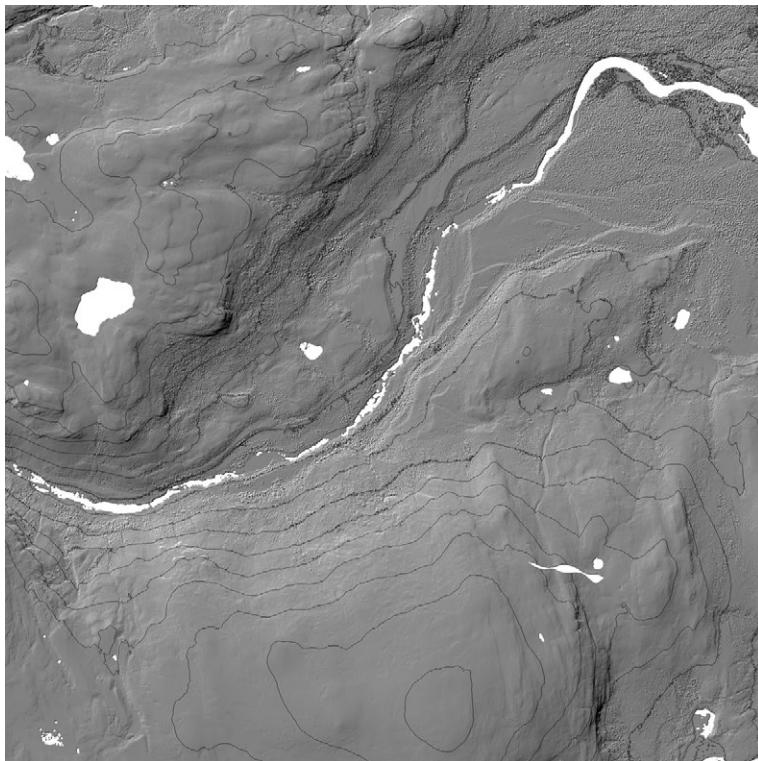


Figure 8.10. Shaded-relief DEM constructed from airborne LiDAR data, showing terrain near Lakselv, Norway. The image covers an area approximately 4.5 km square. Contours have been superimposed with an interval of 50 m. White areas represent regions from which no data were collected (these are water surfaces, which reflect very little infrared radiation).

Summary

A laser profiler transmits short, intense pulses of VNIR radiation and detects the echoes reflected from the Earth's surface. Measurement of the time delay of an echo allows the range to the reflecting point to be determined, and if the location of the platform carrying the instrument, and the direction in which the pulse was transmitted are both known, the location of the reflecting point can be determined. In this way, the Earth's surface topography is measured. Airborne laser profilers are usually scanning instruments, using a similar approach to whiskbroom scanning of VNIR imagers, though the few spaceborne laser profilers that have been deployed to date have produced linear transects along the sub-satellite path.

The simplest laser profilers measure a single echo delay, but more sophisticated instruments can identify a number of echoes or even record the entire time-structure of the return signal (the 'waveform'). In the case of a vegetation canopy, where the

energy of the transmitted pulse can be reflected at different levels of the canopy and also from the ground, this time-resolution of the return signal can allow the height of the canopy, the ground surface topography, and also the structure of the canopy, to be determined.

The effects of atmospheric propagation on the propagation speed of the pulses used in laser profiling are small but not negligible. For a vertically propagating pulse from a spaceborne system the atmosphere increases the apparent range by around 2 m; for an airborne system at an altitude of 2000 m the effect is about one-third as large.

Laser profiling from space currently achieves an along-track sampling interval of around 200 m, with a horizontal spatial resolution of around 40 m and a vertical resolution of a few decimetres. It has been used to measure cloud-top altitudes and (especially) ice-sheet topography. Airborne laser profiling typically provides a horizontal sampling interval of the order of 1 m. The vertical resolution is typically a few decimetres. It has found numerous applications to land-surface topographic mapping, surveying and mapping of man-made structures including cities, and measurement of surface topography and change of land and sea ice. It has proved especially useful for study of the structure of vegetation canopies and in identifying surface features beneath them.

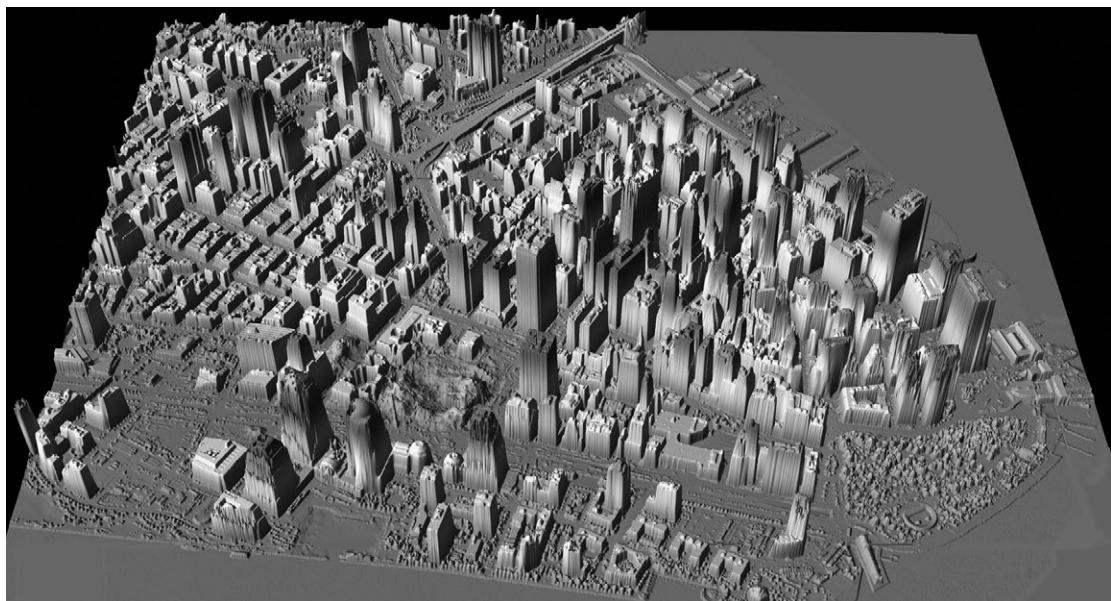


Figure 8.11. Visualisation of airborne LiDAR data collected over Lower Manhattan on 27 September 2001. (The data were acquired by NOAA and processed by the US Army Joint Precision Strike Demonstration. Image downloaded from <http://www.noaanews.noaa.gov/stories/s798.htm> and reproduced by courtesy of NOAA/US Army JPSD.) See also colour plates section.

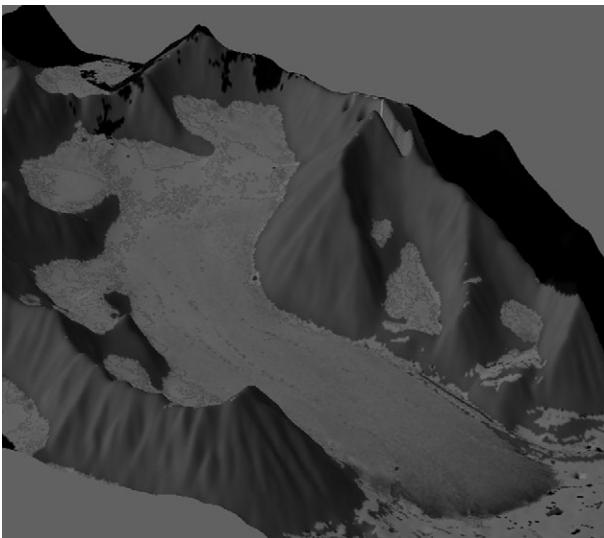


Figure 8.12. Visualisation of the change in surface altitude of a glacier in Svalbard between 2003 and 2005, determined by repeat acquisitions of airborne LiDAR data. Darker shades denote greater loss of altitude.

8.2

Radar altimetry

The radar altimeter is similar in operation to the laser profiler. The basic principle, that of timing a short pulse over its round trip from the instrument to the surface and back again, is the same. Most of the differences between the kinds of information obtainable from the two instruments can be ascribed to the larger beam width of the radar altimeter, which results from the fact that it operates at a much longer wavelength.

Figure 8.13 shows, very schematically, the construction of a radar altimeter. A pulse generator produces extremely short pulses at a frequency of, typically, around 10 GHz. These are fed to an antenna which radiates pulses of microwave electromagnetic radiation towards the Earth's surface. The same antenna collects the reflected pulses, and the signal is fed to a detector for subsequent analysis. The primary variable of interest is the time delay between the transmitted and received pulses but, as we have already noted for the laser profiler, the shape of the received pulse also contains useful information.

8.2.1

Simple model of the waveform

We first develop a very simple model of the operation of a radar altimeter, to illustrate the main features. In this model, which is based on the Brown model (Brown 1977), we assume that the Earth's surface is flat, and that it consists of a uniform density of isotropic, incoherent, point-like scatterers. We also neglect the operation of the inverse square law, which we can justify if the ranges of all the scatterers that make a significant contribution to the received signal do not differ very much, and the fall-off in sensitivity away from the axis of the antenna's main lobe. With these assumptions, the power that

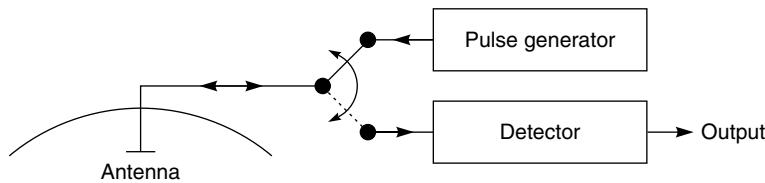


Figure 8.13. Operation of a radar altimeter (schematic).

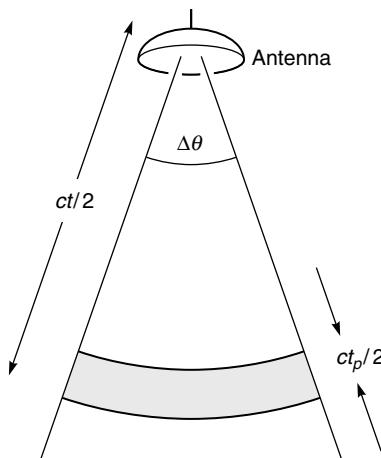


Figure 8.14. A radar altimeter emits a pulse of duration t_p beginning at time $t=0$. $\Delta\theta$ is the beam width of the antenna. Any scatterer within the scattering zone (shaded) will contribute to the signal received at time t .

would be received if the antenna were to transmit *continuously* would just be proportional to the area of the Earth's surface illuminated by it. Of course, the antenna does *not* transmit continuously (it is pulsed), and we therefore need to consider the time-structure of the received pulse.

It is clear that the return signal, if any, received at a time t after the emission of a pulse must arise from those scatterers situated at a distance $ct/2$ from the altimeter (for simplicity, we are assuming that the pulses travel at the speed of light c rather than at the group velocity). It will be convenient to describe this in terms of a 'scattering zone' that propagates away from the altimeter at a speed of $c/2$, such that any scatterer within the scattering zone contributes to the received signal at that time. This is shown schematically in Figure 8.14.

If the distance from the altimeter to the surface is H , it is clear that no return signal will be received until

$$t = t_0 = \frac{2H}{c}. \quad (8.10)$$

A short time Δt after this, the intersection of the scattering zone with the surface will be a circular disc of radius r , as shown in Figure 8.15. Provided that $r \ll H$, this radius is given by

$$r \approx \sqrt{cH\Delta t},$$

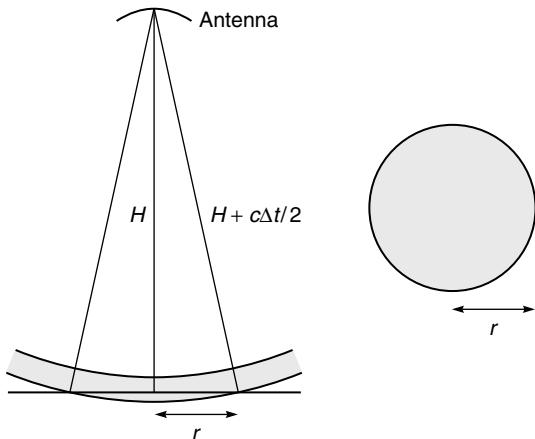


Figure 8.15. At a time $\Delta t < t_p$ after the first return signal is received, the scattering zone (shaded) intersects the surface in a disc of radius r . Left: elevation; right: plan view.

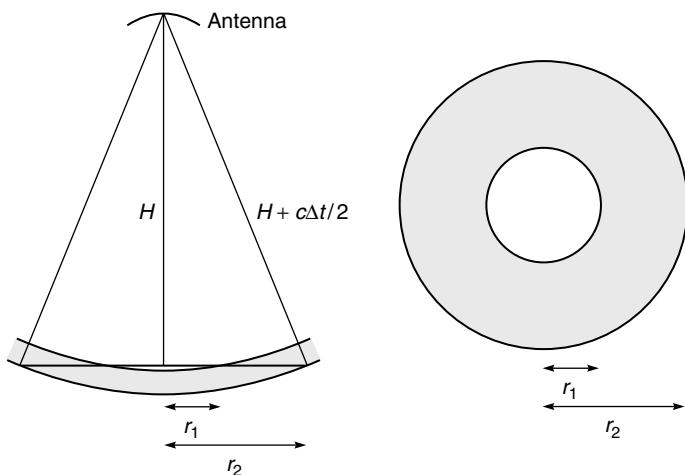


Figure 8.16. At a time $\Delta t > t_p$ after the first return signal is received, the scattering zone intersects the surface in an annulus. Left: elevation; right: plan view.

so the area of the disc is $\pi cH\Delta t$. According to our simplifying assumptions, the received power is proportional to this area and hence to Δt . Thus the received power will at first increase linearly with time. However, at time $t = t_0 + t_p$, where t_p is the duration of the pulse, the trailing edge of the scattering zone will just reach the surface. At times later than this the scattering zone will intersect the surface in an annulus, as shown in Figure 8.16.



The inner radius of this annulus is

$$r_1 \approx \sqrt{cH(\Delta t - t_p)}$$

and the outer radius is

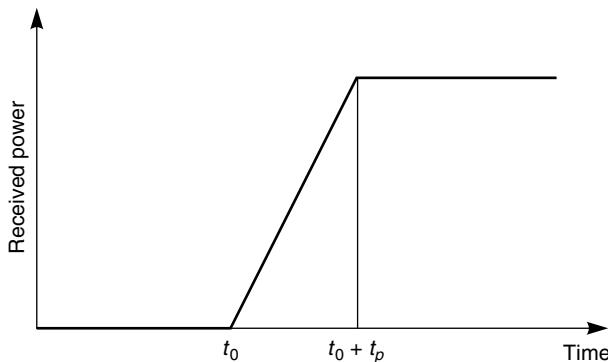


Figure 8.17. Time-dependence of the power received by a radar altimeter above a flat surface, according to the simple model derived in the text.

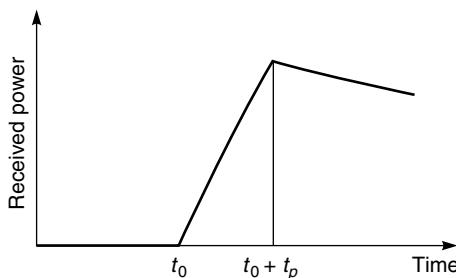


Figure 8.18. Modification of Figure 8.17 to take account of the antenna power pattern and the increasing distance of scatterers from the antenna.

$$r_2 \approx \sqrt{cH\Delta t},$$

so the area of the annulus is $\pi c H t_p$. This is independent of the time, so we can now see that the prediction of our simplified model is that the received power will increase linearly from zero at $t = t_0$ until $t = t_0 + t_p$, whereafter it will remain constant. This is shown in Figure 8.17.

Inspection of Figure 8.17 shows that the received power contains useful information only during the period $t_0 \leq t \leq t_0 + t_p$. It is clear that no further information is obtained at later times. This means that, in effect, the footprint of the instrument is a disc of radius r_p , where

$$r_p = \sqrt{cHt_p} \quad (8.11)$$

is the radius of the scattering disc at time $t_0 + t_p$.

In deriving this model, we have assumed that the beam width of the antenna is sufficiently large that variations in its response at increasingly large angles from the beam axis can be neglected. In general this will not be true, and the effect of the declining power pattern of the antenna (and also of the increasing distance of the scatterers from the antenna) as the time increases beyond t_0 will be to cause the received power to be less than predicted by our simple model, by a factor that increases with time. Thus Figure 8.17 should in fact be modified to look something like Figure 8.18.

We can, however, distinguish two limiting cases. The first of these is where the reduction in power as a result of the beam power pattern is negligible. The condition for this to be true is that

$$H\Delta\theta \gg 2r_p,$$

in which case the altimeter is said to be *pulse-limited* and Equation (8.11) correctly describes its spatial resolution. However, if

$$H\Delta\theta \ll 2r_p,$$

the altimeter is *beam-limited* and the spatial resolution is just $H\Delta\theta$. Most radar altimeters are pulse-limited, while laser profilers are beam-limited. As an example, we can consider a radar altimeter with a pulse length of 3 ns operating from a height of 800 km. Equation (8.11) shows that the radius of the beam-limited footprint will be approximately 850 m. From Section 2.7 we know that the angular beamwidth (in radians) of an antenna of diameter D at wavelength λ is approximately $1.22 \lambda/D$, so if $\lambda = 3$ cm and $D = 1$ m the radius of the beam-limited footprint will be about 30 km. In this case, clearly, the effect of the beam power pattern can be ignored and the altimeter is pulse-limited.

8.2.2 Effect of the Earth's curvature

Our model of the operation of a beam-limited altimeter involved a number of simplifications that we should now examine. One obviously incorrect assumption is that the Earth's surface is flat, and while this is usually adequate for airborne radar altimeters it is not self-evident that it is valid for a spaceborne instrument. Fortunately, the effect of the Earth's curvature can be dealt with rather simply. Figure 8.19 shows how Figure 8.15 can be modified to take the Earth's curvature into account.

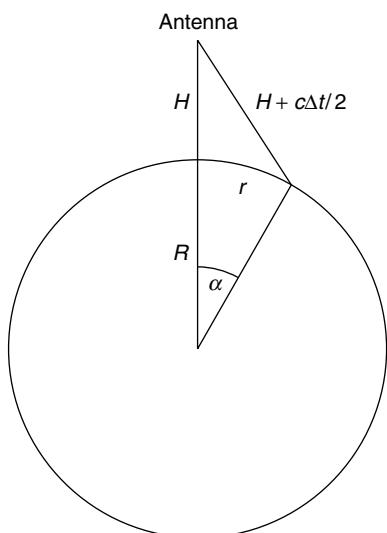


Figure 8.19. Geometry of a radar altimeter measurement when the Earth's curvature is taken into account.

The Earth is assumed to be a sphere of radius R , in which case simple trigonometry shows that

$$(H + c\Delta t/2)^2 = (H + R)^2 + R^2 - 2R(H + R)\cos\alpha.$$

Assuming that $\alpha \ll 1$, this can be simplified to give

$$\alpha^2 \approx (R^2/H + R)^{-1}c\Delta t,$$

and since $r = R\alpha$, we must have

$$r^2 \approx (H^{-1} + R^{-1})^{-1}c\Delta t.$$

Thus the radius of the pulse-limited footprint can still be calculated using Equation (8.11), provided that the range H is replaced by the effective height

$$H_{\text{eff}} = (H^{-1} + R^{-1})^{-1}. \quad (8.12)$$

Using the same example as before ($H = 800$ km, $t_p = 3$ ns), we see that the effective height is about 711 km so the value of r_p is reduced from 850 m to about 800 m.

8.2.3 Effect of coherence: range accuracy

Our model of the response of a radar altimeter is still rather crude. We can improve on the assumption that the surface consists of a uniform distribution of isotropic scatterers by incorporating the form of the BRDF, if it is known. However, there is a more important respect in which the model fails. We have assumed that the power received by the altimeter is proportional to the number of scatterers (and hence the area) visible to it, and have thus added together the powers scattered by the various scatterers. If the radiation reaching the antenna from two scatterers is *coherent*, that is, if there is a definite phase relationship between the two waves, the signals are capable of *interfering* with one another. They should thus be added as vector (or more precisely phasor) quantities, with due regard to amplitude and phase.

Two points will be coherently illuminated with respect to each other if the difference between their distances from the source of illumination (the radar altimeter) is less than the *coherence length* l_c of the radiation and their separation measured in a direction perpendicular to the propagation direction of the radiation is less than the *coherence width* w_c of the radiation. These quantities are given by

$$l_c \approx \frac{c}{\Delta f} \quad (8.13)$$

and

$$w_c \approx \frac{cH}{Df} \quad (8.14)$$

where c is the speed of light, f is the frequency of the radiation, Δf its bandwidth (which must be at least $1/t_p$ (see Section 2.3) and is usually equal to it), and D is the diameter of the antenna. As before, H is the distance from the antenna to the surface. A typical spaceborne radar altimeter will have $l_c \approx 1$ m and $w_c \approx 10$ km, so in practice most of the

scattering zone will be coherently illuminated. The consequence of this is that the power received from a flat surface will not have the simple form shown in Figure 8.17, but will instead be very noisy, in the manner of Figure 8.3 but more so, with a signal-to-noise ratio of the order of 1. However, by averaging together many pulses, something resembling Figure 8.17 can be obtained. An important consequence of the fact that the signal-to-noise ratio for a single pulse is ~ 1 is that the range accuracy of a single pulse is approximately

$$\Delta \approx \frac{ct_p}{2}, \quad (8.15)$$

consistent with Equations (8.1) and (8.2).

8.2.4 Response from a rough surface

The last of the simplifying assumptions that we examine is that the surface is flat. Let us suppose instead that the surface is rough, for example an ocean surface with waves. Considering Figure 8.15 again, we can see that the first return signal will now be received earlier than before, when the scattering zone just touches the tops of the waves. Similarly, the time taken for the received pulse to rise to its maximum value will be increased, because this will now correspond to the time taken for the trailing edge of the scattering zone to reach the lowest scatterers (the lowest troughs of the waves). This is illustrated in Figure 8.20.

The time taken for the received signal to rise from zero to its maximum value will thus be increased beyond the value of t_p (the duration of the transmitted pulse) that we derived in Section 8.2.1. The shape of the time variation (the waveform) of the received signal will also be altered, depending on the distribution of the surface scatterers in height. In general, it can be shown that (at the same degree of approximation that we used to derive Figure 8.17)

$$\frac{dP_r}{dt} \propto \int_{-\infty}^{\infty} P_t(t + 2(h - H)/c)f(h)dh \quad (8.16)$$

where $P_r(t)$ and $P_t(t)$ are respectively the received and transmitted powers at time t ,

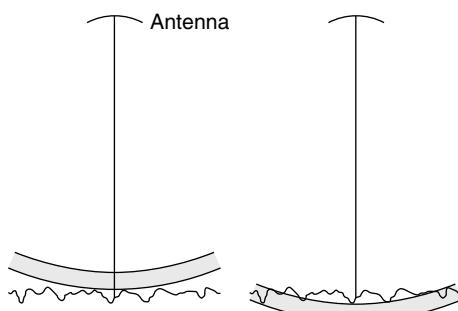


Figure 8.20. The scattering zone of a radar altimeter encounters a rough surface. Left: the instant when the first return signal is received; right: the instant when the received power reaches its maximum value.

and $f(h)dh$ is the proportion of scatterers between heights h and $h+dh$ above the mean height of the surface.

The right-hand side of Equation (8.16) is in fact a *convolution* of the surface height distribution and the shape of the transmitted pulse. Approximately, we may state that if the surface height is distributed over a range Δh , the time t'_p taken for the received power to increase from near zero to near maximum will be given by

$$t'^2_p \approx t_p^2 + k \frac{\Delta h^2}{c^2}, \quad (8.17)$$

where k is a dimensionless constant that depends on how Δh , ‘near zero’ and ‘near maximum’ are defined. Of course, Equation (8.16) allows a more quantitative statement to be made if the shapes of the transmitted pulse and of $f(h)$ are known. As an example, Figure 8.21 shows the waveform for a rectangular transmitted pulse (i.e. one in which the pulse is switched on abruptly, remains constant for time t_p , then is switched off again) incident on a surface for which $f(h)$ is a Gaussian distribution.

An important consequence of the broadening of the pulse received from a rough surface, from t_p to t'_p , is that the radius r_p of the pulse-limited footprint is no longer given by Equation (8.11). Instead, the effective resolution is coarsened to

$$r'_p = \sqrt{cHt'_p}. \quad (8.18)$$

In the case of Figure 8.21, for example, the rise time of the received pulse is of the order of 8 ns, so if $H_{\text{eff}} = 711$ km the radius of the pulse-limited footprint has been broadened from 800 m to about 1300 m. Pulse-broadening also implies a coarsening of the range resolution, which can be seen by substituting t'_p for t_p in Equation (8.15).

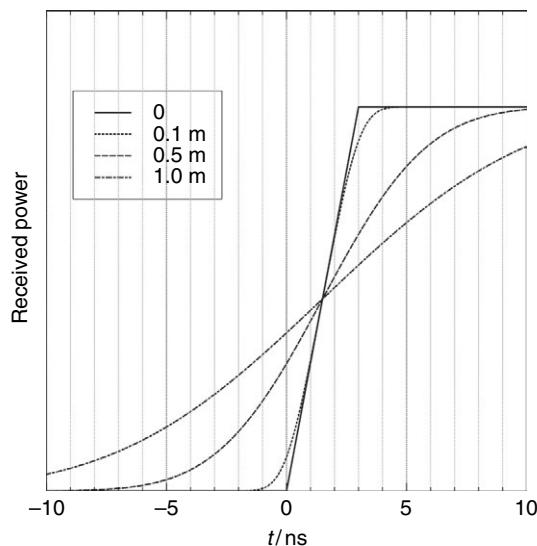


Figure 8.21. Theoretical waveform of the received power for a rectangular pulse of duration 3.0 ns incident on a flat surface and surfaces with a Gaussian distributions of heights with standard deviations of 0.10, 0.50 and 1.0 m.

8.2.5

Applications of radar altimetry

Radar altimeter measurements are used extensively for characterising the topography of the ocean surface (Wunsch and Stammer 1998). The long-term average of the ocean surface (i.e. the average after the effects of tides, surface gravity waves, atmospheric pressure variations and wind-driven disturbances have been removed) is, with a proviso we shall discuss later, coincident with the *geoid*. This is a surface of constant gravitational potential (an equipotential surface) that is a close approximation to an *ellipsoid* of revolution (sometimes called a *spheroid*) with its shortest axis along the Earth's polar axis and circular symmetry about this axis. The geoid differs from this ellipsoid by distances of the order of 100 m, these variations being due to variations in the density of the Earth's mantle and lithosphere and to variations in the topography of the Earth's solid surface. Determination of the geoid is thus important for understanding the structure of the Earth and its gravity field, and also for making accurate predictions of satellite orbits (see Chapter 10).

As we have already mentioned, the derivation of the mean sea surface requires the averaging of repeated measurements to reduce the magnitude of time-dependent effects. Where these phenomena vary more or less randomly in time, it is merely necessary to take the average over a time that is much longer than the correlation time of the phenomenon. However, for phenomena that are strongly periodic, such as tides, a potential problem is introduced by the fact that the sampling is also periodic, at least for a spaceborne system. It is clear that if the sampling frequency exactly matches the frequency of the tide, the tidal variation will be sampled at the same point in its cycle for each measurement, and consequently no information at all will be obtained about the temporal variability of the tide. This is another example of the phenomenon of *aliasing*, which will be discussed in greater detail in Section 10.3.2.4.

Figure 8.22 is a visualisation of the geoid (Pavlis *et al.* 2008) determined principally from data collected from the GRACE satellite mission (i.e. not using radar altimeter data). It clearly shows topographic variations at a wide range of scales, but perhaps the most strikingly obvious features are those that correspond to the deep ocean-floor trenches, for example the South Sandwich Trench to the east of Cape Horn. The surface features corresponding to these trenches are typically 10 m deep and 200 km wide (Figure 8.23), so that surface slopes are a few tenths of a minute of arc. They reflect the fact that the gravitational field strength is reduced above a trench, so that the equipotential surface moves closer to the Earth's centre in this region. Mapping of ocean floor gravitational anomalies in this way is valuable for studies of plate tectonics, locating sedimentary basins (e.g. for oil exploration) etc.

The one situation in which long-term averaging of the sea surface topography does not result in a surface that corresponds to the geoid is where the surface has a steady motion as a result of an ocean current (Fu 2010). In the northern hemisphere, the surface of a stream of water moving at a constant velocity v relative to the Earth's surface will be tilted such that the right-hand edge of the stream is higher than the left (in the southern hemisphere the left side is higher than the right), the angle of tilt being given by

$$\frac{2\Omega v \sin \phi}{g}$$

$$g$$

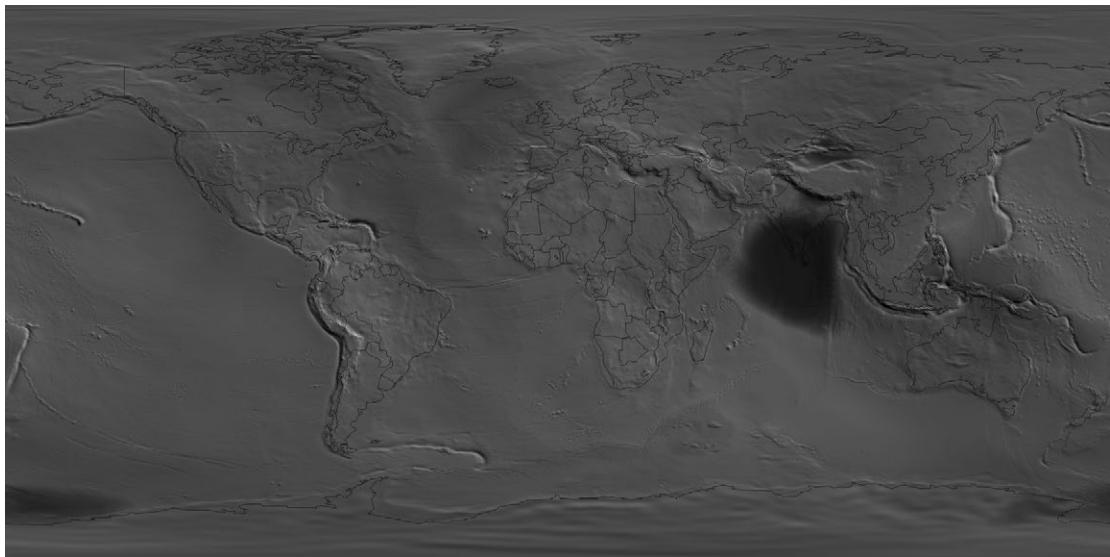


Figure 8.22. Visualisation of the global geoid EGM2008. The colour scale shows the height of the geoid above the WGS84 ellipsoid, ranging from -106 m (blue) to $+86\text{ m}$ (orange).
 (Source: <http://earth-info.nga.mil/GandG/wgs84/gravitymod/egm2008/oceano.html>).

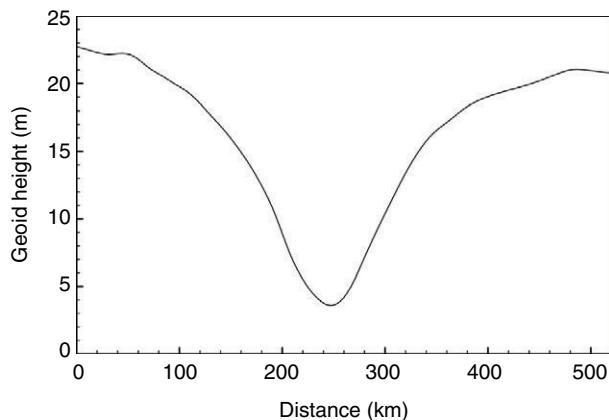


Figure 8.23. Transect of the geoid through the South Sandwich Trench, from 58.0° S 27.5° W to 54.25° S 22.5° W .

where Ω is the Earth's angular velocity, φ is the latitude and g the gravitational field strength. This phenomenon, called *geostrophic balance*, is a consequence of the Coriolis force, and the slopes are very small. For example, the Gulf Stream has a typical velocity of 2 m s^{-1} , so at a latitude of 45° N the slope is 2×10^{-5} radians, but over a typical stream

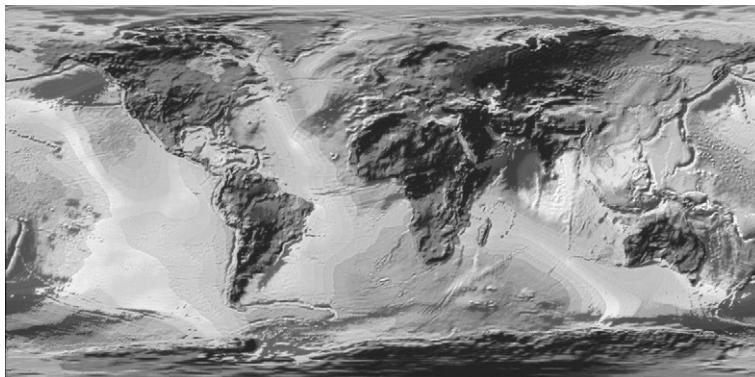


Figure 8.24. Visualisation of the global mean ocean surface, using data from the ERS-1 radar altimeter. (© ESA 1994)

width of 100 km this accounts for a difference in height of 2 m across the stream. This surface tilt can be discerned in radar altimeter data in Figure 8.24.

The sensitivity of the waveform (the shape of the received pulse) to surface roughness, discussed in Section 8.2.4, means that radar altimeter measurements can be used to determine the sea state. The simplest and most widely reported measure of sea state is the *significant wave height* $H_{1/3}$, defined as the mean height (from trough to crest) of the highest third of the waves. It is approximately related to the variance σ^2 of the surface height distribution by

$$H_{1/3} \approx 4\sigma.$$

The significant wave height is related to the wind speed (Young, Zieger and Babanin 2011), and can be used to determine it, although it is also dependent on the *fetch* (the distance from land, measured in the direction in which the wind is blowing) and the *duration* (the time for which the wind has been blowing). For sufficiently large fetch and duration, the sea is said to be *fully developed*, in which case the significant wave height depends only on the wind speed, and is given by

$$\frac{H_{1/3}}{\text{m}} \approx 0.02 \left(\frac{v_w}{\text{ms}^{-1}} \right)^2$$

where v_w is the wind speed 10 m above the surface. On this basis, wind speeds over oceans can be determined to an accuracy of typically $\pm 2 \text{ ms}^{-1}$.

In fact, determination of the significant wave height is also important for measuring the sea surface topography. The reason for this is that the height distribution $f(h)$ of scatterers, introduced in Section 8.2.4, is not symmetric about the mean surface height. The consequence of this is that the mean surface height is underestimated by an amount that is typically 2 to 3% of the significant wave height. This effect is known as *electromagnetic bias* or *sea-state bias* (Hausman and Zlotnicki 2010).

Radar altimeter measurements are also important for monitoring change in the global mean sea level (MSL) (Nerem *et al.* 2010). Figure 8.25 shows the MSL calculated at 10-day intervals from 1992 to 2011.

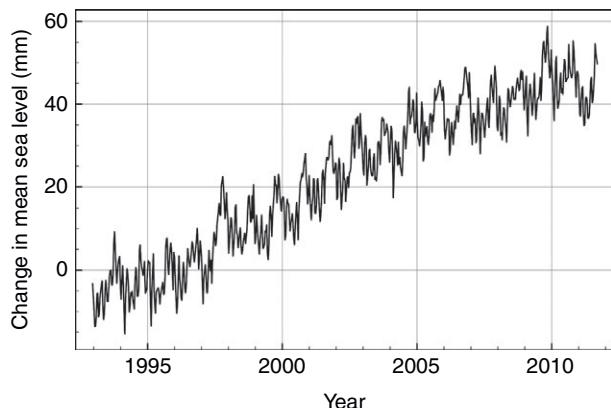


Figure 8.25. Variation over time in global mean sea level, calculated using radar altimeter data from TOPEX (to 2002), Jason-1 (2002–8) and Jason-2 (2008–11) radar altimeter data. (Data processed by Colorado University Sea Level Research Group and downloaded from <http://sealevel.colorado.edu/content/2011rel4-global-mean-sea-level-time-series-seasonal-signals-retained>.)

Topographic measurements over land surfaces are considerably more difficult to make using a spaceborne radar altimeter. There are essentially two technological reasons for this, both related to the fact that land surfaces exhibit considerably larger slopes than ocean surfaces. The first of these relates to a phenomenon usually referred to as ‘loss of lock’ or ‘loss of tracking’, and it can be explained roughly as follows. If we have a spaceborne radar altimeter at an altitude of 800 km, the time interval between the emission and reception of a pulse is of the order of 5 million nanoseconds. If the instrument is to be capable of resolving the waveform of the returned pulse, it will have to sample the pulse at intervals of the order of 1 nanosecond or even less. In order to keep the volume of data collected by the instrument within manageable limits, the receiving and detecting part of the instrument is therefore only activated a short time before the expected return of the pulse. This clearly requires an accurate prediction of the arrival time of the next pulse, based on the last pulse or last few pulses received. Over very smoothly varying surfaces with small slopes this is not much of a problem. However, land surface slopes can often be so steep that the next pulse will arrive well outside the time during which the instrument is ready to accept it.

The second problem that can arise in radar altimetry of comparatively steep slopes is that of *slope-induced error*. In effect the return pulse is derived, not from the nadir point (the point directly beneath the instrument), but from the closest point to the altimeter. This is illustrated in Figure 8.26.

When the radar altimeter is located at the position R_1 , the closest point of the surface is S_1 . If no correction is made for the slope-induced error, the scattering point will be assumed to be at the position S'_1 , where the distances R_1S_1 and $R_1S'_1$ are both equal to the range H and S'_1 is directly below R_1 . For a small slope α (radians), the horizontal error is approximately $H\alpha$ and the vertical error is approximately $H\alpha^2/2$. Thus, for a spaceborne altimeter ($H \approx 800$ km) observing the ocean surface, where slopes are unlikely to exceed 10^{-4} radians, the errors are at most 80 m horizontally and 4 mm vertically, and can safely

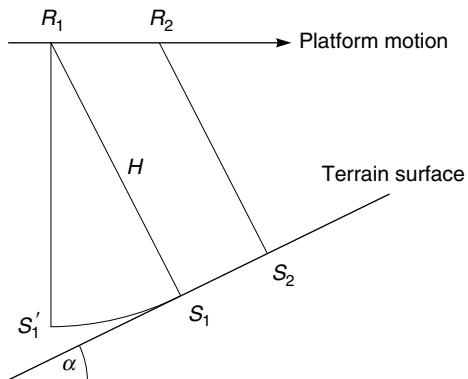


Figure 8.26. Illustration of slope-induced error in one dimension. When the radar altimeter is located at R_1 , scattering actually occurs from the point S_1 but is assumed to occur at the point S'_1 .

be ignored. However, over a land surface with a slope of, say, 0.03 radians, the errors are 24 km horizontally and 360 m vertically, and hence far from negligible.

Provided that the slope α does not change too rapidly, slope-induced error can be corrected. Figure 8.26 indicates how this can be achieved. If we have a second measurement from position R_2 , such that the true location of the scattering point is S_2 , the slope α can be deduced from the rate at which the range H changes with the along-track distance that the instrument has travelled. Once α is known, the scatterers can be assigned to their correct locations.

Two more remarks should be made about slope-induced error. The first is that the problem is in fact a two-dimensional one, so that correction requires two-dimensional coverage of the area rather than the one-dimensional transect illustrated in Figure 8.26. The second is that if the angle α is large compared with the beamwidth of the antenna, little or no signal will be received from the surface unless the antenna is tilted so that its beam axis is normal to the surface.

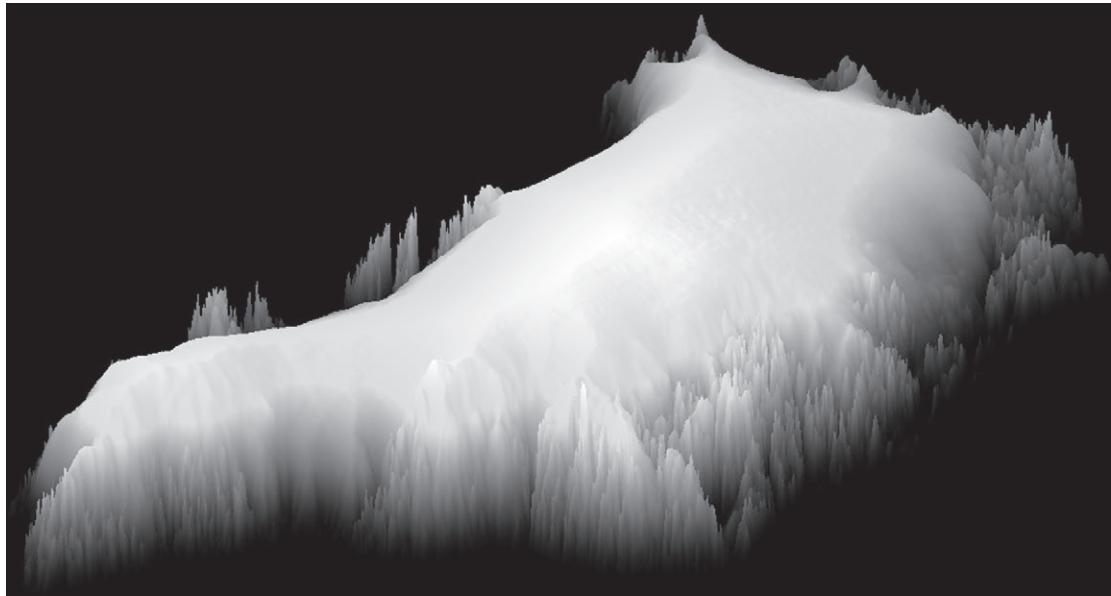
Despite these difficulties, spaceborne radar altimetry has been applied with some success to mapping land surface topography in comparatively flat areas, and with notable success to mapping the Antarctic and Greenland ice sheets where the surface slopes are generally a few degrees at most (Figure 8.27).

8.2.6

Atmospheric and ionospheric correction of radar altimeter data

For simplicity, we have been assuming throughout our discussion of radar altimetry that the pulses propagate at the speed of light c . However, as for the laser profiler, we should really use the appropriate group velocity. The principles of correcting the measured time delay for atmospheric (principally tropospheric) delay are very similar to those for the laser profiler, discussed in Section 8.1.3. There is, though, a major simplification in the case of the microwave frequencies employed for radar altimeters. The tropospheric propagation is practically non-dispersive, which means that the group velocity is equal to the phase velocity and independent of the frequency. The corresponding values of P (defined by Equation 8.8) are 2.33 metres per atmosphere for the dry tropospheric

(a)



(b)

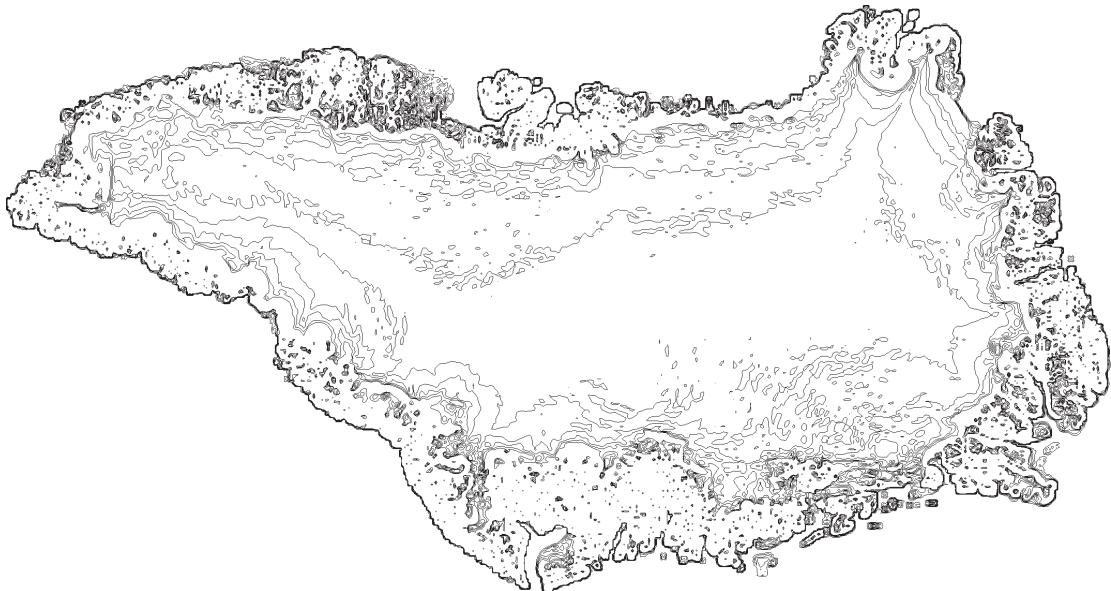


Figure 8.27. (a) Visualisation of the surface topography of Greenland using radar altimeter data; (b) contours of surface slope. Contours are plotted at intervals of 0.2 degrees from 0.2 to 1.0 degrees. (Data provided by the National Snow and Ice Data Center DAAC, University of Colorado, USA. Bamber, Layberry and Gogineni (2001).) See also colour plates section.

8.2 Radar altimetry

component, and 7.1 metres per metre of precipitable water. Thus, the tropospheric delay for a spaceborne observation is typically between 2.4 and 3.6 metres depending on the amount of water vapour present in the atmosphere. Correction for this delay therefore requires that both the atmospheric pressure and water vapour distributions be known.

For a spaceborne observation we must, however, also consider the effect of the ionosphere. This was discussed in Section 4.5. Rearranging Equation (4.21), we find that the quantity P is given by

$$P = \frac{e^2 N_t}{8\pi^2 \epsilon_0 m_e f^2}, \quad (8.19)$$

where e and m_e are respectively the charge and mass of an electron, N_t is the total electron count, and f is the frequency. For a typical daytime total electron count of $3 \times 10^{17} \text{ m}^{-2}$ and a frequency of 10 GHz, we find that $P = 0.12 \text{ m}$. If necessary (for high precision measurements), this effect can be corrected if the total electron content is known, or, more conveniently, by using a dual-frequency radar altimeter.

8.2.7 Example: the Envisat RA-2 radar altimeter

As an example of a typical spaceborne radar altimeter, we consider the RA-2 instrument that is included as part of the payload of the Envisat satellite, orbiting at an altitude of 780 km. Envisat was launched in 2002. RA-2 is a dual-frequency altimeter, operating at both K_u-band (13.575 GHz) and S-band (3.2 GHz), so with an antenna diameter of approximately 1 m it has beam-limited footprints of 26 km (K_u-band) and 90 km (S-band). In its highest-resolution mode of operation, the K_u-band subsystem generates pulses of length 3.1 ns, so the diameter of the pulse-limited footprint over a smooth surface is 1.7 km. The S-band pulse length is 6.3 ns, giving a pulse-limited footprint of 2.4 km. The instrument is thus pulse-limited at both frequencies.

The range resolution for a single measurement over a flat surface is approximately $c t_p / 2$, or about 0.5 m for K_u-band and 1 m for S-band. However, 100 K_u-band waveforms and 25 S-band waveforms are averaged in the instrument, improving these resolutions to approximately 5 cm and 20 cm respectively. The pulse repetition frequency for the K_u-band measurements is 1800 s^{-1} so each 100-waveform average is acquired in 56 milliseconds, during which time the satellite moves a distance of approximately 400 m relative to the Earth's surface. This, therefore, is the spatial sampling interval. For the S-band measurements the PRF is 450 s^{-1} , so the spatial sampling interval is again 400 m. The principal reason for having two operating frequencies is to allow for ionospheric corrections, which are clearly necessary if range accuracies as fine as 5 cm are to be achieved.

The K_u-band subsystem can also be operated in two lower resolution modes, with pulse lengths of 12.5 and 50 ns respectively. These give pulse-limited footprint diameters of 3.5 and 7 km (so they are still pulse-limited), and range resolutions, after averaging, of approximately 20 and 75 cm respectively. The purpose of these modes is to allow tracking of rougher terrain.

In passing, we can make two final remarks on the technical difficulties associated with achieving accuracies of the order of 10 cm in the range measured from a spaceborne radar altimeter. The first of these is a rather obvious one. If it is desired to measure the Earth's

surface topography with an absolute accuracy of 10 cm, it will be necessary to know the position of the satellite to this accuracy. Methods of obtaining this knowledge are discussed in Chapter 10.

The second point concerns the production of the extremely short (nanosecond) pulses needed to obtain high range precision. It would in fact be extremely difficult or impossible to put enough energy into a pulse only a few nanoseconds long, such that the pulse would be detectable after its 1600-km round trip to the Earth's surface. Instead, *pulse compression* is used. The real pulses are, in the case of the Envisat RA-2, 20 μs long, but they are modulated in frequency. We saw in Section 2.3 that a pulse of duration T must contain a range of frequencies $\Delta f \approx 1/T$. Pulse compression effectively uses the converse of this principle, which states that if we wish to construct a pulse of length t_p , we must use a range of frequencies of $\Delta f \approx 1/t_p$. The frequency modulation necessary to achieve a synthetic pulse length of 3.1 ns thus requires a bandwidth of approximately 320 MHz. In practice, this is achieved by generating a '*chirp*', which is a pulse whose frequency rises from $f_0 - \Delta f/2$ to $f_0 + \Delta f/2$ over a period of 20 μs (f_0 is the central frequency, i.e. 13.6 GHz for the K_u-band subsystem).

Summary

The radar altimeter, like the laser profiler, emits short pulses of radiation and uses the time delay of the echo to determine the range to the reflecting point. Because of the much longer wavelengths of microwave radiation than VNIR radiation, the beamwidth of the antenna used to transmit and receive the signal is much greater (typically one degree) than the beamwidth of a laser profiler (a few hundredths of a degree). Most radar altimeters are *pulse-limited*, meaning that the effect of the short pulse on the time-dependence of the echo signal is much more important than that of the antenna's beamwidth. In this case the form of the echo from a smooth surface is a steady increase in power over a time equal to the pulse duration t_p , after which the echo power remains constant. The effective horizontal spatial resolution (the *pulse-limited footprint*) of the system is the area contributing to the echo signal at the instant it reaches its maximum value. This is a disc of radius where c is the speed of light and H' is the effective height above the surface (somewhat smaller than the true height in the case of a spaceborne measurement because of the effect of the Earth's curvature). The vertical resolution of a single pulse is $ct_p/2$. Over a rough surface the rise time of the echo signal is increased, which increases the size of the pulse-limited footprint.

Radar altimetry is well suited to measurement of the ocean surface topography. This is governed by the geoid (the surface of constant gravitational potential, determined by the distribution of mass within the Earth) together with effects such as tides, waves, atmospheric pressure variations and wind-driven disturbances that average to zero over the long term, and steady ocean currents that do not average to zero. The effect of surface roughness on the rise-time of the echo allows radar altimeter data to be used to characterise significant wave height, and hence to infer wind speed, over the ocean. Radar altimeter measurements are also used to measure changes in global mean sea level. Radar altimetry is less well suited to measurements over land surfaces, because it is difficult to follow the rapidly changing range to the surface except over very smoothly varying surfaces.

However, it is well suited to topographic measurements over the large ice sheets of Antarctica and Greenland.

Atmospheric propagation has a significant effect on the range measured by a spaceborne radar altimeter. The atmosphere introduces a delay of typically 2–4 metres, depending on the water vapour content. The ionosphere introduces a delay of typically a decimetre. The ionospheric delay can be measured and corrected by a dual-frequency observation since it is dependent on frequency, but the atmospheric component must be corrected using ancillary data.

A typical spaceborne radar altimeter operates at a frequency of around 10 GHz and transmits pulses of a few nanoseconds duration, giving a pulse-limited footprint of the order of 1 km. The vertical resolution of a few decimetres is improved by averaging pulses, and range accuracies as fine as 5 cm are achievable.

8.3

Other ranging systems

We have now discussed the two main types of ranging system that conform to the definition of remote sensing enunciated in Chapter 1. We should also, however, briefly mention *radio echo-sounding*. This is a technique for measuring the thickness of ice sheets and glaciers, relying on the large attenuation length of VHF (*c.* 100 MHz) radio waves in ice. Since the attenuation length in ice at these frequencies is of the order of 100 to 1000 m (see Figure 3.1), it is feasible to transmit a signal through a large body of ice, and to detect the echo from the bedrock beneath it, even at a range of several thousand metres, which is typical of the Antarctic and Greenland ice sheets. This technique has been extensively and successfully employed (Bingham and Siegert 2007) for mapping ice sheets and glaciers, with a range resolution approaching one metre. However, because of the long wavelength (~ 3 m) in free space, narrow-beam antennas are not yet technologically feasible and such remote sensing has so far been confined to observations from platforms on or relatively close to the ice surface. Satellite observations will be precluded until narrow-beam instruments can be devised and placed in orbit.

Similar techniques are used for determining the thickness of saline ice, although in this case the higher electrical conductivity and inhomogeneous structure of the medium greatly reduce the attenuation length. For this reason, high power systems must be employed, and the distance from the platform to the ice surface must be kept small (Holt *et al.* 2009). Satellite-based remote sensing of sea-ice thickness is an even more distant prospect than for ice sheets and glaciers, whose ice is comparatively pure and homogeneous.

Finally, we mention the use of *soil-sounding radars* (or *ground-penetrating radars*) in archaeological and other investigations (Jol 2009). Again the technique is similar to radio echo-sounding, although the frequencies used are typical radar (microwave) frequencies rather than VHF. It has achieved a limited degree of success over dry soils in which buried artefacts produce a strong electromagnetic contrast.

Summary

Other important classes of instrument, not strictly conforming to the definition of remote sensing used in this book, employ the principle of measuring the time delay in the echo of a short pulse to determine the distance to a reflecting object. These instruments, which include ground-penetrating radar, impulse radar and radio echo-sounding, measure the time delay of a pulse of electromagnetic propagating *through* a solid medium. Radio echo-sounding, for example, relies on the transparency of cold freshwater ice at long radio wavelengths to measure the thickness of the ice to the bedrock beneath it.

Review questions

- Outline the principles of operation of a laser profiler, including an explanation of the factors that control its accuracy in measuring the range to a surface.
- Explain how two-dimensional scanning can be achieved by an airborne laser profiler.
- Explain the advantages of a waveform-resolving laser profiler compared to one that records a single variable (the pulse travel time) for each measured point.
- Describe the applications of airborne and spaceborne laser profiling.
- Outline the principles of operation of a radar altimeter, including an explanation of the factors that control its horizontal and vertical spatial resolution.
- Discuss the main applications of data from a spaceborne radar altimeter.
- Compare the performance of a radar altimeter with other spaceborne techniques for measuring surface topography.

Problems

1. A laser profiler is operated at a wavelength of $1 \mu\text{m}$ from an altitude of $10\,000 \text{ m}$, and views at 45° to the nadir. Estimate the range corrections needed to account for the dry atmosphere and the water vapour component if the atmosphere contains 50 mm of precipitable water. The atmospheric pressure at $10\,000 \text{ m}$ altitude can be taken as 0.26 atmospheres.
2. (Use small-angle approximations throughout this question.) An airborne scanning laser profiler transmits pulses at a frequency f_0 . Its field of view is scanned in a direction perpendicular to the aircraft's velocity through an angle $\pm\theta$ either side of the nadir direction, at a scanning frequency f_s . The aircraft's speed is v and its flying height is H over level ground. The spatial arrangement of samples on the surface is shown schematically in Figure 8.4. Show that the average along-track spacing s_a of the samples is $v/2f_s$ and that the average across-track spacing s_c is $4H\theta f_s/f_0$. The system has $f_0 = 33.3 \text{ kHz}$ and is flown on an aircraft at $v = 70 \text{ m s}^{-1}$. It is subject to the following constraints on its operation:
 $150 \text{ m} < H < 2000 \text{ m}$

$$\begin{aligned}\theta &< 0.35 \text{ radian} \\ f_s\theta &< 14 \text{ Hz} \\ f_s &< 50 \text{ Hz}\end{aligned}$$

It is required to operate the system so that $s_a < 1 \text{ m}$ and $s_c < 1 \text{ m}$, while at the same time maximising the swath width w .

- (i) Calculate the value of the scanning frequency f_s that satisfies these criteria, and the corresponding maximum swath width w .
- (ii) Calculate the minimum flying height consistent with your results in (i) and all the constraints. Suggest why it is desirable to minimise the flying height.
- 3. Explain how it is possible for the horizontal resolution of a radar altimeter to exceed the limit set by diffraction at the antenna, provided that the pulse duration t_p is short enough, and derive the result that the resolution is given by

$$2(cHt_p)^{1/2},$$

where c is the speed of light and H is the range to the surface. (Ignore any effects arising from the Earth's curvature.)

Hence show that the vertical resolution of a pulse-limited radar altimeter is given by

$$\left(\frac{c^3 t_p^3 v^2}{64 H p^2} \right)^{1/4},$$

where v is the speed of the platform relative to the Earth's surface and p is the pulse repetition frequency. Estimate the best vertical resolution achievable from a space-borne radar altimeter if $t_p > 3 \text{ ns}$ and $p < 2 \text{ kHz}$.

- 4. A radar altimeter emits a pulse whose variation of power with time is Gaussian with a length of 3.00 ns between $1/e$ points. The pulse is reflected from a surface whose height distribution is Gaussian with a range of 1.00 m between the $1/e$ points. Calculate the time for the reflected pulse to rise from 8% to 92% of its final value, neglecting coherence effects. [Note that 84% of the area under a Gaussian curve is enclosed between the $1/e$ points.]
- 5. (For enthusiasts). Prove Equation (8.16).
- 6. For microwave radiation propagating vertically through the Earth's atmosphere, the delay to a pulse as a result of the dry component of the atmosphere, expressed as an equivalent distance, is 2.33 metres irrespective of frequency. The corresponding figure for the water vapour component of the atmosphere is 7.10 metres per metre of precipitable water, again irrespective of frequency.
 - (i) Explain why the dry atmosphere delay can be expressed 'per atmosphere' while the water vapour delay is expressed per metre of precipitable water.
 - (ii) Show that a pulse of microwave radiation with a carrier frequency f will be delayed by its passage through the ionosphere by a time equivalent to a free-space distance of

$$\frac{e^2 N_t}{8\pi^2 \epsilon_0 m_e f^2},$$

where e and m_e are respectively the charge and mass of the electron, N_t is the total electron content (TEC) of the ionosphere and ϵ_0 is the permittivity of free space.

- (iii) A dual-frequency satellite radar altimeter measures the range to the Earth's surface at 3.2 GHz and 13.6 GHz. After averaging, the travel times of the pulses are found to be equivalent to free-space ranges of $793\,125.20 \pm 0.10$ m at 3.2 GHz and $793\,123.35 \pm 0.05$ m at 13.6 GHz. The atmospheric pressure at the Earth's surface is 0.970 atmospheres and the water vapour content is 0.15 m of precipitable water. Calculate the true range to the surface and its uncertainty. Also calculate the TEC and its uncertainty.

7. (For Fourier transform enthusiasts only.) A chirp signal, in which the angular frequency rises uniformly from $\omega_0 - \Delta\omega/2$ to $\omega_0 + \Delta\omega/2$ over a time T , can be written as $\exp(i\varphi(t))$, where the phase $\varphi(t)$ is given by

$$\phi(t) = \omega_0 t + \frac{\Delta\omega}{2T} t^2$$

for $|t| \leq T/2$. This signal is then passed through a delay-line, whose effect on a component of angular frequency w can be represented as multiplication by the factor

$$\exp\left(-i\omega(T/\Delta\omega)(\omega_0 + \omega/2)\right).$$

Show that the signal that emerges from the delay-line is a carrier of angular frequency ω_0 , modulated by an envelope of width $2\pi/\Delta\omega$.

9

Scattering systems

In this chapter we complete our survey of the principal types of remote sensing instrument by discussing those active systems that make direct use of the backscattered power. Optical (LiDAR) systems are used for sounding clouds, aerosols and other atmospheric constituents, for characterising surface albedo, and for measuring wind speeds. These are discussed briefly in Section 9.1. However, the bulk of this chapter is concerned with microwave (radar) systems. ('Radar' is an acronym, originally standing for 'radio detection and ranging'. The functions that can be performed by microwave scattering systems now extend far beyond detection and ranging, but the term continues to be in very general use.)

In Section 9.2 the ground-work established in Chapter 3 is extended to a derivation of the radar equation, which shows how the power detected by a radar system is related to the usual measure of backscattering ability, the differential backscattering cross-section σ^0 . The remainder of the chapter discusses the main types of system that employ this relationship. The first and simplest is the microwave scatterometer (Section 9.3), which measures σ^0 , usually only for a single region of the surface but often for a range of incidence angles. As described here, this is not an imaging system, although the distinction between microwave scatterometers and imaging radars is not a precise one.

The last two sections discuss the true imaging radars. Section 9.4 describes the side-looking airborne radar (SLAR), or real-aperture radar, which achieves a usefully high spatial resolution in one dimension by time-resolution of a very short pulse. Resolution in the perpendicular direction is achieved by using an antenna with a narrow beam-width, i.e. a large antenna. This approach is not feasible for satellite-borne radars, since the antenna needed to produce a useful spatial resolution would be impractically large. Instead, a large aperture (antenna) is synthesised, and the technique is thus known as synthetic aperture radar. This technique is discussed in Section 9.5.

9.1

LiDAR

LiDAR techniques were introduced in Section 8.1, where the simplest form, the laser profiler, was discussed. In that case, the delay time of the returned pulse was the principal variable to be measured. However, it is also, of course, possible to analyse the temporal

structure of the returned signal, and this is the principle of the *backscatter LiDAR*. Backscatter LiDARs are used to calculate vertical profiles or column-integrated values of the backscattering coefficients due to atmospheric constituents such as aerosols and cloud particles (Yorks *et al.* 2011). The horizontal resolution of a backscatter lidar is similar to that of a laser profiler, being set by the angular width of the laser beam, but the vertical resolution is somewhat poorer, being typically 10 to 200 m. This is because the backscattered signal must be integrated over a range of heights to give a detectable output.

Enhancements of the basic backscatter lidar are possible. The *differential absorption lidar* (DIAL) uses a tuneable laser to measure the spectral variation of the backscattered signal, so that atmospheric absorption lines can be distinguished. The *Doppler lidar* or *wind lidar* measures the Doppler shift (Section 2.4) of the backscattered signal (Weissmann and Cardinali 2007). This adds to the backscatter lidar the ability to determine the component of the scattering medium's velocity along the line of sight (i.e. the vertical component for a downward-looking lidar).

Lidars have been operated extensively from aircraft for meteorological sounding. The first spaceborne lidar was *Alissa*, a French system carried on the Mir space station. The *CALIOP* (cloud-aerosol LiDAR with orthogonal polarisation) instrument is currently operational from the CALIPSO mission (Hunt 1980). The CALIPSO (Cloud-Aerosol LiDAR and Infrared Pathfinder Satellite Observations) satellite was launched in 2006 into an orbit with a nominal altitude of 705 km. CALIOP is a two-wavelength (532 and 1064 nm) nadir-viewing atmospheric LiDAR which, as its name suggests, can detect returned radiation in two orthogonal polarisations. It is used for profiling aerosols and atmospheric ice (in clouds), and for measuring the geometry of clouds. It is also used for measuring the vertical component of wind velocities. Its vertical resolution is 30 m and its horizontal field of view is 70 m, sampled every 330 m along track.

Summary

LiDARs employ the same essential principles as laser profilers discussed in Chapter 8, but extract more information from the echo signal. The simplest type is the backscatter LiDAR, in which the time-structure of the echo is used to determine the profile of atmospheric constituents such as water droplets or aerosol particles along the line of sight. The ability to adjust the wavelength of the transmitted pulse or to analyse the wavelength of the echo allows the possibility of identifying specific absorption lines or measuring the Doppler shift (and hence velocity) of the scattering medium. LiDARs are used extensively from aircraft for meteorological observations, and a few have also been deployed from spacecraft.

9.2

The radar equation

The remainder of this chapter discusses radar systems, i.e. scattering systems that operate at microwave frequencies. We have already developed, in Chapters 3 and 7, most of the theory necessary to calculate the response of a radar. In this section we develop this theory a little further in order to derive the radar equation.

9.2 The radar equation

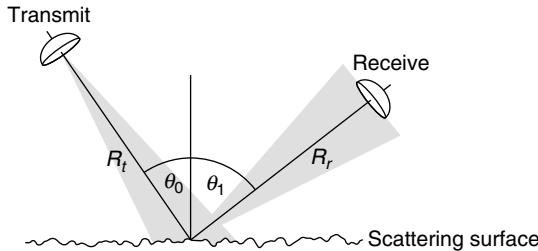


Figure 9.1. Geometry for deriving the bistatic radar equation.

Figure 9.1 shows, schematically, an antenna transmitting microwave radiation towards a surface. Some of this incident radiation is scattered into a range of directions, and some of the scattered radiation is collected by a receiving antenna. The distance from the transmitting antenna to the surface is R_t , and the distance from the surface to the receiving antenna is R_r . We begin by supposing that the transmitting antenna has gain G_t and is transmitting a power P_t . If this power were radiated isotropically, the flux density (power per unit area) at a distance R_t would be given by

$$\frac{P_t}{4\pi R_t^2},$$

so from the definition of antenna gain (Equations 7.5 and 7.6) we can see that the flux density in the direction of the antenna's main beam axis must be given by

$$F = \frac{G_t P_t}{4\pi R_t^2}.$$

If the angle that this radiation makes with the surface normal is θ_0 , the irradiance at the surface is, following the discussion in Section 3.3.1,

$$E = F \cos \theta_0 = \frac{G_t P_t}{4\pi R_t^2} \cos \theta_0.$$

Thus, from the definition of the bistatic scattering coefficient γ (Equation 3.38), the radiance scattered into the direction θ_1 is

$$L = \frac{\gamma E}{4\pi \cos \theta_1} = \frac{G_t P_t \cos \theta_0 \gamma}{(4\pi)^2 R_t^2 \cos \theta_1}.$$

(We should really include the azimuthal components ϕ_1 and ϕ_2 as well as the components θ_1 and θ_2 in specifying the incident and scattered directions. However, we omit them for clarity.)

Next, we suppose that the receiving antenna has an effective area (Section 7.1.1) of A_r and is directed so that its main beam axis points directly at the region illuminated by the transmitting antenna. The solid angle subtended by this antenna at the distance R_r is given by A_r / R_r^2 , so the power that it will collect from an area A of the scattering surface is given by

$$P_r = LA \frac{A_r}{R_r^2} \cos\theta_1 = \frac{A_r G_t P_t \cos\theta_0 \gamma}{(4\pi)^2 R_t^2 R_r^2}. \quad (9.1)$$

This is one form of the *bistatic radar equation*, showing how the received power is related to the transmitted power, the radar geometry and the scattering properties of the surface. However, in all cases that concern us in this chapter we are interested only in the *monostatic radar equation*, where the same antenna is used for both transmitting and receiving radiation. In this case, we can put

$$\begin{aligned} R_t &= R_r = R, \\ \theta_0 &= \theta_1 = \theta, \\ G_t &= G, \\ A_r &= A_e, \end{aligned} \quad (9.2)$$

and the radar equation simplifies to

$$P_r = \frac{A_e G P_t \cos\theta \gamma}{(4\pi)^2 R^4}.$$

This can be rewritten in terms of the backscattering coefficient σ^0 (Equation 3.39) as

$$P_r = \frac{A_e G P_t \sigma^0 A}{(4\pi)^2 R^4}. \quad (9.3)$$

We can simplify this equation even further by noting the relationship between the gain G of an antenna and its effective area A_e (from Equations 7.5, 7.6 and 7.9):

$$A_e = \frac{\lambda^2 G}{4\pi\eta}$$

(where λ is the wavelength and η the efficiency of the antenna), so that

$$P_r = \frac{\lambda^2 G^2 P_t \sigma^0 A}{(4\pi)^3 \eta R^4}. \quad (9.4)$$

Thus, Equation (9.4) (or equivalently 9.2 or 9.3) shows the power received from an area A of scattering surface in the monostatic case. The backscattering coefficient σ^0 , which is dimensionless (it can be thought of as the backscattering cross-section per unit surface area, so its units are m^2/m^2 although it is most commonly expressed in decibels), will in general depend on the incidence angle θ , and possibly also on the corresponding azimuth angle ϕ .

We have made no mention of polarisation in the foregoing discussion. In general, values of σ^0 can be defined for all possible combinations of incident and scattered polarisation states, so that, for example, σ_{HV}^0 is the backscattering coefficient for horizontally polarised incident radiation and vertically polarised scattered radiation. In order to make the argument leading up to Equation (9.4) as general as possible, all polarisation states should be considered. For example, a radar set to receive only horizontally polarised radiation will detect a power proportional to $P_H \sigma_{HH}^0 + P_V \sigma_{VH}^0$ where P_H and P_V are the components of the transmitted power with horizontal and vertical polarisations respectively. A complete description can be given by specifying a *scattering matrix*, showing how the Stokes vector (see Section 2.2) of the radiation is changed.

Summary

The radar equation is fundamental to the operation of imaging radars. It relates the power detected by the radar to the transmitted power, through the gains of the radar antennas, the geometry of the observation, and the ability of the surface to reflect radar radiation. This is most commonly expressed through the backscattering coefficient σ^0 , a pure number that is usually expressed in decibels. It is generally dependent on polarisation. The usual arrangement for a radar measurement is for the same antenna to be used both to transmit and to receive radiation. In this case the *monostatic* radar equation is used.

9.3

Microwave scatterometry

A microwave scatterometer is a non-imaging radar system that provides a quantitative measure of the backscattering coefficient σ^0 , often as a function of the incidence angle θ . It transmits a continuous signal or a series of pulses, the return signal is recorded, and its strength is used in conjunction with the radar equation (e.g. Equation 9.4) to determine the value of σ^0 for that part of the surface which is illuminated. It is especially useful if the scatterometer can be operated in such a way as to yield the value of σ^0 as a function of the incidence angle θ , since this function often allows the surface material to be identified or its physical properties to be deduced (see the discussion of microwave backscattering in Sections 3.3.4 and 3.6.5). There are three principal methods of achieving this. One is to use a narrow-beamwidth scatterometer that can be steered to point at the desired target area. As the platform (aircraft or satellite) carrying the scatterometer moves, the radar tracks the target area and the backscattering curve is built up. The second method is to use *Doppler processing* of the signal.

Let us consider a scatterometer with a power pattern that is broad in the along-track direction but narrow in the perpendicular, across-track, direction (Figure 9.2). The scatterometer beam is inclined so that it looks forward. At any instant, the return signal is derived from a large range of angles $\Delta\theta$ (the beamwidth of the antenna), and hence from a long strip of the surface being sensed. The signal returned from the point X will be Doppler shifted to a frequency $f_0 + df$, where f_0 is the transmitted frequency and df is given, following Equation (2.21), by

$$\delta f = \frac{2f_0 v}{c} \sin\theta_0. \quad (9.5)$$

In this equation, v is the platform velocity and c is the speed of light. This Doppler shift is unique to the incidence angle θ_0 , so by feeding the return signal into a bank of filters tuned to select different Doppler shifts, data from a range of incidence angles can be extracted.

The third method of scanning a range of incidence angles is to transmit very short pulses of radiation, and to analyse the time-structure of the returned signal. Unlike the Doppler method, this does not rely on motion of the scatterometer, and it simplifies the analysis if we assume that the platform is stationary.

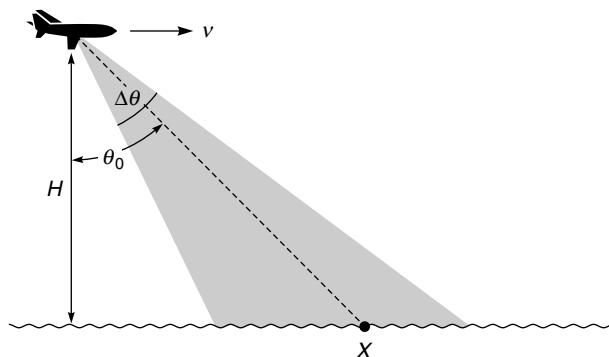


Figure 9.2. Principle of operation of a Doppler scatterometer. The radar emits a broad beam of angular width $\Delta\theta$, but radiation scattered from the point X (at incidence angle θ_0) can be identified by its Doppler shift.

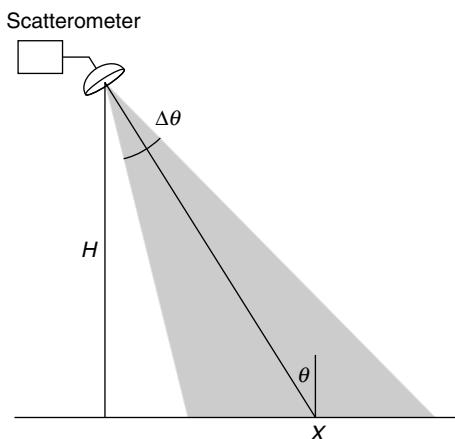


Figure 9.3. Principle of operation of a time-domain scatterometer. The radar emits a short pulse into a broad beam of angular width $\Delta\theta$, but radiation scattered from the point X (at incidence angle θ) can be identified by its time delay.

Again we assume that the antenna power pattern is broad in one dimension (with a beamwidth $\Delta\theta$) and narrow in the perpendicular dimension, although since we are assuming that the scatterometer is stationary the orientation of the beam is unimportant as long as it is obliquely inclined to the surface (Figure 9.3). The two-way propagation time from the scatterometer to the point X and back again is

$$\frac{2H}{c \cos \theta},$$

so by resolving the time-structure of the returned pulse we can uniquely identify the contribution for the incidence angle θ . We may note that the ability to resolve the incidence angle is equivalent to spatial resolution, and that this approach to achieving spatial resolution is very similar to that of a pulse-limited radar altimeter (Section 8.2.1).

9.3.1

Applications of microwave scatterometry

The useful output of a scatterometer, however it is realised, may be regarded as a plot of σ^0 as a function of the incidence angle θ , or, at least, one or more points representing this function. In Chapter 3, and especially in Sections 3.3.4 and 3.6.5, we discussed the principles that relate the function $\sigma^0(\theta)$ to the physical properties of the scattering medium. In general, we can state that the overall level of backscattering will be determined by the dielectric properties of the scattering medium (and its internal structure, if volume scattering is significant), while the dependence on incidence angle will be governed primarily by the surface geometry. A specularly smooth surface should, in principle, show a delta-function in a plot of $\sigma^0(\theta)$, centred at the value of θ that gives specular reflection into the radar. If the surface is horizontal, this will be $\theta = 0$. Real surfaces, however, are unlikely to be specularly smooth, especially at shorter wavelengths (see the discussion of the Rayleigh criterion in Section 3.3.3), and in addition the plot of $\sigma^0(\theta)$ will be convolved with the antenna power pattern. Thus a true delta-function will never in practice be realised, and the plot of $\sigma^0(\theta)$ for a smooth surface will look more like Figure 9.4.

At the opposite extreme, a Lambertian ('perfectly rough') surface has γ proportional to $\cos \theta$ (Sections 3.3.1, 3.3.2), so that σ^0 will be proportional to $\cos^2 \theta$. This has a characteristic shape if σ^0 is plotted logarithmically (that is, as decibel values: Figure 9.5). Real materials lie somewhere between these extremes, unless volume scattering is also important in which case the variation of σ^0 with incidence angle may be more complicated.

Even if a satisfactory physically based model of the backscattering is not available, a plot of $\sigma^0(\theta)$ may still be diagnostic of a particular surface material. Examples of such plots were given in Section 3.6.5. Of course, more information can be obtained if observations can be made at more than one frequency or polarisation. Multiple-frequency observations are difficult to make even from aircraft, because of the technical complexity of providing different 'front ends' for the radar (or the weight burden of carrying several radars), but multiple polarisations are easier to observe since little change needs to be made to the radar hardware.

9.3.1.1

Microwave scatterometry over ocean surfaces

The major application of microwave scatterometry to ocean surfaces is in determining wind velocity (Liu 2002). This is somewhat similar in principle to the determination of

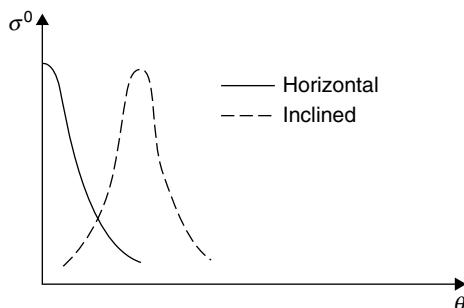


Figure 9.4. Variation of σ^0 with incidence angle for a smooth surface (schematic).

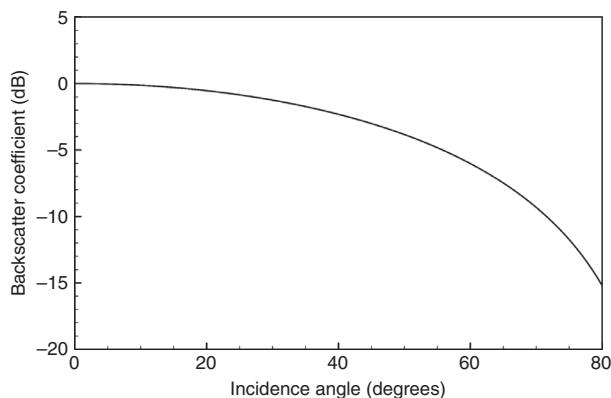


Figure 9.5. Variation of backscattering coefficient with incidence angle for a Lambertian scatterer. The backscattering coefficients are normalised to the value at normal incidence.

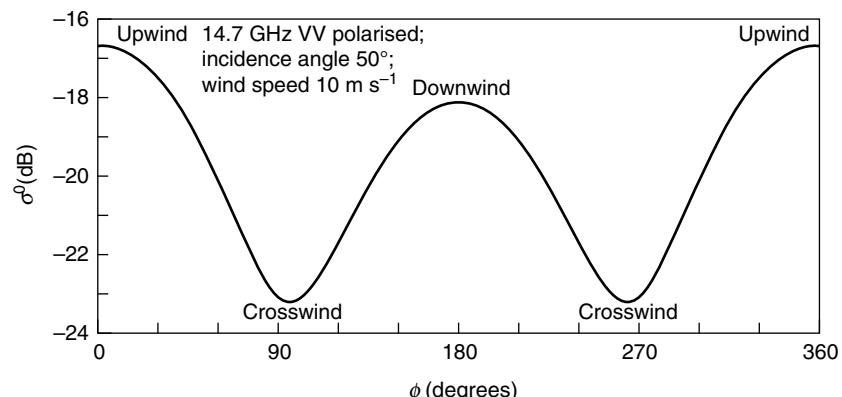


Figure 9.6. Typical variation of microwave backscattering coefficient with azimuth angle ϕ over a rough ocean surface.

wind speed from the significant wave height using radar altimetry (see Chapter 8), although more information is available. The method relies on a model relating the roughness of the sea to the wind speed. The roughness is anisotropic (this fact was mentioned in Section 3.6.5), as might be expected since the crests and troughs of the surface wave field tend to align themselves perpendicularly to the wind direction, and this is the key to determining the wind direction as well as the wind speed.

This is shown in Figure 9.6, which has been derived from a simple empirical model of microwave backscatter from ocean waves rather than directly from experimental data. It shows the effect of varying the azimuth angle ϕ while keeping the incidence angle, frequency and polarisation constant. The azimuth angle is defined such that $\phi = 0$ corresponds to the case when the horizontal component of the direction of the incident microwave radiation is opposite to the wind velocity vector, i.e. when the scatterometer is looking upwind. The figure shows a strong contrast between the upwind and crosswind directions, and a much weaker contrast between the upwind and downwind directions.

9.3 Microwave scatterometry

The azimuthally dependent part of the model of sea surface backscatter used to construct Figure 9.6 can be written as

$$\sigma^0 = A + B \cos\phi + C \cos 2\phi, \quad (9.6)$$

where A , B and C are ‘constants’ that depend on the frequency, polarisation, state, incidence angle and wind speed. By making at least three scatterometer observations in different azimuthal directions, the values of A , B and C , and hence the corresponding wind velocity, can be determined. This is illustrated schematically in Figure 9.7 (in practice, a least-squares method, rather than the graphical approach suggested by the figure, is used). Here it is assumed that three observations have been made, all at the same frequency, polarisation and incidence angle, but with azimuth directions of 0 (i.e. looking north), 90° (east) and 180° (south). In each case, the observed value of σ^0 is consistent with a range of possibilities for the wind velocity. These are plotted as three curves in Figure 9.7. For example, the curve (a) represents all the combinations of wind speed v and direction ψ consistent with the observed backscattering coefficient at azimuth 0. From the figure, we see that there is a mutual intersection of all three curves at a unique point, namely a wind speed of 12.3 m s^{-1} in the direction $\psi = 47^\circ$. We note, however, that there is also a ‘near intersection’ at about 12.5 m s^{-1} , 310° . Thus, if the scatterometer data are rather noisy, an unambiguous determination of the wind velocity may not always be possible.

Microwave scatterometry can give wind speeds, over the ocean, to an accuracy of about 2 m s^{-1} , and wind directions to an accuracy of about 20° , at least in the absence of rainfall. In the presence of rain, scattering from the falling droplets or from the rain-roughened sea can give rise to anomalous results. Nevertheless, scatterometry currently represents the most accurate technique for obtaining wind velocities over oceans (Figure 9.8).

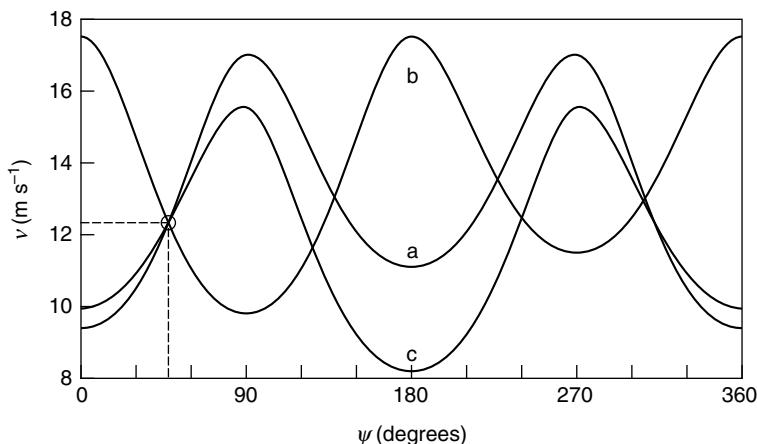


Figure 9.7. Determination of wind velocity over an ocean surface using three microwave scatterometer measurements. Curve (a) shows all values of the wind velocity (expressed as speed v and azimuth direction ψ) consistent with a certain value of backscattering coefficient measured by looking north. Curves (b) and (c) correspond to observations looking east and south, respectively. There is a unique intersection of all three curves at $v = 12.3 \text{ m s}^{-1}$, $\psi = 47^\circ$.



Figure 9.8. Composite image derived from SeaWinds scatterometer data collected by the ADEOS-2 (Midori-2) satellite on 28 January 2003. Variations in backscattering coefficient over land surfaces are represented in shades of green. Over oceans, the colours represent different wind speeds from blue (low) to red (high), and arrows show the inferred circulation of the wind. (Image downloaded from <http://earthobservatory.nasa.gov/IOTD/view.php?id=3248> and reproduced by courtesy of NASA/JPL SeaWinds science team.) See also colour plates section.

A second oceanographic application of microwave scatterometry is the delineation and characterisation of sea ice (Meier and Stroeve 2008). The technique can achieve much higher spatial resolutions than is possible with passive microwave radiometry, discussed in Chapter 7.

9.3.1.2 Microwave scatterometry over land surfaces

Microwave scatterometry has been used extensively for characterising geological materials, using the variation of σ^0 with θ as a signature in much the same way as materials are identified in the optical band by their spectral signatures. The technique has also been used for studying soil surfaces, where the principal retrievable parameters are moisture content (Wagner, Lemoine and Rott 1999), roughness and texture. It should be noted, however, that the greater spatial complexity of most land surfaces, when compared with oceans and sea ice, means that an imaging radar system is usually preferable to a simple scatterometer when interpreting surface types. Microwave scatterometry has also found applications in the remote sensing of vegetation, particularly crops and forests. These can present a substantial theoretical problem because of their complicated geometries and comparatively open structures, with significant volume and surface scattering as well as, in some cases, scattering from the ground. Multitemporal scatterometer observations have proved useful for studying seasonal dynamics of vegetation (Frison and Mougin 1996). Scatterometer data have also been used to study ice and snow properties, frozen ground

9.3 Microwave scatterometry

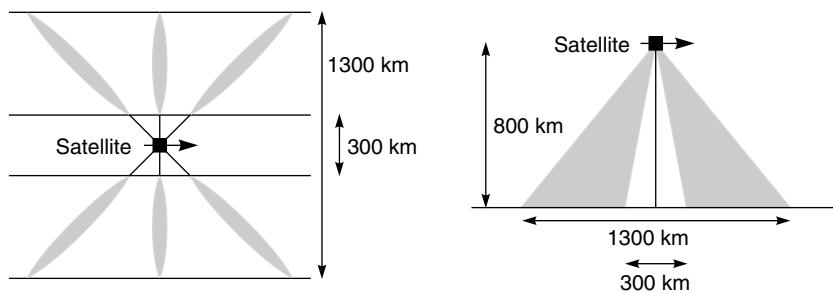


Figure 9.9. Geometry of the ASCAT scatterometer (schematic). Left: plan view; right: side view.

and freeze–thaw processes (Bartsch *et al.* 2007, Drinkwater, Long and Bingham 2001, Kimball *et al.* 2001, Remund, Long and Drinkwater 2000)

9.3.2 Example: ASCAT

In this section we discuss the operation of a typical spaceborne microwave scatterometer. This is the *ASCAT* instrument carried on the *Metop* satellite at an altitude of around 840 km. *ASCAT* is a C-band (5.3 GHz) VV-polarised scatterometer. Like most spaceborne scatterometers, *ASCAT* uses short pulses to obtain its spatial resolution. It emits radiation in six fan-beams (beams narrow in one dimension and broad in the perpendicular dimension), three on each side of the satellite, so that the different azimuths of these beams can be used to obtain wind velocities over the ocean as discussed in Section 9.3.1.1. The geometry of the fan-beams is shown in Figure 9.9. They are arranged so that they intersect the Earth’s surface in swaths 500 km wide, extending from 150 to 650 km from the sub-satellite track. Two of these beams are perpendicular to the direction of the satellite’s motion; these beams cover a range of incidence angles from 12° to 44°. The other four beams are arranged at 45° to the direction of motion, and these intersect the surface at incidence angles from 17° to 55°.

The instrument emits short pulses from each of its six antennas in turn. Time-resolution of the returned signal gives a spatial resolution of about 45 km. Since any point on the Earth’s surface that falls within one of the two swaths of the instrument will be seen by all three of the fan-beams on that side, a ‘triplet’ of σ^0 values can be collected for input into the type of algorithm discussed in Section 9.3.1.1.

Summary

Microwave scatterometry is in general a non-imaging technique that provides quantitative measurement of the backscattering coefficient σ^0 , often as a function of the incidence angle θ . This can be achieved using a narrow-beam radiometer that can be steered to point at the area of interest, or by using a broad-beam radiometer and resolving the return signal by frequency or (if a short pulse is transmitted) by time in order to determine the location of a scatterer within the beam. The dependence of σ^0 on θ is characteristic of the type of

scattering: surface scattering from a very rough surface will give Lambertian scattering for which σ^0 is proportional to $\cos^2\theta$, while a very smooth surface will give high values of σ^0 close to $\theta = 0$. The dependence of σ^0 on the polarisation state may also be diagnostic.

Microwave scatterometry is especially valuable over ocean surfaces, where it is used to infer wind velocity from the anisotropic roughening of the water surface and to measure sea ice with a higher spatial resolution than is possible with passive microwave radiometry. Over land surfaces, microwave scatterometry is valuable for characterising geological materials and soil, vegetation canopies, ice, snow and frozen ground.

9.4

Real-aperture imaging radar

Microwave scatterometers can be considered as imaging systems, although with a rather poor spatial resolution. In effect, they have sacrificed spatial resolution in order to achieve good radiometric resolution. In Sections 9.4 and 9.5 we consider radar systems in which the ability to generate images of reasonably high spatial resolution is a primary consideration. In this section we discuss the real-aperture radars (RARs), usually referred to as *side-looking radars* (SLRs) or side-looking airborne radars (SLARs).

The SLR technique evolved in the 1950s, as a tool for military reconnaissance, from the plan-position indicator (PPI) radars developed during the Second World War. The SLR is a pulsed radar system that looks to one side of the flight direction (hence ‘side-looking’) and is capable of producing a continuous strip image of the target area.

The basic idea of SLR is very similar to that of the fan-beam scatterometer shown in Figure 9.9. Just one of the fan-beams is used, and for a definite example we assume that it is the middle beam on the right of the platform’s motion. High spatial resolution in the along-track direction is achieved by ensuring that the narrower dimension of the fan-beam is as narrow as possible, i.e. by arranging that the antenna is as long as possible in the along-track direction. Spatial resolution in the across-track direction is achieved by transmitting very short pulses (strong pulse compression, as discussed in Section 8.2.7, is used) and time-resolving the returned signal. The main difference between an SLR and a scatterometer like ASCAT, apart from the fact that the latter will usually have several beams, is the use of very short pulses by the SLR.

Figure 9.10 shows the basic geometry of an SLR system. The upper part of the diagram shows the view from behind, so that the platform carrying the instrument is flying ‘into the page’. The antenna, which has a width w , emits radiation into a beam of angular width ψ , which will normally be set by the diffraction limit and therefore given by

$$\psi \approx \frac{\lambda}{w}. \quad (9.7)$$

The intersection of this beam with the Earth’s surface defines the width of the imaged swath. The transmitted radiation consists of a short pulse. We can use the same concept of a scattering region, introduced in Section 8.2.1, that propagates away from the antenna at a speed of $c/2$. This region is shown shaded in Figure 9.9, and its intersection with the Earth’s surface is the region from which backscattered radiation is instantaneously detected.

The lower part of Figure 9.10 shows a plan view of the same situation. The intersection of the scattering region with the surface is approximately rectangular. Its length in the

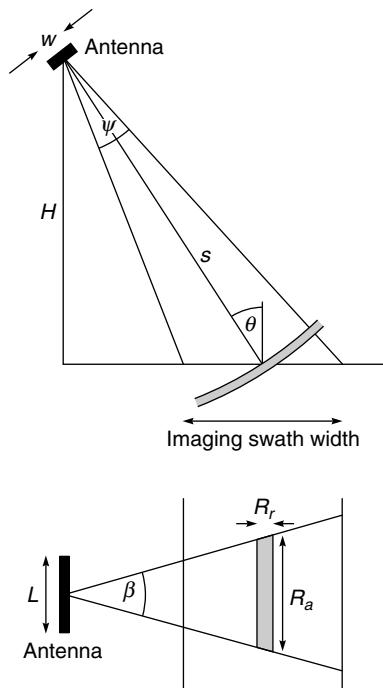


Figure 9.10. Geometry of an SLR system. Above: viewed from behind (the platform is flying ‘into the page’); below: viewed from above (the platform is flying ‘up the page’).

along-track direction (also referred to as the *azimuth direction*) is governed by the beamwidth β of the antenna in this direction, and this is set by the length L of the antenna. The diffraction limit will normally mean that

$$\beta \approx \frac{\lambda}{L}. \quad (9.8)$$

The length of the scattering strip in the azimuth direction is the azimuth resolution R_a of the system. Taking the azimuthal beamwidth as β (assumed to be much less than 1 radian) and the *slant range* from the antenna to the scattering region to be s , we must have $R_a \approx s\beta$, so we can see that the resolution will vary across the swath, being poorer at the farther edge and better at the nearer edge. If the effect of the Earth’s curvature can be neglected, we can write the slant range in terms of the height H and the incidence angle θ :

$$s = \frac{H}{\cos\theta}, \quad (9.9)$$

so, with the approximation of Equation (9.8), the *azimuth resolution* becomes

$$R_a \approx \frac{H\lambda}{L \cos\theta}. \quad (9.10)$$

If the length (duration) of the radar pulse is t_p , the *slant-range resolution* Δs is $ct_p/2$ – in other words, time-resolution of the returned signal allows two scatterers to be

discriminated if their distances from the radar differ by at least this amount. By simple trigonometry, we can show that the *range resolution* (in the range direction, or across-track direction) is then given by

$$R_r \approx \frac{ct_p}{2\sin\theta}. \quad (9.11)$$

As an example, we consider an SLR system operating at $\lambda = 1$ cm, with an antenna length of 5 m and a pulse length of 30 ns from an aircraft at an altitude of 6000 m. At a distance of 10 km from the ground track the incidence angle $\theta = 59^\circ$, so $R_a = 23$ m and $R_r = 5.2$ m. 25 km from the ground track, the incidence angle is 77° , the azimuth resolution has coarsened to 42 m but the range resolution is virtually unchanged at 4.7 m.

We note from Equation (9.11) that the range resolution is independent of the platform height H , and can be made as small as 10 m, or less, provided that the incidence angle θ is not too small. The azimuth resolution R_a , on the other hand, is proportional to the platform height. Thus, although resolutions of the order of 10 m can be achieved from airborne systems, much poorer resolutions are available from spaceborne systems. This difficulty can be circumvented by the use of SAR systems, discussed in Section 9.5.

9.4.1 Image distortions

The oblique imaging geometry, and the fact that the range (across-track) coordinate is determined from the slant range, introduces characteristic geometrical and radiometric distortions into SLR images. The simplest of these is *slant-range distortion*. It arises only in the simplest form of signal processing, when the image is presented in such a way that it is the slant range, rather than the ground range, that increases uniformly across the image. It is illustrated schematically in Figure 9.11a.

Slant-range distortion is relatively straightforward to correct, since it can be described by a small number of variables. Some SLR systems incorporate a correction within the radar's signal processing unit itself.

If the Earth's surface has appreciable relief, further geometrical distortions will occur as a result of the imaging method. These are the phenomena *layover* and *shadowing*.

Layover arises from the fact that the pulse delay time is used to determine the across-track coordinate of a scatterer. Figure 9.11b shows a simple topography with five

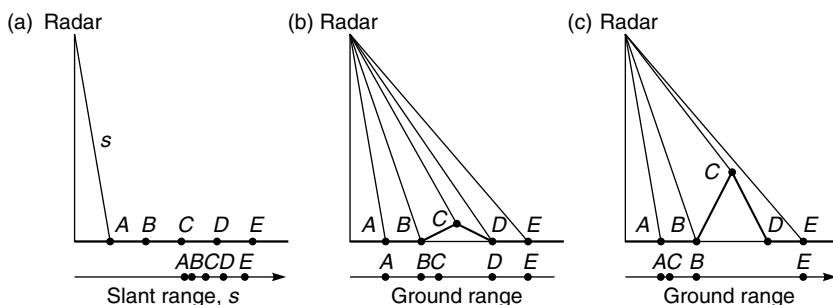


Figure 9.11. Image distortion phenomena in SLR imaging: (a) slant-range distortion; (b) layover; (c) shadowing.

9.4 Real-aperture imaging radar

scatterers labelled *A* to *E*. Scatterers *A*, *B*, *D* and *E* are at the same altitude as one another, and this is equal to the altitude of the ground surface assumed for the slant-range to ground-range correction. However, scatterer *C* is located above the reference plane. The reduced slant-range is erroneously interpreted as a reduced across-track coordinate, and this is shown schematically in Figure 9.11b by a displacement of the point *C* to the left of its correct position.

Layover can be corrected if the surface topography is known, although the procedure is a laborious one. An uncorrected image of an area of high relief, e.g. a mountainous area, has a characteristically strange appearance in which the mountains appear to lean towards the radar. Figure 9.11b also demonstrates a radiometric consequence of layover. We can observe that, although the distances *BC* and *CD*, measured along the terrain surface, are similar to the distances *AB* and *DE*, the distance *BC* in the image will be shortened, and the distance *CD* lengthened, relative to flat terrain. As a result, the number of scatterers per unit length, measured on the image, is higher than normal in *BC*, with the consequence that this region will appear unusually bright. This phenomenon is known as *highlighting*, and is a purely geometrical effect (although it is likely to be enhanced by the fact that the backscattering coefficient σ^0 is usually significantly larger at small local incidence angles). Conversely, the density of scatterers within *CD* is lower than normal, so this region of the image will appear unusually dark. Although the term does not seem to be in widespread use, an obvious name for this phenomenon is *lowlighting*. Layover and highlighting effects can be clearly seen in Figure 9.12, which is in fact a Synthetic Aperture Radar (SAR) image rather than an SLR image.

Shadowing is the phenomenon in which one part of the surface is hidden from the radar's view by another. The shadowed region receives no radar illumination, and consequently no signal is returned from it. Unlike optical shadows, radar shadows are

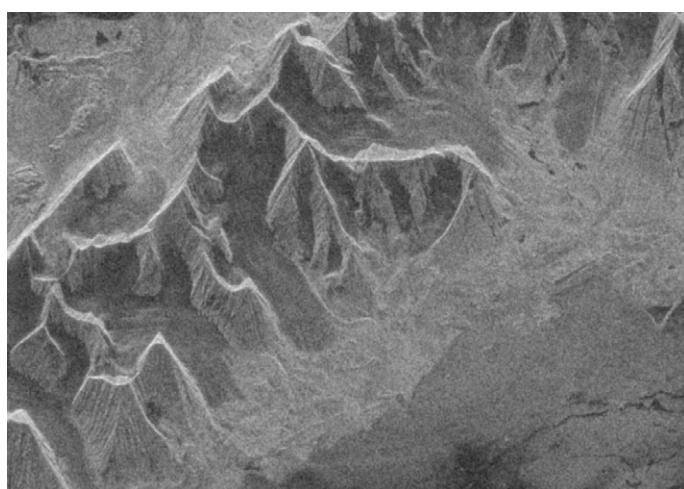


Figure 9.12. Extract of an ERS-1 radar image of the area around Ny Ålesund, Svalbard, showing the phenomena of layover (the mountains appear to lean towards the top of the image) and highlighting (the sides of the mountains facing the top of the image are unusually bright). The image was acquired on 20 July 1999. (Image © ESA 1999.)

completely dark, because scattering of microwave radiation from the atmosphere into the shadowed region is entirely negligible. Shadowing is illustrated schematically in Figure 9.11c, where the scatterer at *D* is obscured by the terrain in the vicinity of *C* and so does not appear in the image. It is clear from the figure that shadowing is a phenomenon associated with steep slopes away from the radar. However, the figure also shows that steep slopes *towards* the radar will introduce image distortions. Here, the layover of the scatterer at *C* is so large that the points *B* and *C* have been imaged in the reverse order. This can be thought of as an extreme example of the highlighting phenomenon.

9.4.2 Instruments and applications

As we remarked earlier, SLR systems are generally suitable for use only from airborne platforms, since the azimuth (along-track) resolution of a spaceborne system is normally unacceptably poor. Airborne SLRs are extensively used for military and other reconnaissance (for example, detection of icebergs). Many applications are essentially equivalent to those of SAR systems, and we therefore defer a more detailed discussion to Section 9.5.6.

Some SLRs have been deployed in space. The Ukrainian *RLSBO* instrument was carried on the *Okean-O* satellites from 1994 to 2002 and the *Sich* satellites from 1995 to 2006. These satellites had an orbital height of around 650 km. The RLSBO used an 11.1 m antenna and operated at a frequency of 9.7 GHz (V polarised), achieving an azimuth resolution of 2.1–2.8 km and a range resolution of 0.7–1.2 km. The swath width was around 470 km (Figure 9.13). RLSBO imagery was used for measurements of ocean, sea ice and ice sheet surfaces.

Summary

A real-aperture imaging radar, or side-looking radar (SLR), transmits a beam of microwave radiation that is narrow in the direction of motion of the platform (the along-track, or azimuth direction) and broad in the perpendicular (cross-track, or range) direction. The beam is transmitted to one side of nadir. High spatial resolution in the azimuth direction is achieved through the narrowness of the beam (in turn achieved by using a long antenna) while high spatial resolution in the range direction is achieved by transmitting short pulses and time-resolving the return signal. The image is subject to a number of distortions. *Slant-range distortion* is a consequence of the fact that the time of flight from the radar to a scatterer on the ground and back to the radar does not vary in a linear manner with the distance of the scatterer from the nadir, but it is easily corrected. It occurs even over flat terrain, while *layover* is a form of distortion that is a consequence of relief in the terrain. This can be corrected if a suitable digital elevation model is available. Relief also causes radiometric variations called *highlighting*, and some regions may be shadowed by the terrain, in which case no signal at all is returned from them.

Although the cross-track resolution of an SLR system is controlled by the pulse length, the along-track resolution depends on the range from the antenna to the surface. This implies that SLR systems are more suitable for airborne than spaceborne use, although the technique has sometimes been used from space, where it generally achieved a spatial resolution of a few kilometres.

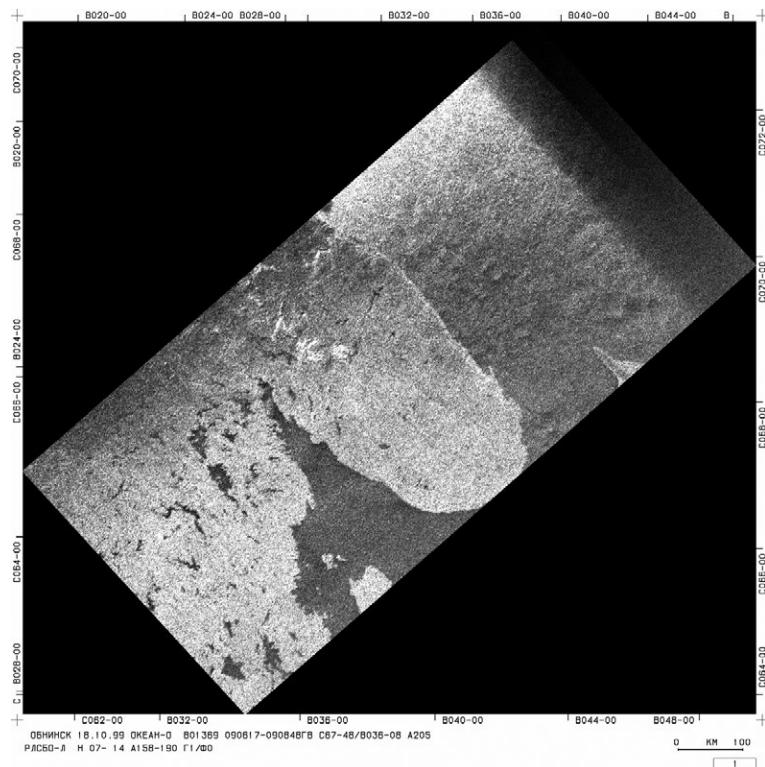


Figure 9.13. RLSBO SLR imagery of north-western Russia, recorded on 18 October 1999. (Image downloaded from <http://damocles.ntsomz.ru/database/index.php> and reproduced by courtesy of the Research Centre for Earth Operational Monitoring (NTsOMZ), Moscow.)

9.5

Synthetic aperture radar

The synthetic aperture radar (SAR) technique overcomes the problem of the altitude-dependence of the azimuth resolution of an SLR system (Equation 9.10). In external appearance, a SAR system is indistinguishable from an SLR. The same geometry applies, so that Figure 9.10 describes both systems, and the same technique of emitting a very short pulse and analysing the temporal structure of the return signal is used to obtain resolution in the range direction. Higher resolution in the azimuth (along-track) direction is achieved by sophisticated signal processing.

Equation (9.10) shows that improved azimuth resolution can be obtained from a longer antenna. Instead of making an antenna that is physically longer, the SAR technique relies on the motion of the platform. During some time interval T , the antenna is carried through a distance vT (where v is the platform velocity), so if we record the signal collected at the antenna during this interval, it ought to be possible to use it to reconstruct the signal that would have been collected from an antenna of length vT . This is the idea of the ‘synthetic aperture’. In order to discuss how this is achieved in practice, we consider the simple

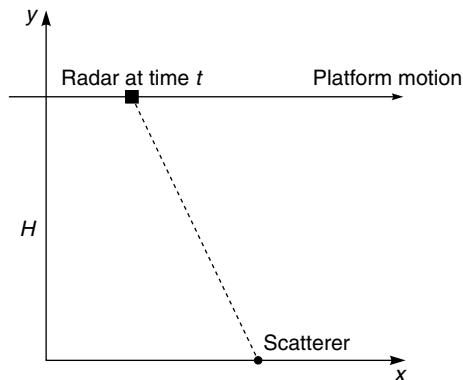


Figure 9.14. Simple geometry for considering SAR imaging. At time t , the radar is at $(vt, H, 0)$ and a scatterer has the fixed coordinates $(x, 0, 0)$.

imaging geometry shown in Figure 9.14. In this somewhat unrealistic case, the radar is looking vertically downwards (i.e. it is not side-looking), and we ignore the Earth's curvature. The radar is moving at a constant velocity v parallel to the x -axis of a Cartesian coordinate system, such that its position at time t is given by $(vt, H, 0)$. The radar's beam is broad in the azimuth direction, so that the radar can 'see' a large range of values of x . Somewhere within this range there is a scatterer with coordinates $(x, 0, 0)$.

When the time t is less than x/v , the radar is approaching the scatterer and the return signal will therefore be Doppler-shifted upwards. At $t = x/v$ the Doppler shift is zero, and at later times it is negative. Thus, by extracting the component of the return signal that has just the right variation of frequency with time, we can resolve any desired value of x . (Note that this is very similar to the Doppler processing for microwave scatterometry that was discussed in Section 9.1). Clearly, the ability to extract different frequency components from the return signal requires that both its amplitude and phase should be stored, not just its intensity. This means that the transmitted radiation should be *coherent* (i.e. it should have a definite phase) and that it should be detected coherently. The path length from the radar to the scatterer at time t is $(H^2 + (vt - x)^2)^{1/2}$, so the phase delay for the two-way path from the radar to the scatterer and back to the radar is

$$\Delta\phi = 2k(H^2 + (vt - x)^2)^{1/2}, \quad (9.12)$$

where k is the wavenumber of the radiation (Figure 9.15). To 'focus' the data on a particular value of x , the phase variation given by Equation (9.12) is subtracted from the data.

The length of the synthetic aperture is vT , where v is the platform velocity and T the time during which data are coherently collected for subsequent processing to generate the image. Since we expect the azimuth resolution to improve as the length of the synthetic aperture is increased, it appears that there should be no limit to the azimuth resolution. This is, however, not true, and we can estimate the best (finest) resolution as follows. If the length of the real antenna is L , it will have an angular beamwidth in the azimuth direction of roughly λ/L . Referring to Figure 9.14, we see

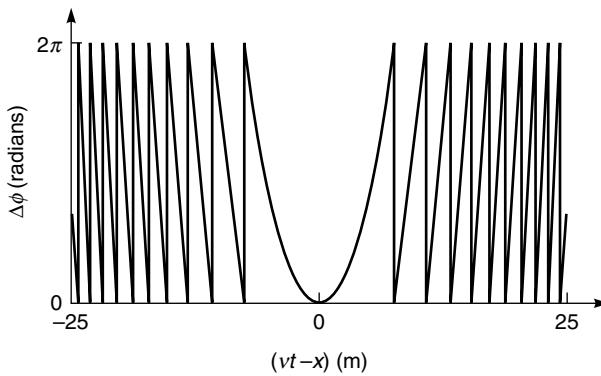


Figure 9.15. Phase delay $\Delta\phi$ given by Equation (9.12) for the case when $H = 1000$ m, at a frequency of 5 GHz. Values of $\Delta\phi$ greater than 2π have been reduced to the range 0 to 2π by subtracting integer multiples of 2π .

that this implies that a scatterer located at x will only be within the radar beam, and hence visible to the radar, for times t between

$$\frac{x}{v} - \frac{H\lambda}{2Lv}$$

and

$$\frac{x}{v} + \frac{H\lambda}{2Lv}.$$

Thus, the maximum useful length of the synthetic aperture is $H\lambda/L$. The angular beamwidth, again in the azimuth direction, of this aperture is therefore roughly $\lambda/(H\lambda/L) = L/H$, giving a linear resolution at the surface of roughly L .

This is only an approximate calculation, although the result is almost correct (it should actually be $L/2$). A more exact calculation is given in Section 9.5.1. However, our simple derivation shows that the best possible resolution is independent of the distance from the antenna to the surface, and that finer resolutions are, at least in principle, achievable using smaller antennas. Both of these results are contrary to the results we have obtained for other imaging systems. We can now see that spatial resolutions of the order of 10 m are feasible even from spaceborne systems.

We should note that, if the optimum resolution of $L/2$ is to be achieved, it is necessary to preserve the coherence of the transmitted radar signal for a time $2H\lambda/Lv$. In practice, the maximum useful length of the synthetic aperture may be limited by the *coherence time* of the radiation rather than by the beamwidth of the antenna. (The coherence time t_c is related to the coherence length l_c , introduced in Equation (8.11), by $l_c = ct_c$; c is the speed of light.) For example, suppose we consider a spaceborne SAR system designed to achieve an azimuth resolution of 5 m. The beamwidth criterion merely requires that the antenna should not be longer than 10 m. However, if we assume that $H = 800$ km, $\lambda = 6$ cm and $v = 7 \text{ km s}^{-1}$, we see that the coherence time of the transmitted radiation must be at least 1.4 seconds. This would require that the transmitted radiation should have a bandwidth of no more than about 1 Hz.

9.5.1

More exact treatment of the azimuth resolution

The argument presented in the previous section was somewhat approximate. In this section we re-derive the result that the best possible azimuth resolution is given by $L/2$, using a somewhat more rigorous argument. No new principles are introduced here, and the section may be omitted by readers who do not share my enthusiasm for Fourier transforms. It does, however, illustrate in greater detail the idea of ‘phase unwrapping’ discussed in the previous section.

Figure 9.16 shows the necessary geometry. The antenna moves such that its position A at time t is given by $(x_a, H, 0)$ where $x_a = vt$, and the data will be processed so as to focus on the origin O . We wish to calculate the response of the system to a nearby point P with coordinates $(x_p, 0, 0)$. The signal $a(x_a, x_p)$ received from P when the antenna is at A is just

$$a(x_a, x_p) = f(\beta) \exp(2ik\Delta r), \quad (9.13)$$

where the distance AO exceeds the distance AP by Δr . (The reason that Δr , rather than the distance AP , appears in this expression is that the phase delay introduced by the path from A to O and back again is subtracted in the process of focussing the data.) $f(\beta)$ is the amplitude response of the antenna in the direction β , and k is the wavenumber of the radiation.

When the signals from different positions of the antenna are combined, Equation (9.13) is integrated over x_a . For the present, we will suppose that the limits of the integration are given by $|x_a| \leq X/2$, in other words that the length of the synthetic aperture is X . Using the approximations $\Delta r \approx \beta x_p$ and $x_a \approx \beta H$, we thus obtain the expression

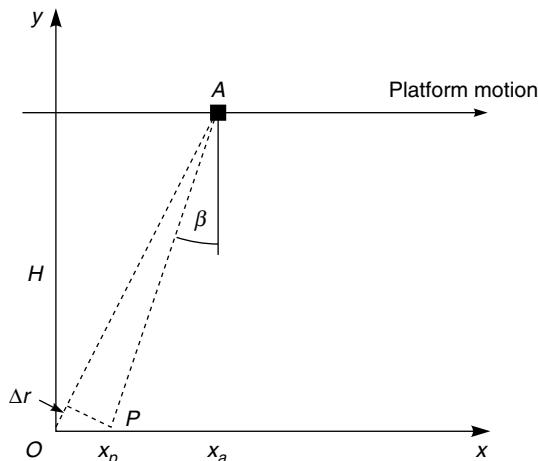


Figure 9.16. Geometrical construction for calculating the azimuth resolution of a SAR. The antenna A has coordinates $(x_a, H, 0)$ and the data are processed to focus on the origin O . The text shows how the response to a scatterer at the point P can be calculated.

$$a(x_p) = \int_{-x/2H}^{x/2H} f(\beta) \exp(2ik\beta x_p) d\beta \quad (9.14)$$

for the total amplitude $a(x_p)$ collected from the point P (we are neglecting constant factors that can be taken outside the integral).

Now the antenna's amplitude response $f(\beta)$ is obtained from the Fraunhofer diffraction pattern of its amplitude distribution. We can write this amplitude distribution as $A(y)$, where y is distance measured in the along-track direction from the centre of the antenna, so that, for example, a uniformly illuminated antenna of length L will have $A(y) = 1$ for $|y| \leq L/2$, 0 otherwise. Whatever the form of $A(y)$, we have from Equation (2.41) that

$$f(\beta) = \int_{-\infty}^{\infty} A(y) \exp(iky\beta) dy \quad (9.15)$$

provided that the antenna is long compared with the wavelength of the radiation.

Now we can substitute Equation (9.15) into (9.14). We also assume that X is infinite, so as to calculate the best possible azimuth resolution of the system. Thus

$$a(x_p) = \int_{-\infty}^{\infty} A(y) \int_{-\infty}^{\infty} \exp(ik\beta(y + 2x_p)) dy d\beta.$$

From the definition of the Dirac delta-function given in Section 2.3, we can see that this simplifies to

$$a(x_p) = \int_{-\infty}^{\infty} A(y) \delta(k\beta(y + 2x_p)) dy$$

(we are again ignoring constant factors that can be taken outside the integral), and hence to

$$a(x_p) = A(-2x_p). \quad (9.16)$$

We have arrived at the desired result. Equation (9.16) shows that the response of the SAR system has exactly the same shape as the antenna's aperture function, but is half as wide. For example, the case of a uniformly illuminated antenna of length L gives $a(x_p) = 1$ for $|x_p| \leq L/4$, 0 otherwise. Thus, we have located the factor of 2 missing from the treatment in Section 9.5, and we have also discovered the shape of the SAR response pattern.

9.5.2 Speckle

We noted, in Section 9.5, that SAR is necessarily a *coherent* imaging technique, meaning that both the amplitude and phase of the received signal, and not just the intensity, are significant. An important consequence of this fact is that SAR images contain a characteristic type of granularity or image noise termed *speckle*. This adds to the uncertainty with which the backscattering coefficient σ^0 can be determined from a SAR observation. In this section we develop a simple one-dimensional model of this image speckle.

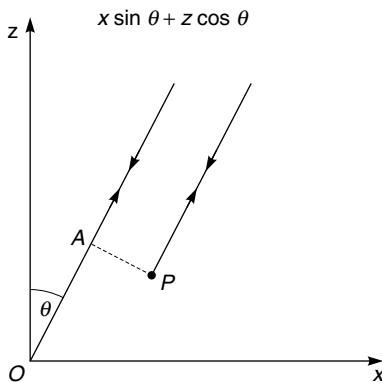


Figure 9.17. Geometry for calculating the effects of image speckle.

We begin by assuming that the radar system observes a surface that is nominally flat and uniform, consisting of isotropic point scatterers (this is the same starting point that we used to consider the behaviour of a radar altimeter in Section 8.3). These scatterers are at various heights z above some datum, in order to model a rough surface. Figure 9.17 shows the necessary geometry.

The radar is located in the direction θ . The scattering surface does not necessarily pass through the point O , the origin of the coordinate system; this is just a convenient reference point. However, there is a scatterer at the point P , with coordinates (x, z) . The ray that travels from the radar to P and back again is shorter than the ray from the radar to O and back again by twice the distance OA . Since OA is just

$$x \sin \theta + z \cos \theta,$$

the phase $\phi(x)$ of the ray returning to the radar from P , relative to that from O , is

$$\phi(x) = -2k(x \sin \theta + z \cos \theta),$$

where k is the wavenumber of the radiation. The (complex) amplitude received in the direction θ is found by integrating over the whole surface:

$$a(\theta) = \int \exp(i\phi(x)) dx.$$

To simplify this, we will find the speckle pattern near $\theta = 0$, so that we may put $\sin \theta \approx \theta$ and $\cos \theta \approx 1$. Thus

$$a(\theta) = \int \exp(-2ik(x\theta + z)) dx. \quad (9.17)$$

We recognise this as the Fourier transform of the function $\exp(-2ikz)$, where z is a function of x .

The exact nature of the speckle pattern will depend on the properties of the function $z(x)$ (in two dimensions, the function $z(x, y)$), which will in general be defined only statistically. Even if the r.m.s. value of $2kz$ is much less than 1 (i.e. the surface is very smooth), any realistic function $z(x)$ will generate a function $a(\theta)$ that changes sign on a small angular scale. Thus the ‘expected’ image will be multiplied by a spatial intensity

variation whose statistical properties will depend on the nature of $z(x)$. We can describe the image statistics by Equation (9.18):

$$Q(j) = R(j)S(j). \quad (9.18)$$

Here, $Q(j)$ is the amplitude (or intensity) of pixel j in the image, $R(j)$ is the amplitude (or intensity) that it would have had in the absence of speckle, and $S(j)$ is the multiplicative speckle noise term, drawn at random from an appropriate probability distribution.

When the surface is rough (so that the r.m.s. value of $2kz$ is much greater than 1), Equation (9.17) shows that the signal in a given direction is found by adding together a large number of components each of which has the same amplitude but a phase drawn at random from a uniform distribution from 0 to 2π . In this case, which is called *fully developed speckle*, the probability distribution $p(S)$ for an intensity image is a negative exponential function:

$$p(S) = \exp(-S). \quad (9.19)$$

For an amplitude image, the corresponding form of $p(S)$ is a Rayleigh distribution:

$$p(S) = 2S \exp(-S^2). \quad (9.20)$$

Figure 9.18 shows an example of ‘pure’ speckle. It is a simulated amplitude image corresponding to the case where $R(j)$ in Equation (9.18) is constant, i.e. to a completely homogeneous region. The characteristic granularity can clearly be seen. Real, as opposed to simulated, speckle can be seen in Figures 9.12 and 9.23, for example.

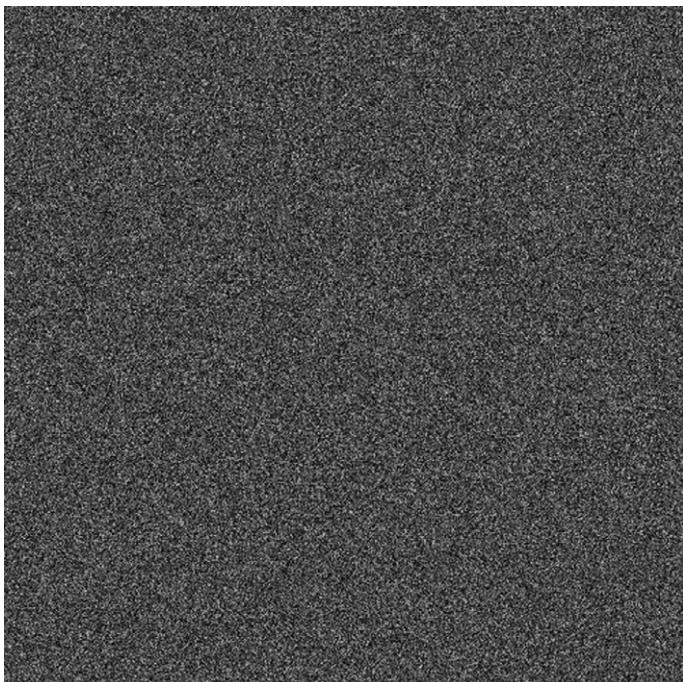


Figure 9.18. ‘Pure’ fully developed speckle in an amplitude image.

Image speckle is an undesirable consequence of the coherent imaging mechanism that is needed to obtain the high azimuth resolution of a SAR. The effects of speckle can be reduced by some form of spatial averaging of the data (e.g. Rees and Satchell 1997). Often this is performed at the stage of generating the radar image from the raw amplitude and phase data. This is referred to as a multi-look image, where the number of ‘looks’ refers to the number of samples that are combined incoherently to process a given pixel. A multi-look image will have reduced speckle at the expense of poorer spatial resolution.

9.5.3 Distortions of SAR images

SAR images are subject to the same geometrically induced distortions as SLR, discussed in Section 9.4.1. However, if the ‘target’ is moving, a further source of distortion is introduced. This arises from the complicated way in which the data are processed to generate the image, and can most easily be thought of in terms of the Doppler frequency analysis presented briefly in Section 9.5. We noted there that the Doppler shift of the signal received from a given scatterer is first positive, falling to zero at the instant when the radar has the same along-track coordinate as the scatterer, then becoming negative. If the scatterer is in motion, an extra Doppler shift will be added to that due to the platform motion. The total Doppler shift will thus fall to zero at a different value of the along-track coordinate, and the processor will assign this wrong value to the along-track coordinate of the scatterer.

We can model this phenomenon quite simply. Figure 9.19 shows a ‘target’ scatterer located at the origin of a Cartesian coordinate system, moving at speed u in a direction that makes an angle ψ with the x -axis. The radar moves parallel to the y -axis such that its position is (x, y, H) where x and H are constants, and y increases uniformly with time at the rate v . At the instant illustrated in the figure, the velocity of the target with respect to the radar is

$$\mathbf{v}' = (u \cos \psi, u \sin \psi - v, 0)$$

and the position of the target with respect to the radar is

$$\mathbf{r}' = -(x, y, H).$$

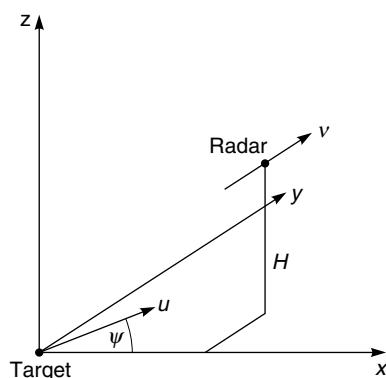


Figure 9.19. Geometry for discussing azimuth shift in a SAR image. The target scatterer is located at the origin with velocity coordinates $(u \cos \psi, u \sin \psi, 0)$; the radar is at (x, y, H) with velocity coordinates $(0, v, 0)$.

9.5 Synthetic aperture radar

The Doppler shift falls to zero when these two vectors are perpendicular (this follows from Equation 2.20), i.e. when $\mathbf{v}' \cdot \mathbf{r}' = 0$. This gives

$$y = \frac{ux \cos \psi}{v - u \sin \psi}. \quad (9.21)$$

This is the value of the y -coordinate (i.e. the along-track or azimuth coordinate) that will be assigned to the target, instead of its correct value of zero.

Equation (9.21) shows that the azimuth shift is zero when $\cos \psi = 0$, i.e. when the target is moving parallel or antiparallel to the radar track. If the target is moving perpendicularly to the radar track and towards it, $\cos \psi = 1$ and the displacement is in the direction in which the radar moves. Conversely, when the target moves away from the radar track, the displacement is opposite to the direction of the radar. These effects can be quite significant. For a spaceborne SAR, for which $v \gg u$, equation (9.21) can be approximated as $y \approx ux \cos \psi / v$. Taking $u = 10 \text{ m s}^{-1}$, $x = 300 \text{ km}$, $\cos \psi = 1$ and $v = 7 \text{ km s}^{-1}$ gives an azimuth shift y of over 400 m. Figure 9.20 shows an example of azimuth shift in a SAR image. In this case, the image of a moving ship is displaced from the image of its turbulent wake, which is stationary with respect to the sea surface. The phenomenon is valuable here, since it can be used to infer the presence of a moving ship and to estimate its velocity. Azimuth shift is less valuable in the case of imaging ocean waves, since it scrambles the image.

The azimuth shift will be fully developed only if the motion of the target is maintained throughout the coherence time of the SAR. If the period during which the target's motion can be considered steady is shorter than the coherence time, the image of the target will be blurred in the azimuth direction.

Other forms of motion-induced distortion and blurring can occur in SAR images. *Range walk* occurs when the target's range-direction coordinate changes by more than the range resolution during the time taken to acquire the image. This will obviously cause blurring in the range direction, but in fact also in the azimuth direction

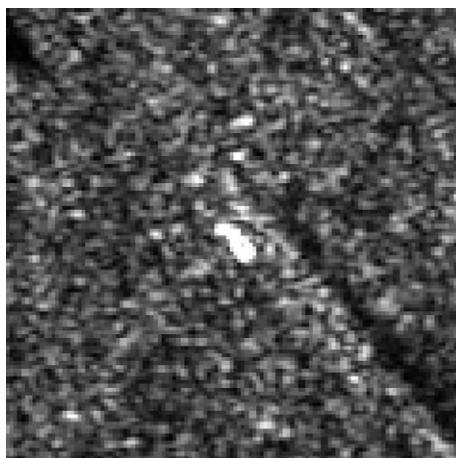


Figure 9.20. Azimuth shift in a SAR image. The image of the moving ship (the bright rectangular region at centre) is displaced from the image of its wake (the dark diagonal stripe). (Image reproduced from Lin *et al.* (1997) by courtesy of the European Space Agency. <http://earth.esa.int/workshops/ers97/papers/lini/>.)

(see Robinson 2010). For a platform velocity v , the time t taken to acquire an image with an azimuth resolution R_a is of the order of

$$t = \frac{S\lambda}{2R_a v}, \quad (9.22)$$

where λ is the wavelength and S the slant range to the target. The condition for range walk to occur is thus

$$|u_r| > \frac{2R_r R_a v}{S\lambda}, \quad (9.23)$$

where u_r is the range-component of the target's velocity and R_r is the range resolution. For a typical spaceborne system, one might have $R_a = R_r = 10$ m, $S = 350$ km, $\lambda = 6$ cm and $v \approx 7$ km s $^{-1}$, giving an upper limit on the range velocity (to avoid range walk) of about 70 m s $^{-1}$. At similar target velocities to those given by Equation (9.23), the phenomenon of *azimuth defocusing* can also occur. This happens when the rate of change of the Doppler shift of the returned signal is significantly different from the rate of change expected from a stationary target.

9.5.4 Limitations imposed by ambiguity

In Section 8.2 we introduced the concept of range ambiguity, and the desirability of avoiding it, for a pulsed system. In the case of a SAR system this produces some rather unexpected limitations on its performance.

Figure 9.21 shows a rear view (the radar is flying ‘into the page’) of a SAR imaging a swath of width W from a height H . For simplicity, we assume that H is sufficiently small that the Earth’s curvature can be ignored. The time taken for a round-trip journey by electromagnetic radiation from the radar to the near edge of the imaged swath and back again is just

$$\frac{2H}{c} \sec \theta_1,$$

and similarly for the far edge. To avoid ambiguity, the time interval between successive pulses must be greater than the difference between these two times, so the pulse repetition frequency p must satisfy

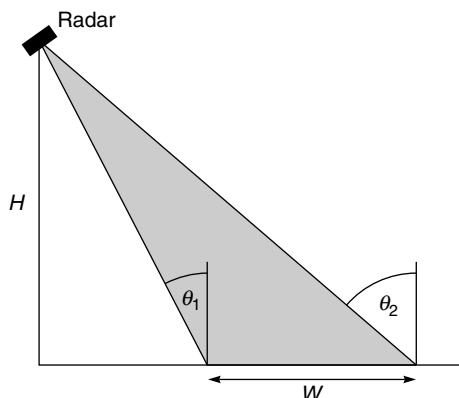


Figure 9.21. Relationship between the near-swath and far-swath incidence angles θ_1 and θ_2 , and the swath width W , for a flat-Earth geometry.

$$\frac{1}{p} > \frac{2H}{c} (\sec\theta_2 - \sec\theta_1). \quad (9.24)$$

On the other hand, a pulse repetition frequency that is too low will degrade the azimuth resolution of the system. If the platform velocity is v and the azimuth resolution is R_a , the condition that the azimuth resolution is not degraded by excessively slow sampling is

$$\frac{1}{p} < \frac{R_a}{v}. \quad (9.25)$$

Combining the two inequalities (9.24) and (9.25) gives

$$\frac{\sec\theta_1 - \sec\theta_2}{R_a} < \frac{c}{2Hv}. \quad (9.26)$$

Since $\sec\theta_2 - \sec\theta_1$ is related to the swath width W , we see that the swath width and the azimuth resolution cannot be varied independently of one another. To make this more definite, we suppose that $\theta_2 - \theta_1$ is small. In this case, the inequality (9.26) can be expressed as

$$\frac{W}{R_a} < \frac{c}{2v \sin\theta}. \quad (9.27)$$

Thus, for example, a spaceborne SAR system (with $v \approx 7 \text{ km s}^{-1}$) designed to operate at an incidence angle θ near 30° cannot have a swath width greater than approximately 4×10^4 times the azimuth resolution.

9.5.5 SAR interferometry

A SAR determines the across-track position of a target from the slant range, which depends on both the ground range coordinate and the height of the target. As we have already seen in discussing the geometrical distortions in SLR images (Section 9.5.1), the contributions from these two components cannot be disentangled from a single observation. However, if *two* SAR images are available, acquired from positions separated by a small distance, it is possible to separate the ground range and height effects, and hence to derive information about the surface topography. This is the idea of SAR interferometry (Gens and van Genderen 1996, Crosetto *et al.* 2011, Dong and Huang 2011), which shares some similarities with stereophotography (Section 5.3.2).

9.5.5.1 Geometry

The basic idea behind SAR interferometry can be understood from Figure 9.22. Two SAR images are acquired, from the positions M and S respectively (these denote ‘master’ and ‘slave’). We consider a particular target scatterer located at A , such that its slant range from M is r_M . If no more information than this is available, all we can say about the position of the scatterer is that it lies somewhere on BB , the locus of points that are distance r_M from M . Now we also consider the observation from S . The slant range to A is r_S , and the locus of points that are this distance from S is CC . In principle this is enough to distinguish between different positions, but we can see from Figure 9.22 that if M and S

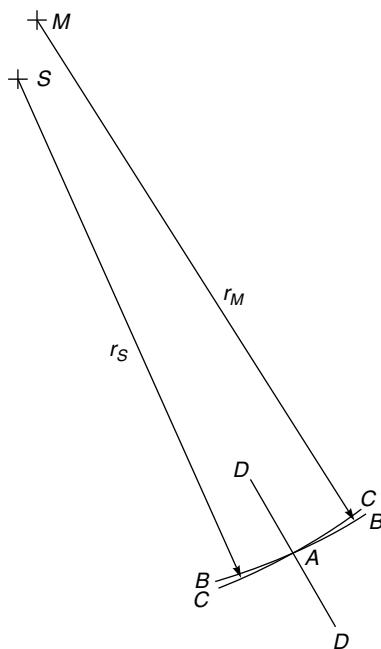


Figure 9.22. Essential geometry of SAR interferometry. Two SAR observations are made of the region near the target A , from positions M ('master') and S ('slave') respectively. M and S are very close to one another. BB is the locus of points that are the same distance r_M from M as A ; CC is the locus of points that are the same distance r_S from S as A ; and DD is the locus of points for which the difference between the distance to M and the distance to S is equal to $r_M - r_S$.

are very close together the loci BB and CC will be almost coincident. Once we take into account the finite thickness of BB and CC (as a result of the finite range resolution of the SAR), we can see that in practice the two curves are coincident for a significant part of their lengths, centred at A . However, if we now consider the locus DD of points for which the *difference* between the distance to M and the distance to S has the same value as for the point A , we see that it is *perpendicular* to the locus BB (or CC). Thus we can see that a measurement of the slant range from M , and the difference between the slant ranges from M and from S , will in general be enough to identify the location of a particular scatterer.

The slant-range difference is in practice determined by comparing the phases of the signal in the two images, which of course means that both the amplitudes and phases of the backscattered signal must be available (in which case the images are referred to as 'complex'). The phase of a single SAR image is not particularly useful, but comparing the phases of two images contains useful geometrical information. In general, we can write the complex amplitude detected from a given pixel in the 'master' image as

$$a_M = a_1 \exp(i\phi_1) \exp(2ikr_M),$$

where a_1 and ϕ_1 are real numbers denoting the backscattered amplitude, and k is the wavenumber of the radiation. The same pixel observed in the 'slave' image has

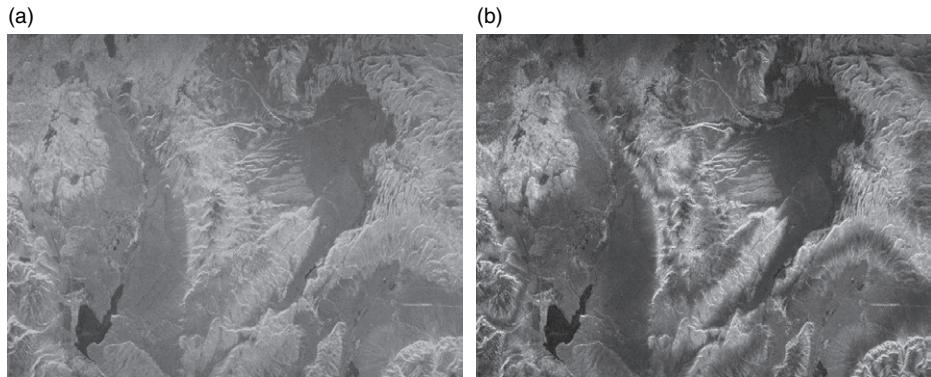


Figure 9.23. (a) One of a pair of SAR images used to construct an interferogram; (b) the corresponding interferogram. The images were acquired by the L-band SIR-C system carried on the Space Shuttle Endeavour, in April and October 1994, with a baseline of around 100 m. They show an area 59×34 km of Long Valley, California. The intensity of the interferogram is derived from the radar images, while the hue represents the phase. (Images downloaded from <http://www.archive.org/details/VE-IMG-626> and reproduced by courtesy of NASA Jet Propulsion Laboratory.) See also colour plates section.

$$a_S = a_2 \exp(i\phi_2) \exp(2ikr_S).$$

If the local observing geometry at the pixel, and its backscattering properties, are practically identical between the two observations, we will have

$$a_1 = a_2$$

and

$$\phi_1 = \phi_2,$$

which means that if we multiply one image by the complex conjugate of the other we will obtain

$$a_M a_S^* = |a_1|^2 \exp\left(2ik(r_M - r_S)\right).$$

This represents an *interferogram*, i.e. an intensity image with interference fringes superimposed (Figure 9.23). The fringes contain the information about how $r_M - r_S$ changes between the two images, with one complete fringe corresponding to a change of half a wavelength. Since the wavelength of a SAR system is typically only a few centimetres, we can see that the technique has the potential to achieve extremely high resolutions.

A convenient way of characterising the potential offered by a particular imaging geometry for the generation of interference fringes is to calculate the ambiguity distance e , as shown in Figure 9.24. This is defined by the point B , which has the same distance from M as the reference point A , but for which the difference between the distances to the M and S has changed by half a wavelength. Since this changes the difference in the round-trip distances by one wavelength, it means that the use of interferometry cannot distinguish between the positions A and B . Simple flat-Earth geometry shows that

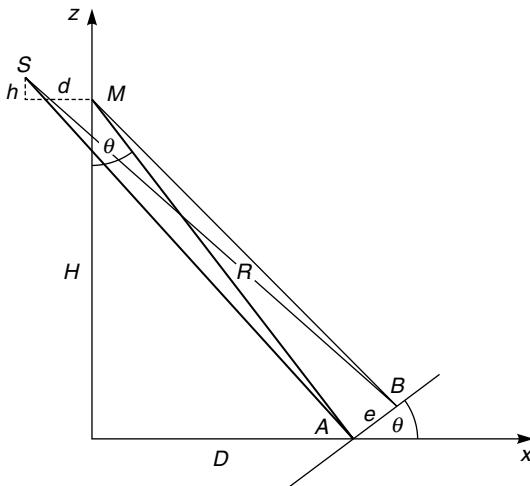


Figure 9.24. The ambiguity distance e is defined such that $SB - MB$ differs from $SA - MA$ by $\lambda/2$, where λ is the wavelength.

$$e = \frac{R^2 \lambda}{2(Hd - hD)}. \quad (9.28)$$

For example, if we take $H = 800$ km, $D = 350$ km (and hence $R = 873.2$ km), $h = 0$ and $d = 1$ km for a SAR with $\lambda = 6$ cm, we obtain $e \approx 30$ m. The ambiguity height, $e \tan \theta$, is about 11 m in this case, which means that if the signal-to-noise ratio in the complex image is high enough to detect, say, one-tenth of an interference fringe, altitudes can be determined from the interferogram with an accuracy of about 1 m.

Equation 9.28 shows that when $Hd = hD$, the ambiguity distance e is infinite. In this case no interference fringes are formed, and it arises when the target, master and slave are collinear. Values of e that are too small also pose a problem, since if e is much smaller than the slant-range resolution of the SAR system a single pixel will contain many fringes and it will be impossible to count the number of fringes. Equation (9.28) shows that small values of e arise from large values of the baseline (h or d) between the master and slave, so we can see that this baseline must not be too large. For typical spaceborne SAR systems this maximum baseline is of the order of 1 km.

The ambiguity of 2π in determining the phase difference between the two images imposes some practical difficulties in determining the geometry from the interferogram, since it is necessary to count the number of fringes to determine the true value of the phase. This problem, which is somewhat analogous to determining the correct labels to put on the contours of an unlabelled contour map, is termed *phase unwrapping* and a number of approaches have been developed to solve this problem. If more than two SAR images are available, giving more than one baseline, the problem is considerably easier to solve.

In general, SAR interferometry is a difficult technique which can yield excellent results when it works. Because it is very unlikely that one will have sufficiently accurate a-priori information about the locations from which the master and slave images were obtained, it is usually necessary to use a cross-correlation method on the images first, in order to determine which pixels in the slave image correspond to those in the master image. Only

if the correlation coefficient is high enough will further processing be worth undertaking. Similarly, the baseline vector (d, h) cannot usually be predicted and must be found from ground control points.

9.5.5.2 Acquiring suitable SAR images

The commonest method of obtaining two (or more) SAR images of the same location is to use the same instrument and to acquire the images at different times. In the context of spaceborne observations this is normally referred to as *repeat-orbit interferometry*. In this case the orbit must be such that the satellite repeats its path exactly at regular intervals and it is shown in Section 10.3.2.4 that this interval cannot normally be less than 3 days. Shorter repeats are possible in a *tandem* observation, in which two satellites are placed in the same orbit. This was first demonstrated in space in 1995, when the ERS-2 satellite was placed in the same orbit as the earlier ERS-1 satellite, following it at an interval of one day. Simultaneous observations are possible if the two satellites are very close in orbit. This is achieved by the TerraSAR-X/TanDEM-X mission (2010 onwards), in which these two satellites are only a few hundred metres apart in the same orbit.

If the SAR images that are used to construct an interferogram are not acquired simultaneously, differences in the ranges r_M and r_S to a given scatterer can be caused either by topographic relief, as described above, or by motion of the Earth's surface. (Differences in the corresponding phase differences can also be caused by different atmospheric and ionospheric propagation conditions at the times of acquisition.) SAR interferometry can thus be used to measure bulk translation of parts of the surface between images obtained at different times (for example, to monitor the motion of a glacier or the bulging of a volcano or earthquake zone). A single interferogram (i.e. acquired using just two SAR images) cannot distinguish between topography and motion as causes of phase difference, so further information is needed to resolve this ambiguity if it is believed to be present. This can take the form of an existing digital elevation model, or further interferograms with different baselines.

9.5.5.3 Coherence between images

The ability to form a useful interferogram between two SAR images depends on each image having a well-defined phase. This can be expressed in terms of the *coherence* between the images. The coherence γ can be defined quantitatively through Equation (9.29):

$$\gamma = \frac{\langle a_M \ a_S^* \rangle}{\sqrt{\langle |a_M|^2 \rangle \ \langle |a_S|^2 \rangle}}. \quad (9.29)$$

The angle brackets $\langle \rangle$ denote a spatial average over a small area of the interferogram. The coherence γ ranges between 0 and 1: a value of zero indicates that no useful information can be obtained from the interferogram. In general, the coherence is decreased by long baselines, by long times between the acquisition of the images, and by motion of the surface. The coherence itself can be a useful product: for example, vegetated areas show lower coherence than non-vegetated land surfaces, and the coherence observed over a water surface depends on the wind strength over the surface.

9.5.6

Major applications of radar imaging

In considering the applications of SLR and SAR images, it is helpful to compare them with the optical systems discussed in Chapter 6. SLR and SAR images are generally produced in a format similar to that of a black and white aerial photograph, the brightness of the displayed image being dependent on the value of σ^0 . Such images may be visually interpreted singly or as stereo pairs (although the side-looking geometry complicates the stereo effect). Even simple visual interpretation can often reveal important spatial relationships in the data, although increasing use is now made of digital imagery which can be processed and analysed using a computer, making the maximum use of the quantitative nature of the data.

The spatial resolution achievable by imaging radar systems is often comparable to that obtained from optical imaging systems. Since SLR and SAR are active remote sensing techniques, in which the illumination of the target is supplied by the instrument itself and not by solar radiation, they can be applied at night as well as during the daytime. Furthermore, since the propagation of microwave radiation is little affected by the presence of cloud or even of rainfall (unless it is very intense), radar images can be acquired during most weather conditions. These are obviously major advantages.

On the other hand, the physical processes that modulate the ‘brightness’ (i.e. the value of the backscattering coefficient) of a radar image are things for which most image analysts do not have a strong intuitive feeling. The relevant factors include the microwave dielectric constant of the surface material, its surface roughness and internal structure, as well as parameters of the observing system itself such as the frequency, polarisation and incidence angle. In some cases, the counter-intuitive nature of the relationship between a surface material and its backscattering coefficient can be quite strong. An example of this is a situation in which a medium that is optically thick at visible wavelengths is practically transparent to microwave radiation, such as dry snow. It is clear that in such cases *modelling* of the interaction between the microwave radiation and the target material will be especially important. Further difficulties are introduced by geometrical and topographic effects such as highlighting, layover and shadowing, and by speckle.

Despite these complications, imaging radar data are being applied to an increasingly wide range of tasks (Oliver and Quegan 2004). A number of these are suggested by Figure 9.25. Over land surfaces, applications include cartography (where the main task is to recognise and delineate natural and man-made features), the detection and characterisation of geomorphological lineaments, vegetation mapping (Kasischke *et al.* 2011) (including the identification and monitoring of different agricultural crops, change and damage detection, and the monitoring of soil moisture content), and highly ‘applied’ tasks such as monitoring flooding events (Martinez and Le Toan 2007) and (using radar interferometry) subsidence (Jiang *et al.* 2011) and landslides. SAR data have also been routinely applied to the task of generating digital elevation models (DEMs). The Shuttle Radar Topography Mission (SRTM) was flown from the Space Shuttle for 11 days during February 2000 (Farr *et al.* 2007). Two antennas, separated by a 60-m boom, were used to acquire images suitable for SAR interferometry, and the data have been processed to provide an almost complete DEM of the Earth’s land surface between latitudes 56° S and 60° N, with a sampling interval of 3 seconds of arc (approximately 90 m at the equator) (Figure 9.26). The data are freely accessible. Since 2010, the TanDEM-X mission has involved two orbiting SAR systems separated by 250–500 m which should provide DEM data with greater coverage, finer sampling and higher precision.



Figure 9.25. An L-band SAR image showing the area around Lake Biwa, Japan. The image was obtained in 2006 by the PALSAR instrument carried on the ALOS satellite. Note the features visible in the water, some of which are probably a manifestation on the surface of the bottom topography. Note also the many features of the land surface that are visible, including roads, rivers, canals, built-up areas (including the city of Kyoto, bottom left) and agricultural fields. Layover of the mountains is clearly visible. (Reproduced by permission of JAXA.)

Imaging radar also finds many applications in the marine environment. Surface wave fields can be imaged (Sun and Kawamura 2009) and their power spectra deduced, although with some difficulty as a result of the motion-dependent distortions discussed in Section 9.5.3. The azimuth shift phenomenon can be used positively to identify ship wakes and hence to monitor shipping (Wu and Wang 2008), for example in an area where fishing restrictions apply. Diffraction of waves by coastal features, and refraction by variations in bottom topography, are often clearly visible, and the latter phenomenon has been used as a bathymetric technique (Fan *et al.* 2011). There is also some evidence for the imaging of internal waves. Small-scale surface roughness is reduced by the presence of natural and artificial slicks (Solberg, Brekke and Husoy 2007), and these have been

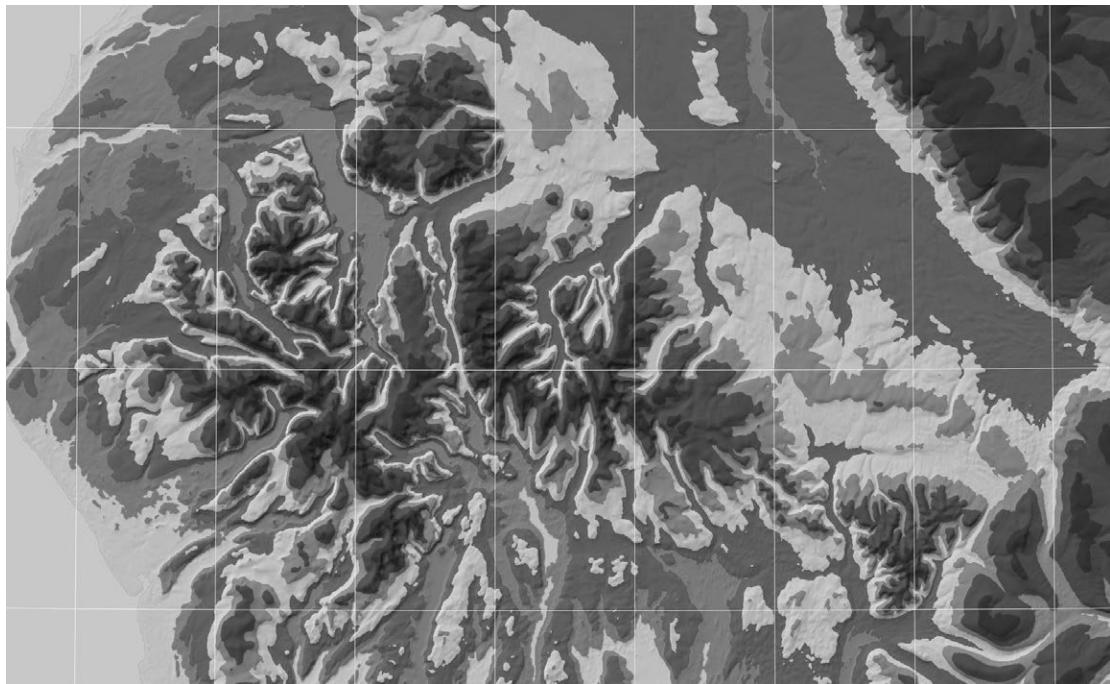


Figure 9.26. Visualisation of SRTM data of north-western England. The graticule has intervals of 10 minutes of latitude and longitude. See also colour plates section.

detected and monitored in radar imagery. The natural slicks can sometimes be indicators of unexploited oil reserves.

Imaging radar finds extensive applications in the delineation and monitoring of snow and ice (Høgda, Storvold and Lauknes 2010). The presence of a snow cover can be detected provided the snow is not completely dry (in which case the microwave penetration depth can be very large, thus rendering the snowpack effectively transparent). The boundaries of ice sheets and glaciers can be delineated, and different surface zones, corresponding to different thermodynamic regimes, can often be located. Year-to-year monitoring of these zones, and of the glacier's total area, can be used to infer its *mass balance*, i.e. whether it is growing, shrinking or in dynamic equilibrium. If the glacier or ice sheet calves icebergs, these can often be detected and tracked using radar imagery.

Imaging radar is also applied extensively to the study of sea ice. The delineation of boundaries between ice floes and open water is usually straightforward, and comparison of consecutive images of the same area of ocean allows the motion of ice floes to be tracked. A rather more difficult problem is the determination of ice type from radar images.

9.5.7

Example: Radarsat-2

We conclude this chapter by briefly describing a spaceborne SAR system, namely the instrument carried on the Radarsat-2 satellite. This is a Canadian satellite, launched in December 2007 into an orbit with an altitude of 800 km, with a mission entirely dedicated

Table 9.1. Observing modes of Radarsat-2 SAR

Mode	Incidence angle (°)	Swath width (km)	Resolution (m)		Looks		Polarisation
			range	azimuth	range	azimuth	
Fine	37–49	50	10	9	1	1	selective
Standard	20–49	100	25	28	1	4	
Low incidence	10–23	170	40	28			
High incidence	50–60	150	25	28			
Wide	20–45	100	25	28			
ScanSAR narrow	20–46	300	50	50	2	2	
ScanSAR wide	20–49	500	100	100	4	4	
Fine quad-polarisation	20–41	25	11	9	1	1	polarimetric
Standard quad-polarisation			25	28	1	4	
Ultrafine	30–40	20	3	3	1	1	selective
Multi-look fine	30–50	50	11	9	2	2	single

to SAR imaging. It is a successor to the earlier Radarsat mission (launched 1995). The SAR system operates at C band (5.4 GHz) in all polarisation modes (HH, VV, HV and VH) and can view either side of nadir. The SAR carried by Radarsat-1 had substantial flexibility in observing modes, and this principle was followed by Radarsat-2. Table 9.1 summarises the observing modes of Radarsat-2.

The column headed ‘polarisation’ indicates that the instrument can be operated in three different polarisation modes. In ‘selective single polarisation’, the transmitted signal can be either H or V and the received polarisation can be either H or V, while in ‘selective polarisation’ the received polarisation can in addition be both H and V states. In the ‘polarimetric’ mode the radar transmits H and V alternately, and receives both H and V.

Summary

Synthetic aperture radar (SAR) overcomes the difficulty of achieving a high along-track resolution from an SLR by synthesising the effect of a very long antenna from the motion of the platform carrying the antenna. This is equivalent to using the Doppler shift of the return signal to deduce the along-track coordinate of a scatterer in the antenna’s beam. The best possible along-track resolution achievable using this technique is half the length of the antenna, independent of the range from the antenna to the surface.

A SAR must necessarily use coherent radiation. One consequence of this is that the image is subject to *speckle*, a noise-like spatial variation in the image brightness. SAR images are subject to the same distortions as SLR images, and in addition, the images of moving scatterers are displaced relative to the positions they would occupy if they were stationary. The use of pulsed radiation to achieve high resolution in the cross-track

direction means that range ambiguity imposes some limitations on the performance of a SAR: one consequence is that the swath width and the along-track resolution cannot be varied independently.

Two or more SAR images can be combined in a technique called *SAR interferometry*, which measures differences in the geometry of the two (or more) image acquisitions. If the images are acquired at essentially the same time, but from different locations, the geometrical differences arise from the topography of the surface. However, if there is relative motion of the surface between the times at which the images are acquired this will also contribute to the geometrical differences. Both phenomena are useful. The ability to perform SAR interferometry using a pair of SAR images depends fundamentally on the coherence between the images. This is affected by a number of factors, including the baseline (separation between the two locations from which the images are obtained), time interval between the images, and the nature of the surface.

Radar images from SLR and SAR systems find a wide range of applications. In some senses they are comparable to VIR images or aerial photographs, often achieving similar spatial resolutions and with the added advantage of being able to be acquired at night and through cloud. On the other hand they are subject to a number of geometrical and radiometric distortions and are often not intuitive to interpret. Over land surfaces, SAR imagery is used cartographically, for studying vegetation and soils etc. SAR interferometry is extensively used for generating digital elevation models and for studying motion of the Earth's surface. In the marine environment, SAR is used to study surface wave fields, bathymetry, surface slicks, and to monitor shipping. SAR is also extensively used in terrestrial and marine glaciology.

Review questions

Outline the use of LiDAR systems for profiling atmospheric constituents.

Give a qualitative explanation of the form of the monostatic radar equation (Equation 9.4).

Explain what is meant by microwave scatterometry.

Discuss the use of microwave remote sensing methods for measuring wind speed over oceans.

Explain how range (cross-track) and azimuth (along-track) resolution is achieved by an SLR and by an SAR.

Describe the distortions present in side-looking radar imagery. What other types of distortion are present in SAR imagery?

Why does azimuth shift occur in SAR imagery? What are its advantages and disadvantages?

Explain the nature of speckle in SAR imagery. Why does it arise? How can it be reduced?

What limitations are imposed by range ambiguity on the SAR imaging process?

Give an outline of the use of SAR interferometry for determining topography. What determines its accuracy?

Explain how SAR interferometry can be used to measure the motion of the Earth's surface, for example changes of a few centimetres in elevation following an earthquake.

Problems

1. A radar transmits a power of 4 kW at a wavelength of 5 cm and over a range of 800 km. If the radar can detect a received signal of 10–16 W, calculate the required gain of the antenna, and hence estimate the area of the antenna, if it is to be able to detect values of σ^0 down to –25 dB from an area of 10^3 m^2 . Assume that the antenna has unit efficiency.
2. A very simplified model of the radar backscattering coefficient of a sea surface is

$$\sigma^0 = A + B\cos 2\psi$$

where ψ is the angle between the wind direction and the radar look azimuth, and A and B are as follows for K_u band scattering at 40° incidence angle, HH-polarisation:

$$A = 0.8v - 30,$$

$$B = 3.5 - 0.1v$$

v is the wind speed in ms^{-1} and σ^0 is given by these expressions in dB.

A scatterometer observation measures σ^0 values of –22.9 dB and –21.1 dB looking north and east respectively. Find the wind velocity. Is there any ambiguity in your answer? If so, could it be removed by a third observation?

3. Show that the amplitude of fully developed speckle has a Rayleigh distribution. Assume that the signal is composed of a very large number of components, each of which has the same amplitude but a phase randomly selected from the range 0 to 2π .
 4. A SAR operates at a frequency of 5.3 GHz. Its satellite platform is 775 km above the Earth's surface and has a speed relative to the surface of 6.73 km s^{-1} . The swath width is 100 km and the incidence angle at the centre of the swath is 23°. The along-track resolution is 30 m. Estimate (i) the coherence time of the radiation transmitted by the radar, and (ii) the difference in two-way travel time between pulses reaching the near edge and the far edge of the swath. Use your answer to (ii) to show that the along-track resolution and the swath width do not conflict in this case. [Assume the Earth to be flat for simplicity.]
 5. A SAR operates at a wavelength of 6 cm and is used for interferometric SAR observations. In position M , the radar has (x, y) coordinates $(-400\,000, +800\,000)$ m, while in position S its coordinates are $(-400\,100, +800\,000)$ m. A scatterer A located near the origin has coordinates (x, y) .
 - Show that the range from M to A is given approximately by $0.447\,2136x - 0.894\,427\,2y$ plus a constant, and that the difference between the range from S to A and the range from M to A is given approximately by $0.0000894x + 0.0000447y$ plus a constant.
 - Hence find the coordinates of A if its distance from M is the same as that from the origin, but the difference $SA - MA$ is greater than $SO - MO$ by half a wavelength.
 - If the system is capable of resolving 1/50 of a fringe, what is its effective height resolution?
- [Assume that $z = 0$ throughout this question.]

10

Platforms for remote sensing

In this chapter we consider aircraft and satellites as platforms for remote sensing. There are other, less commonly used, means of holding a sensor aloft, for examples towers, balloons, model aircraft and kites, but we do not discuss these. The reason for this, apart from their comparative infrequency of use, is that most remote sensing systems make direct or indirect use of the relative motion of the sensor and the target, and this is more easily controllable or predictable in the case of aircraft and spacecraft. Figure 10.1 shows schematically the range of platforms, and their corresponding altitudes above the Earth's surface.

The spatial and temporal scales of the phenomenon to be studied will influence the observing strategy to be employed, and this in turn will affect the choice of operational parameters in the case of an airborne observation or of the orbital parameters in the case of a spaceborne observation. After a brief introduction to the use of aircraft as platforms for remote sensing, this chapter focusses on the use of artificial satellites.

10.1

Aircraft

Aircraft of various types provide exceptionally convenient and operationally flexible platforms for remote sensing, carrying payloads ranging from a few tens of kilograms (considerably less in the case of UAVs) to many tonnes (Figure 10.2). With a suitable choice of vehicle a range of altitudes can be covered from a few tens of metres, where atmospheric propagation effects are generally negligible, to many thousands of metres, above most of the Earth's atmosphere. The choice of flying altitude will obviously have an impact on the scale, spatial coverage and spatial resolution of the data collected. It is also important in the case of pulsed systems such as laser profilers, where the question of range ambiguity arises. The range of available platform speeds is more or less continuous from zero (in the case of a hovering helicopter) to several hundred metres per second. It is particularly important to match the flying speed to the characteristics of a scanner or a SAR system, and flying speed also has an obvious role in determining the total area from which data can be collected. Flying routes and times can also be chosen with great flexibility, subject of course to any restrictions on the use of air space or those imposed by weather.

The main disadvantages of aircraft as platforms for remote sensing, when compared with spacecraft, are as follows: First, a typical airborne observing mission has a duration

10.1 Aircraft

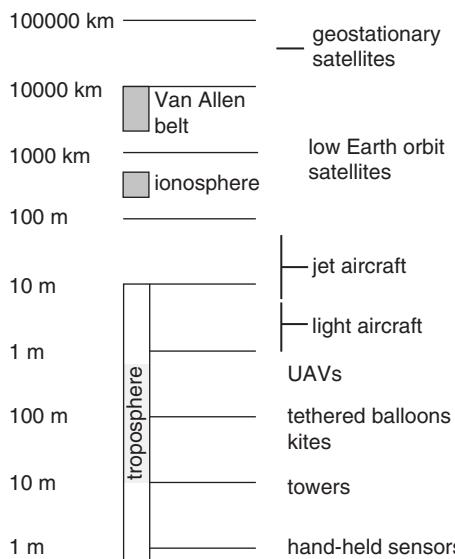


Figure 10.1. Remote sensing platforms arranged by typical altitude above the Earth's surface. The ionosphere (Section 4.5) and Van Allen belts (Section 10.3.2.4) are regions of high concentrations of charged particles, unfavourable for satellites. UAVs are unmanned aerial vehicles.

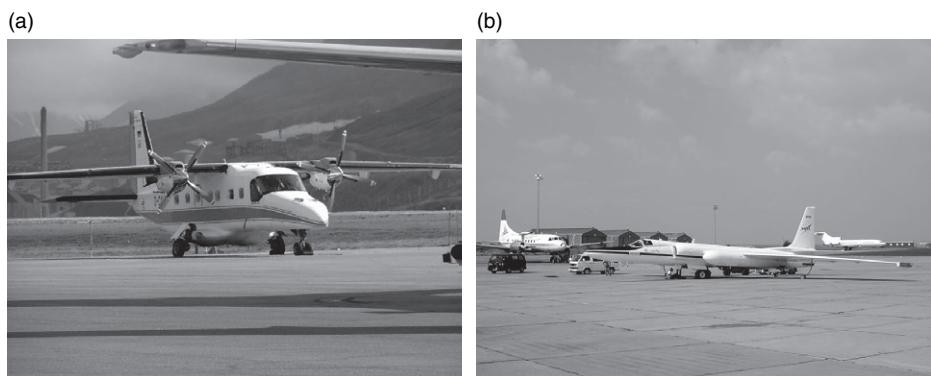


Figure 10.2. (a) Small turboprop aircraft used for remote sensing observations. This unpressurised plane has an operational ceiling of around 4600 m. (This is a Dornier 228 used by the UK Natural Environment Research Council's Airborne Research and Survey Facility: Author's photograph.) (b) Turbofan aircraft used for high altitude remote sensing observations. This plane (NASA-operated Lockheed ER-2) can fly high in the stratosphere, having an operational ceiling of above 21 km, and can carry a payload of over 1 tonne. (Photograph courtesy of NASA. <http://earthobservatory.nasa.gov/IOTD/view.php?id=771>.)

of only a few hours, as compared with a few years for a spaceborne mission. This means that it is much more difficult to provide continuity of data, for example for a 10-year monitoring programme. Second, since airborne observations are acquired from much lower altitudes than spaceborne observations, the spatial coverage of the data is smaller

and airborne observations are obviously unsuitable for studying very large areas. (On the other hand, of course, they are much more suitable than spaceborne observations for detailed investigations of smaller areas.) Finally, since aircraft necessarily operate within the Earth's atmosphere, and the atmosphere is in motion, neither the position nor the motion of the aircraft may be exactly what was intended. (These are problems for spacecraft too, just less severe.)

An aircraft's true position can be determined in two ways, often in combination. The recent enormous improvements in radiolocation methods, notably the Global Positioning System (GPS), mean that it is now easily possible to determine the position of an aircraft, with an accuracy of the order of one decimetre or better, at the time. The second approach is through the use of ground control points (GCPs). These are features whose exact locations on the ground are known and which can also be observed and precisely located in the image collected by the airborne sensor. They can thus be used to correct the position, orientation and scale of the image, and also to correct any distortions that may be present. This procedure is discussed in Section 11.2.1.2. Suitable GCPs can be 'naturally' occurring – for example, coastal features or road intersections – or can be emplaced especially for the purpose. In the latter case, suitable GCPs might be reflective markers fixed to the ground or, for active microwave sensors, passive or active radar transponders.

In addition to uncertainty in its position, an aircraft may also be subject to uncertainty or variation in its motion. The most important of these are roll, pitch and yaw, which are oscillations about three mutually perpendicular axes: roll is oscillation about the longitudinal axis, pitch about an axis parallel to the wings, and yaw about an axis that is nominally vertical. These motions are particularly inconvenient in the case of scanner imagery. Pitch and yaw can lead to over- or under-sampling in the along-track direction, while roll changes the imaged swath (Figure 10.3). Many scanners for airborne use inertial sensors which provide at least partial correction for these motions, although the use of GCPs may also be necessary.

Summary

Aircraft provide convenient and operationally flexible platforms for remote sensing, enabling instruments to be supported at a range of heights above the Earth's surface from essentially zero to 20 km or more, and with a range of speeds relative to the Earth's surface from zero to several hundred metres per second. Large aircraft can carry payloads of many tonnes. Irregularities in the motion of an aircraft may require corrections to the position and geometry of images acquired from them.

10.2

Satellites

Placing a satellite in orbit about the Earth is clearly more expensive than mounting an airborne remote sensing campaign, but the advantages, in terms of the increased platform speed and potential swath width as well as continuity of observations, are substantial. In general, the spatial data coverage from a satellite mission is better than that obtainable

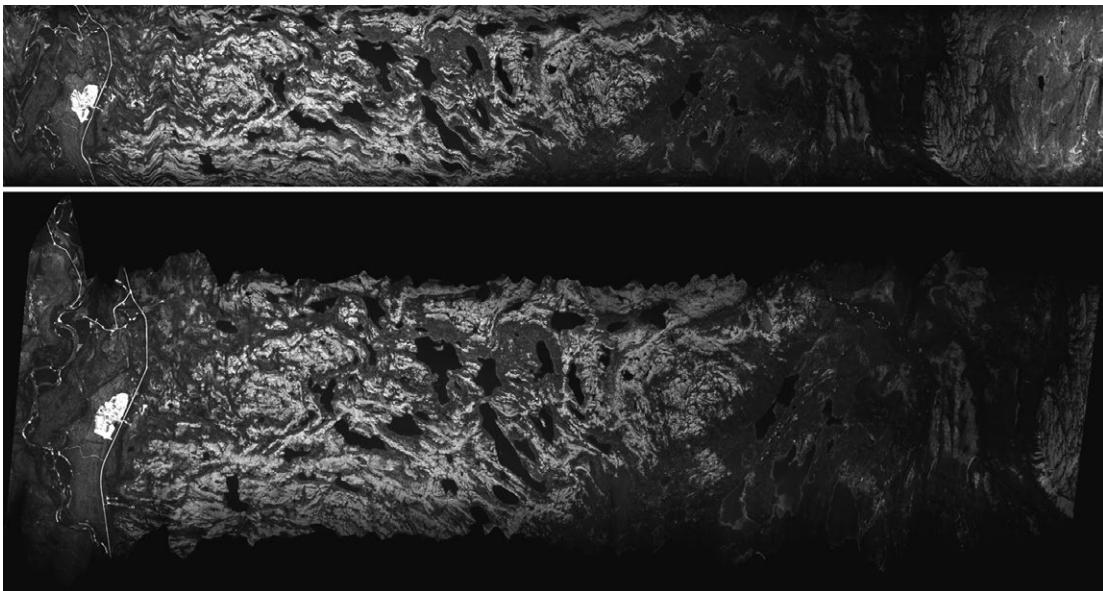


Figure 10.3. True-colour representation of airborne scanner imagery before (above) and after (below) correction for the effects of roll, pitch and yaw. The aircraft was flying from left to right, as can be seen from the fact that the initial roll was subsequently corrected by the pilot. (Airborne Thematic Mapper image of Porsangmoen, Norway, collected by UK Natural Environment Research Council's Airborne Research and Survey Facility in 2004.) See also colour plates section.

from an airborne mission, and the fact that a spaceborne sensor may continue to function for three years or more (sometimes much more) means that temporally homogeneous datasets can be obtained.

The cost-effectiveness of satellite remote sensing was discussed briefly in Chapter 1, where it was suggested that the economic benefits of the data generated by an operational satellite (as opposed to one launched purely for the purpose of research) normally more than justify the cost of launching and operating it. An obvious advantage of using satellites for remote sensing, but one which poses interesting legal, security and moral questions, is the fact that the laws of orbital dynamics do not respect political boundaries. The development of very high resolution observing systems has intensified some of these questions in recent years (Purdy 1999, Harris 2009).

10.2.1

Launch of satellites

This section provides a brief introduction to the considerations that apply to the placing of a satellite in orbit about the Earth. A more detailed treatment is given by Maini and Agrawal (2010).

To place a satellite in a stable (apart from the effects of atmospheric friction and the solar wind) orbit, it is necessary to overcome the Earth's gravitational attraction and, to a lesser extent, the resistance of the lower atmosphere. This is achieved using a *rocket*,

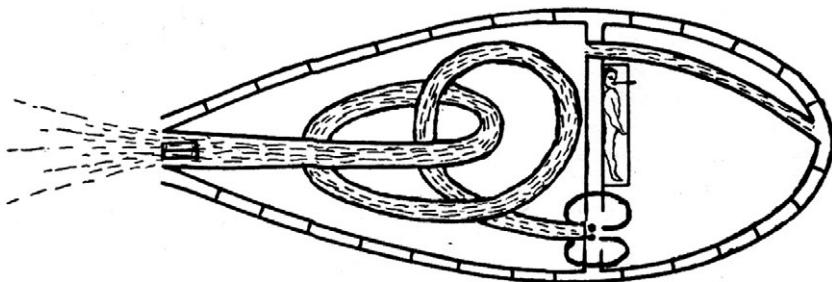


Схема „ракеты“.

Figure 10.4. Sketch showing the principle of rocket propulsion, from Konstantin Tsiolkovsky's paper of 1903.



Figure 10.5. Launch of a Dnepr rocket from Baikonur Cosmodrome, Kazakhstan. The Dnepr is a converted intercontinental ballistic missile and it is launched from an underground silo. It has been used to launch more than a dozen remote sensing satellites. (Photograph from <http://www.kosmotras.ru> by courtesy of ISC Kosmotras.) See also colour plates section.

which is a vehicle that carries all its own fuel, including the oxidising agent, deriving a forward thrust from the expulsion backwards of the combustion products (Figures 10.4 and 10.5). Elementary classical mechanics shows that a rocket of total initial mass M_i burning a mass M_f of fuel will increase its speed, in the absence of gravitational and friction forces, by

$$\Delta v = u \ln \frac{M_i}{M_i - M_f}, \quad (10.1)$$

where u is the speed of the exhaust gases relative to the rocket. This is usually called the Tsiolkovsky equation, after Konstantin Tsiolkovsky who published its

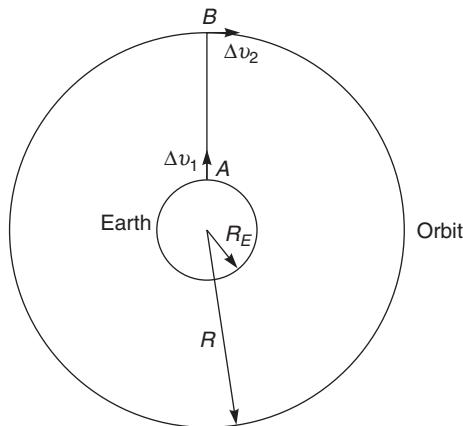


Figure 10.6. Schematic illustration of one method of placing a satellite in orbit around the Earth.

derivation in 1903 (Tsiolkovsky 1903), although an earlier derivation is due to William Moore in 1813 (Johnson 1995).

Figure 10.6 illustrates a very simple model of the insertion of a satellite into a circular orbit of radius R around the Earth. The Earth is assumed to be a non-rotating uniform sphere of mass M and radius R_E . The first part of the process is to launch the rocket vertically upwards (i.e. radially away from the Earth's surface), with an essentially instantaneous ‘burn’ of the rocket motor at A . This gives the rocket a speed Δv_1 . The rocket then travels ‘ballistically’, i.e. under the influence of gravity alone (we ignore air resistance), until it comes to rest at B . It is straightforward to show that the velocity increment Δv_1 needed to achieve this is given by

$$\Delta v_1^2 = 2GM \left(\frac{1}{R_E} - \frac{1}{R} \right), \quad (10.2)$$

where G is the gravitational constant. For example, if we take $R_E = 6400$ km and $R = 7200$ km (corresponding to an orbital altitude of 800 km), Equation (10.2) gives $\Delta v_1 = 3.7$ km s $^{-1}$.

The final part of the process is to give the satellite the velocity necessary for a circular orbit of radius R . Since the rocket starts from rest at B , the required velocity increment Δv_2 is just the orbital velocity and hence is given by

$$\Delta v_2^2 = \frac{GM}{R}. \quad (10.3)$$

For the example of $R = 7200$ km this gives $\Delta v_2 = 7.5$ km s $^{-1}$, so the total velocity increment needed for this type of launch is 11.2 km s $^{-1}$.

Figure 10.7 illustrates a more efficient procedure. In this case, the rocket is launched tangentially to the Earth’s surface at A , with a velocity Δv_1 sufficient to put it into an elliptical *transfer orbit* that just grazes the desired orbit at B . At B , a further velocity increment Δv_2 is applied to inject the satellite into its circular orbit. It can be shown that the required velocity increments in this case are

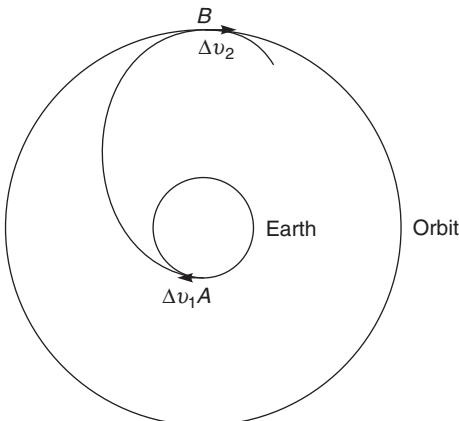


Figure 10.7. Injection of a satellite into a circular orbit using an elliptical transfer orbit.

$$\Delta v_1 = \sqrt{\frac{2RGM}{R_E(R + R_E)}} \quad (10.4)$$

and

$$\Delta v_2 = \sqrt{\frac{GM}{R}} \sqrt{\frac{2R_E GM}{R(R + R_E)}}. \quad (10.5)$$

For our example of $R = 7200$ km, Δv_1 is 8.2 km s^{-1} and Δv_2 is 0.2 km s^{-1} , giving a total velocity increment of 8.4 km s^{-1} . However, if the satellite revolves in its orbit in the same sense as the Earth rotates (this is called a *prograde* orbit – see Section 10.3.2), some of the required velocity is imparted by the Earth's rotation. Since the tangential speed of the Earth's surface is roughly 0.5 km s^{-1} , this can reduce the total velocity increment to about 7.9 km s^{-1} .

Since a typical value of u (the velocity of the exhaust relative to the rocket) is about 2.4 km s^{-1} , and we have seen that the rocket needs to accelerate the satellite to a speed of at least (roughly) 8 km s^{-1} , we can estimate from Equation (10.1) that a rocket capable of inserting the satellite into our chosen orbit must initially consist of at least 96% fuel. Naturally the payload (the satellite) can account for only a fraction of the remaining 4%, reduced still further when the effect of atmospheric friction is taken into account. For this reason, single-stage rockets are capable of placing only rather small masses into orbit. Instead, multiple-stage rockets, with three or four stages, are used, and these are capable of putting payloads of a few tonnes into low Earth orbit (LEO) and somewhat smaller masses into geostationary orbit. For example, the Dnepr rocket shown in Figure 10.4 has a mass at launch of 211 tonnes and can place a payload of up to 4.5 tonnes in LEO. The heavy launch vehicle Ariane-5 has a mass at launch of 777 tonnes and can place 21 tonnes in LEO or around 7 tonnes in geostationary orbit.

In fact the model of the transfer orbit that we have just outlined is somewhat simplified. The rocket is in practice launched not horizontally but vertically, although its subsequent trajectory brings it into a horizontal flight a couple of hundred kilometres above the Earth's surface.

Summary

Artificial satellites are placed in orbit around the Earth using rockets. The typical orbital speed of a satellite in a low Earth orbit (LEO) is around 7 km s^{-1} , although the rocket has to impart a greater total change of speed to the satellite in order to complete the various manoeuvres necessary to insert it into the desired orbit. A rocket capable of placing a satellite in LEO will have a mass at launch of typically 40 times the mass of the satellite; for a geostationary orbit the ratio is around 100.

10.3

Description of the satellite orbit

If the Earth were a spherically symmetric mass with no atmosphere, and there were no perturbing influences on the motion of a satellite from the Sun, Moon, other planets, solar wind etc., the motion of a small satellite would in general follow an elliptical path with the Earth's centre as one focus of this ellipse. It will be useful to begin our investigation of satellite orbits by considering this idealised case, and Figure 10.8 illustrates some of the terms we need to describe it.

The points P (the *perigee*) and A (the *apogee*) are respectively the nearest and furthest points of the orbit from the Earth's centre E . (In general, these points in an orbit are called the *apsides*, which is the plural of the Greek word *apsis*.) The line AP is called the major axis, and its perpendicular bisector is the minor axis. The terms semimajor and semiminor axes are normally used to refer to half the lengths of these axes, i.e. a and b respectively. The *eccentricity* e of the ellipse is the ratio of the lengths CE to CA ; it is also related to a and b through

$$b^2 = a^2(1 - e^2). \quad (10.6)$$

According to Newton's law of gravitation, the period of the orbital motion (i.e. the time interval between successive passes through the same point in the orbit) is given by

$$P_0 = 2\pi\sqrt{\frac{a^3}{GM}} \quad (10.7)$$

where, as before, G is the gravitational constant and M is the Earth's mass. Although neither G nor M has yet been measured particularly accurately, their product GM , called

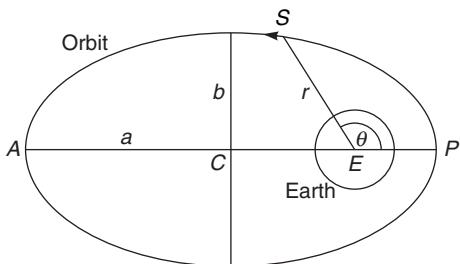


Figure 10.8. Description of a satellite S in orbit about the Earth.

the geocentric gravitational constant, has been measured to very high precision by observing the orbits of artificial satellites about the Earth. A recent value (Ries *et al.* 1992) is

$$GM = (3.986004415 \pm 0.000000008) \times 10^{14} \text{ m}^3 \text{ s}^{-2}.$$

The position of the satellite S in the orbital plane can be specified by the angle θ that it has rotated from the perigee, and its distance r from the Earth's centre E . These variables are related by

$$r = \frac{a(1 - e^2)}{1 + e \cos\theta}, \quad (10.8)$$

which tells us the shape of the ellipse but not the rate at which the satellite travels along it. The position of the satellite in the orbit, specified by the angle θ (between $-\pi$ and $+\pi$), is related to the time t since it passed through the perigee by the following equation:

$$\frac{2\pi t}{P_0} = 2 \arctan \left[\frac{(1 - e)\tan(\theta/2)}{(1 - e^2)^{1/2}} \right] - \frac{e(1 - e^2)^{1/2}\sin\theta}{(1 + e\cos\theta)}. \quad (10.9)$$

Equation (10.9) is actually rather inconvenient, since it gives the time in terms of the position, and we more commonly want to know the position in terms of the time (Figure 10.9). Unfortunately the equation is not analytically invertible, in other words it cannot be rewritten as a closed-form expression for θ in terms of t . It can however be inverted into a series expansion that is useful for small values of the eccentricity e , as shown in Equation (10.10):

$$\theta = \frac{2\pi t}{P_0} + 2e \sin\left(\frac{2\pi t}{P_0}\right) + \frac{5e^2}{4} \sin\left(\frac{4\pi t}{P_0}\right) + \dots \quad (10.10)$$

Use of the first three terms of this expansion, as shown, results in a maximum error in θ of approximately $4e^3/3$ radians.

In fact, most artificial satellites used for remote sensing are placed in orbits that are nominally circular and in practice have very small eccentricities, typically less than 0.01, and we may continue for the time being to develop our description of orbital motion on the assumption that $e = 0$. In this case, Equation (10.9) or (10.10) shows that θ increases uniformly with time. The orbit must be concentric with the Earth (which we are still assuming to be spherically symmetric) but need not be parallel to the equator. The angle between the plane of the orbit and the plane of the equator is called the *inclination* of the orbit (see Figure 10.10). The inclination i is by convention always positive, and is less than 90° if the orbit is *prograde* (i.e. in the same sense as the Earth's rotation about its axis), greater than 90° if the orbit is *retrograde*. An exactly *polar* orbit, in which the satellite passes directly over the Earth's poles, has an inclination of 90° . Near-polar orbits give the greatest coverage of the Earth's surface, and are widely used for low orbit meteorological and other remote sensing satellites. It is however in general more expensive to inject a satellite into a near-polar orbit than into a prograde orbit, because of the advantage that can be taken of the Earth's rotation during the launch phase in the latter case.



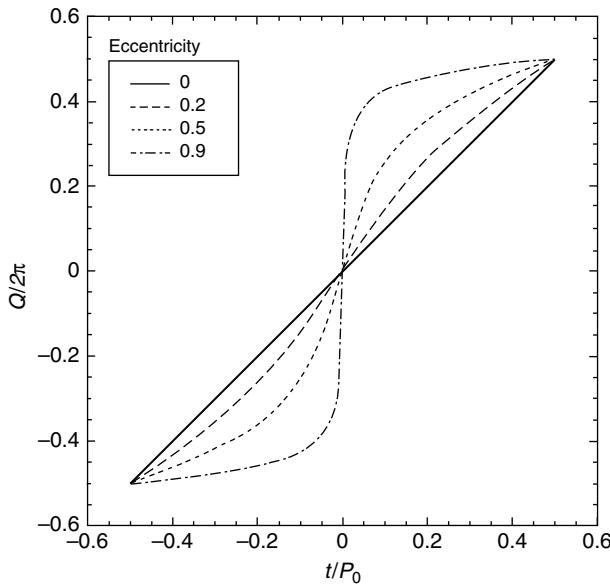


Figure 10.9. Dependence of angle in orbit on time in orbit for different values of the eccentricity e . Note the higher angular velocities near perigee for large values of e .

The position of the satellite in space is in general specified by six variables, although there are various equivalent sets of variables. For example, we may specify the direction of Ω , the values of a , e and i , the instantaneous value of φ and the value of φ at perigee. These variables are usually called the *Kepler elements* of the orbit. For a circular orbit, the number of useful variables is four, since the eccentricity is zero and the orbit has no perigee.

In many cases, we want to define the position of the *sub-satellite point*, i.e. the point on the Earth's surface that is directly below the satellite. If we continue to assume that the Earth is spherical, the latitude b and longitude l of this point can be calculated using spherical trigonometry. (The coordinates are defined so that north latitudes and east longitudes are positive.) The point Ω in Figure 10.10 is the *ascending node* of the orbit, and is the point where the satellite crosses the equatorial plane from south to north. If the angle $SE\Omega$ is φ , as shown in Figure 10.10, and the instantaneous longitude of Ω is l_0 , the position of the sub-satellite point is given by

$$\sin b = \sin \phi \sin i, \quad (10.11.1)$$

$$l = l_0 + \text{atan2}\left(\frac{\tan b}{\tan i}, \frac{\cos \phi}{\cos b}\right). \quad (10.11.2)$$

The atan2 function

Atan2 is a convenient trigonometrical function that avoids the ambiguity involved in taking inverse sines, cosines and tangents. It is nearly always defined such that $\text{atan2}(y, x)$ is the

angle whose cosine is $x/\sqrt{x^2 + y^2}$ and whose sine is $y/\sqrt{x^2 + y^2}$, as shown in the diagram. Some popular spreadsheet programs reverse the order of the arguments.

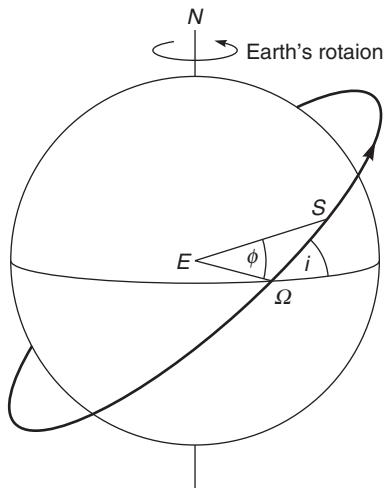
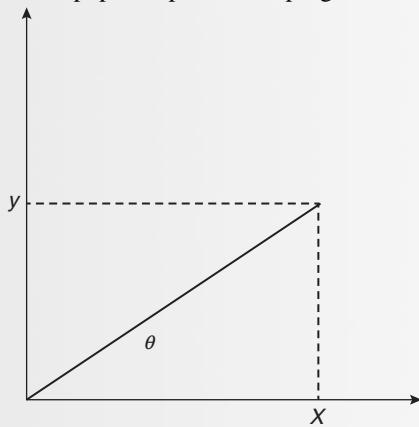


Figure 10.10. Circular satellite orbit in relation to the Earth's surface. E is the centre of the Earth, N the North Pole, S the instantaneous position of the satellite, and Ω is the ascending node. i is the inclination of the orbit and ϕ is the angle $SE\Omega$.

Even if the plane of the satellite's orbit were fixed in space (it is not, in general, as we shall see in the next section), the sub-satellite track would not describe a great circle. This is because of the Earth's rotation. When the satellite has completed one orbit, the Earth will have turned to the east and so the orbit will appear to drift to the west. This is true of both prograde and retrograde orbits, and of course also polar orbits. The rotation of the Earth may be taken into account in Equation (10.11.2) by noting that it is equivalent to a uniform rate of change of l_0 . Typical satellite tracks are illustrated in Figure 10.11. This figure clearly shows the westward drift of the orbits. It also shows the fact, which can be deduced from Equation (10.11.1), that a prograde orbit of inclination i reaches maximum north and south latitudes of i , and a retrograde orbit of inclination i reaches maximum latitudes of $180^\circ - i$.

10.3 Description of the satellite orbit

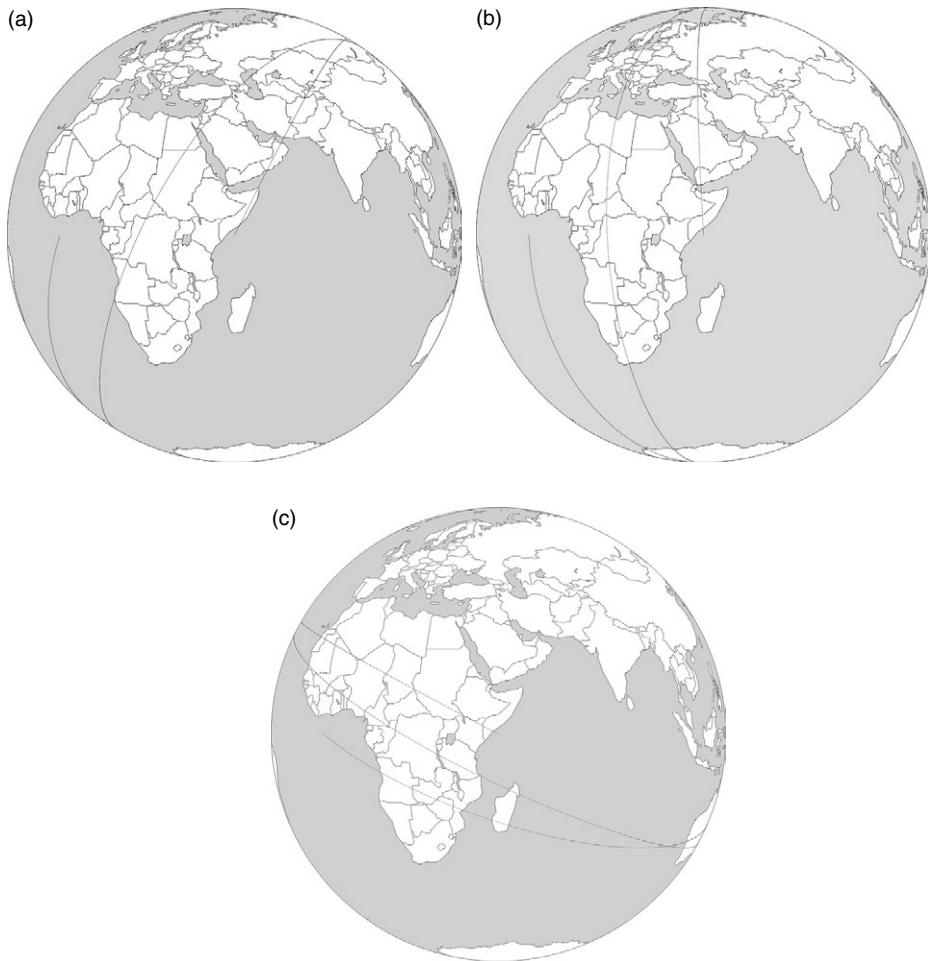


Figure 10.11. Typical satellite tracks for circular orbits of inclination 60, 89 and 150 degrees. All the tracks begin at the equator, travelling into the northern hemisphere. The period of the orbits is 100 minutes, and two complete orbits (200 minutes) have been plotted.

10.3.1 Effects of the Earth's asphericity

Thus far, we have assumed the Earth's mass to be distributed with spherical symmetry. It is not; roughly speaking, it is an oblate spheroid (i.e. the equator bulges outwards) as we noted in Section 8.2.5. The most convenient way to describe mathematically the effect of this non-spherical Earth on the motion of a satellite is to write the gravitational potential as a (possibly infinite) sum of spherical harmonics. As we might expect, the longitudinal variations are quite small when compared with the latitudinal variations, and in fact we normally need only to consider the latitudinal terms. In this case, the gravitational potential V per unit mass at latitude b and distance r from the Earth's centre can be written as

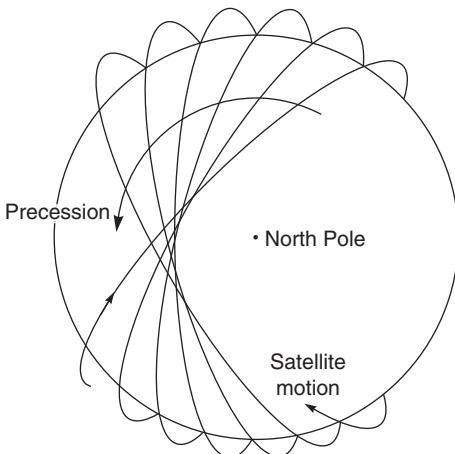


Figure 10.12. Precession of a satellite orbit around the Earth's polar axis, viewed from a fixed location in space looking down onto the North Pole. The rate of precession has been greatly exaggerated.

$$V = -\frac{GM}{r} \left(1 - \frac{a_e^2 J_2}{2r^2} (3\sin^2 b - 1) + \dots \right), \quad (10.12)$$

where a_e is the Earth's equatorial radius, $a_e \approx 6\,378\,135$ m. The dimensionless term J_2 , often called the *dynamical form factor*, expresses the equatorial bulge and has a value of $J_2 \approx 0.001\,082\,6$. It has three important effects on the orbit of a satellite. First, it increases the orbital period relative to the value P_0 given by Equation (10.7). The nodal period (i.e. the time between successive ascending or descending nodes) is given by

$$P_n = 2\pi \sqrt{\frac{a^3}{GM}} \left(1 + \frac{3J_2 a_e^2}{4a^2} \left\{ 1 - 3\cos^2 i + \frac{1 - 5\cos^2 i}{(1 - e^2)^2} \right\} \right) \quad (10.13)$$

and the rate at which even a circular orbit is described is no longer exactly uniform.

Second, it causes the orbital plane to rotate (precess) about the Earth's polar axis, so that the plane is not fixed in space. This is illustrated in Figure 10.12. The precession occurs at an angular velocity of

$$\Omega_p = \frac{3J_2 \sqrt{GM a_e^2} \cos i}{2a^{7/2} (1 - e^2)^2} \quad (10.14)$$

where a positive value indicates prograde precession. Because of the negative sign in Equation (10.14), prograde precession can only occur when $\cos i$ is negative, i.e. when the orbit itself is retrograde.

Finally, if the orbit is elliptical, the Earth's asphericity will cause the ellipse to rotate (precess) in its own plane, as illustrated in Figure 10.13. The angular velocity of this precession is given by



10.3 Description of the satellite orbit

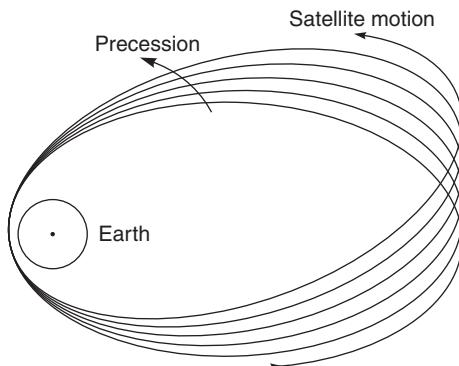


Figure 10.13. Precession of an elliptical orbit in its own plane (exaggerated for clarity).

$$\omega_p = \frac{3J_2\sqrt{GM}a^2(1 - 5\cos^2 i)}{4a^{7/2}(1 - e^2)^2}. \quad (10.15)$$

As before, a positive sign indicates prograde precession.

10.3.2 Special orbits

The dependence of the motion of a satellite on several parameters introduces the possibility of ‘tuning’ these parameters to give orbits with especially useful characteristics. In this section we discuss the most important of these special orbits.

10.3.2.1 Geostationary orbits

A satellite in a geostationary orbit is, as its name suggests, at rest with respect to the rotating Earth. This is achieved by putting the satellite into a circular orbit above the equator, with a nodal period P_n equal to the Earth’s rotational period P_E . The period P_E is not equal to 24 hours. This is because, in a period of 24 hours, the Earth does indeed rotate once with respect to the Sun, but since it is also orbiting the Sun in the same direction as it rotates on its axis, it has in fact rotated by slightly more than one complete turn with respect to a fixed reference system. The Earth takes approximately 365.24 days to orbit the Sun, so in 24 hours it makes $1+1/365.24$ complete turns. Thus we can see that P_E must be about 23.9345 hours or 86 164 seconds. This is called a *sidereal day* from the Latin *sidus*, a star, because it is the time the Earth takes to rotate once with respect to the stars.

The requirements for a geostationary orbit are therefore that its inclination and eccentricity should be zero, and its nodal period should be 86 164 seconds. The value of the semimajor axis a that satisfies these requirements in Equation (10.13) is about 42 170 km (about 6.6 times the Earth’s radius), so geostationary satellites are located approximately 35 800 km above the equator. Such orbits are used for geostationary meteorological satellites such as GOES and METEOSAT, as well as for telephone and television relay satellites.

Figure 10.14 shows a typical view of the Earth from a geostationary satellite. The part of the Earth’s surface that is visible from a geostationary satellite is a small circle centred on the sub-satellite point and having a radius of just over 81° , but in practice the useful



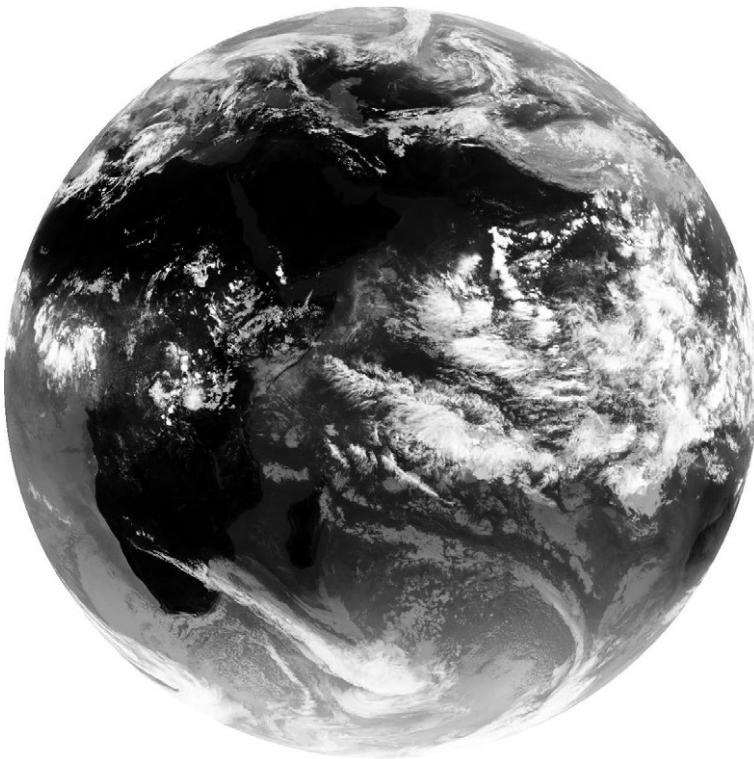


Figure 10.14. View of the Earth from a geostationary satellite. The image was recorded by a METEOSAT satellite located at longitude 57° and above the equator, and it shows thermal infrared radiation with an inverted colour table (higher radiances are shown in darker shades of grey). Image recorded at 1200 UTC on 27 September 2010. (Courtesy of Natural Environment Research Council.) **PERMISSION NEEDED FROM DUNDEE.**

coverage for quantitative analysis is assumed to be within a radius of 55° , and 65° for qualitative analysis. Figure 10.14 shows clearly that, while the coverage of latitudes reasonably close to the equator is excellent, the imaging of higher latitudes is impaired by the oblique viewing angle.

It would be convenient, but of course impossible, to place a satellite in a geostationary orbit above a point that is not located on the equator. If a satellite is placed in a circular orbit that has a nodal period of one sidereal day and a non-zero inclination i , the sub-satellite path traces a *lemniscate*, or figure-of-eight pattern, crossing at a fixed point on the equator and reaching maximum north and south latitudes of i , as shown in Figure 10.15. This can be called a *geosynchronous orbit*. The sub-satellite track is traced at an approximately uniform rate, so that for roughly half a day the satellite is above the ‘wrong’ hemisphere. Consequently, such orbits are not used in remote sensing (except, of course, that a real geostationary satellite is unlikely to have an inclination that is exactly equal to zero, so that in practice the sub-satellite point will make small figure-of-eight oscillations). An ingenious partial solution to this problem is the *Molniya orbit*, discussed in the next section.



Figure 10.15. Sub-satellite track for a geosynchronous satellite with an inclination of 63.4 degrees. The lemniscate or figure-of-eight pattern is described in one sidereal day.

10.3.2.2 Molniya orbits

The Molniya (молния: Russian for ‘lightning’) orbit provides a partial solution to the problem of placing a satellite above a fixed point on the Earth’s surface that is not on the equator. The orbit is highly eccentric, with the apogee (furthest distance) positioned above the desired point. Since a satellite’s angular velocity is much smaller when it is far from the Earth than when it is near to it (this follows from Equation 10.9), it can, with a suitable choice of orbital parameters, be arranged to spend much longer ‘on station’ than in the wrong hemisphere.

It is clear that a Molniya orbit must not rotate in its own plane, otherwise the position of the apogee would change over time. From Equation (10.15) we see that this requires that the inclination i must satisfy the equation

$$1 - 5\cos^2 i = 0,$$

so $i = 63.4^\circ$ or 116.6° . In practice, therefore, the latitude of the apogee must be fixed at 63.4° north or south. The nodal period is chosen to be *half* a sidereal day, so from Equation (10.13) we see that the semimajor axis is about 26 560 km. (If the nodal period were one sidereal day, the semimajor axis would be of the order of 42 000 km, similar to that for a geostationary orbit, and the large eccentricity of the orbit would result in an unhelpfully large apogee distance.) The eccentricity is chosen to give a minimum altitude above the Earth’s surface of the order of 500 km. For example, if we set $e = 0.740$, the satellite will have a perigee distance of about 6900 km, giving a minimum altitude of about 500 km, and an apogee distance of about 46 200 km. Figure 10.16 shows the sub-satellite track of an orbit with these parameters. It can be seen that the satellite spends a large proportion of its time near the perigee point; for practical purposes, it can be assumed to be ‘on station’ for eight hours, so that three satellites in the same orbit can provide continuous coverage.

Molniya orbits are used for telephone relay satellites by the former Soviet Union. They have also been used for some Soviet spy satellite missions.





Figure 10.16. Sub-satellite track of a satellite in a Molniya orbit with an inclination of 63.4 degrees and an eccentricity of 0.74. It repeats itself at intervals of 24 sidereal hours; after 12 sidereal hours, the track advances exactly 180 degrees in longitude. The subsatellite point is within 100 km of the point at maximum latitude for 3 hours, and within 1000 km for more than 8 hours, in each 12-hour period.

10.3.2.3 Orbits for global navigation satellite systems

A global navigation satellite system (GNSS) is a system of satellites broadcasting very precise radio timing signals that can be used to determine the location of a suitable receiver. While a GNSS is not itself a remote sensing system, it has become an essential component of many aspects of remote sensing. The best known GNSS is the Global Positioning System (GPS) operated by the US Department of Defence. This system uses a constellation of 24 Navstar satellites arranged in six uniformly spaced orbital planes. The inclinations of the orbits are 55° and the orbital periods are exactly half a sidereal day, so that the sub-satellite track repeats itself each day (Figure 10.17). This is not an essential requirement of the GPS system but was convenient during its development in the 1980s. The semimajor axis of its orbit is around 26 600 km. The Russian GLONASS GNSS, and other systems currently in development, use similar though not exactly synchronous orbits.

10.3.2.4 Low Earth orbits

Placing a remote sensing satellite in low Earth orbit (LEO) has the obvious advantage of improving the potential spatial resolution with respect to what can be achieved from a geostationary orbit, at the expense of reduced coverage. Such orbits are very widely used in spaceborne remote sensing.

The useful range of orbital altitudes for LEO satellites is constrained by the Earth's atmosphere and by the Van Allen belts. If a satellite orbits too low above the Earth's surface, it will experience an unacceptably high degree of atmospheric friction and will spiral out of its orbit and towards the Earth (see Section 10.4). Except for short satellite missions and very high resolution military reconnaissance satellites, the minimum useful altitude is effectively about 350 km.

10.3 Description of the satellite orbit



Figure 10.17. Sub-satellite track of a Navstar satellite.

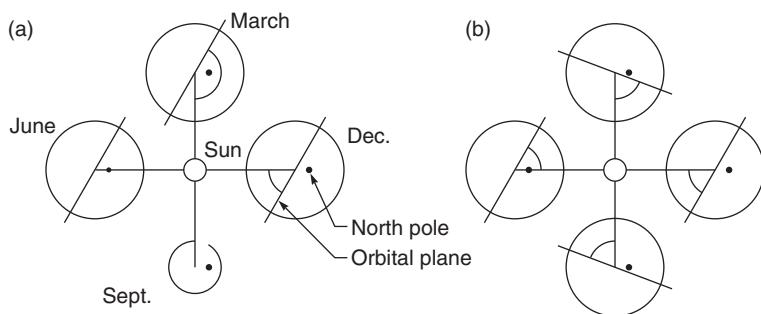


Figure 10.18. Schematic illustrations of (a) a non-precessing orbit and (b) a sun-synchronous orbit. In each case the view (not to scale) is normal to the plane in which the Earth rotates around the Sun.

The *Van Allen belts* are rings of high energy charged particles (mostly electrons and protons) in the Earth's equatorial plane, at altitudes between about 2000 and 5000 km and between 13 000 and 19 000 km. They have probably been 'captured' from the solar wind by the Earth's magnetic field. They represent a particularly difficult environment in which to operate a remote sensing satellite, so in practice there is an upper limit of about 2000 km to the altitude of a LEO satellite.

Careful choice of the orbital parameters of a satellite in LEO can give a range of useful properties. One particularly important type of LEO orbit is the *sun-synchronous orbit*. Such an orbit has the property that it precesses about the Earth's polar axis at a rate (one revolution per year) that matches the Earth's average angular speed around the Sun. This is illustrated in Figure 10.18.

The figure shows that the angle between the normal to the orbital plane and the line joining the centres of the Earth and the Sun is constant for a sun-synchronous orbit, whereas it increases at the rate of 360 degrees per year for a non-precessing orbit.

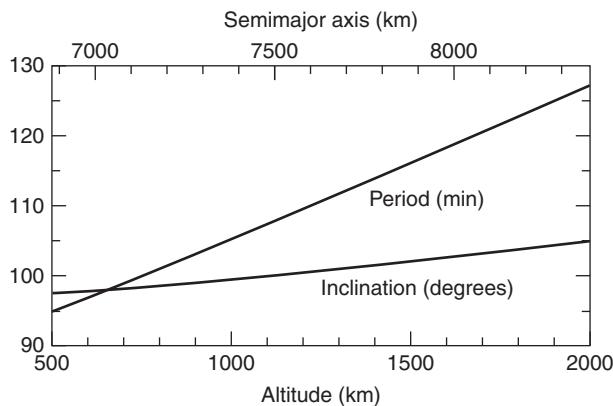


Figure 10.19. Inclination and nodal period for circular sun-synchronous orbits.

The mean angular speed Ω_s at which the Earth orbits the Sun is 2π radians per year or equivalently

$$\Omega_s = 1.991 \times 10^{-7} \text{ s}^{-1}.$$

Substituting $\Omega_p = \Omega_s$ into Equation (10.14), and setting $e = 0$ for simplicity, we can thus find the relationship between the inclination i and the semimajor axis a for a sun-synchronous orbit. This is plotted in Figure 10.19, along with the corresponding nodal period calculated from Equation (10.13).

For example, the figure shows that a satellite in a circular sun-synchronous orbit with an altitude of 800 km will have inclination of about 99° and a nodal period of about 101 minutes. The fact that sun-synchronous LEOs have inclinations close to 90° , and are hence near-polar orbits, is convenient since it means that a satellite in such an orbit has the potential to view a large fraction (all of it if the swath width of the instrument is wide enough) of the Earth's surface.

The most useful consequence of putting a satellite in a sun-synchronous orbit is that it will cross the same latitude, in a given sense (i.e. northbound or southbound), at the same local solar time, regardless of the longitude or the date. This can be seen by considering the northbound crossing of the equator, i.e. the ascending node, although the argument can be generalised to any latitude. Let us suppose that the satellite has a nodal period P_n and that it passes through the ascending node, at longitude zero, at time zero. This is the *Universal Time* (UT), which is equivalent to the Greenwich Mean Time to sufficient accuracy for the purpose of this discussion. At $UT = P_n$ the satellite has made one orbit, so it is again at its ascending node. During this time, the Earth has rotated through an angle (expressed in radians) of $P_n \Omega_E$ to the east, where Ω_E is the Earth's angular velocity of rotation, given by

$$\Omega_E = \frac{2\pi}{P_E}$$

and P_E is one sidereal day. However, the satellite orbit has precessed through an angle $P_n \Omega_s$ to the east, and thus the longitude of this ascending node is, in radians,

10.3 Description of the satellite orbit

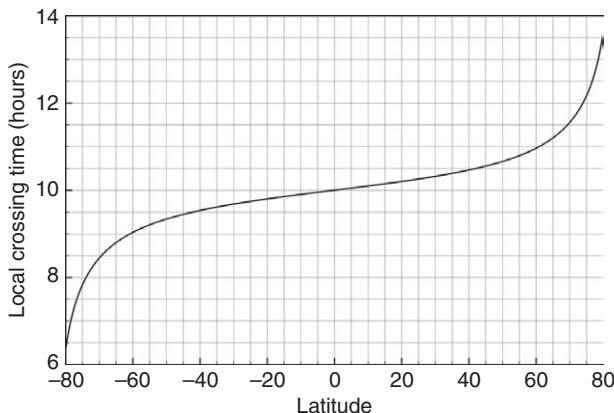


Figure 10.20. Local crossing time as a function of latitude for the descending pass of the Landsat-7 orbit, assuming an equatorial crossing time at the descending node of 10:00.

$$-P_n(\Omega_E - \Omega_s),$$

where the minus sign indicates a longitude that is west of the Greenwich meridian. Now although the UT is P_n , the local solar time must be corrected for the longitude. The correction is 24 hours per 2π radians of longitude, so we can write the local solar time as

$$P_n - P_n(\Omega_E - \Omega_s) \frac{P'_E}{2\pi}$$

where P'_E is one solar day of 24 hours. Finally, we observe from the definition of the sidereal day given in Section 10.3.2.1 that

$$\Omega_E - \Omega_s = \frac{2\pi}{P'_E} \quad (10.16)$$

so the local solar time at the new ascending node is zero, and our demonstration is complete.

The fact that a sun-synchronous orbit crosses a given latitude at the same solar time is particularly useful for satellites carrying passive optical or infrared sensors, and is widely used. For example, the *Landsat-7* satellite is in a sun-synchronous orbit with a descending node at a local solar time of 10:00, i.e. in the late morning. Thus, depending on the time of year, data can be collected from most of the descending (southbound) pass of the satellite while it is above the sunlit side of the Earth. This is illustrated in Figure 10.20, which shows how the local solar time varies with the latitude for the descending pass.

Exactly repeating orbits

It is often convenient to arrange that the sub-satellite track will form a closed curve on the Earth's surface, in other words that it will repeat itself exactly after a certain interval of time. This allows images having the same viewing geometry to be acquired on many occasions during the satellite's lifetime, and makes available a particularly simple method of referring to the location of images, for example the 'path and row' system used for the Landsat World Reference System (WRS).

In order for the sub-satellite track to repeat itself, it is clear that the Earth must make an integral number of rotations, say n_1 , in the time required for the satellite to make an integral number of orbits, say n_2 . However, we must also allow for the precession of the satellite orbit. Thus we can write the condition for an exactly repeating orbit as

$$P_n(\Omega_E - \Omega_p) = 2\pi \frac{n_1}{n_2}. \quad (10.17)$$

Provided that the fraction n_1/n_2 is expressed in its lowest terms (i.e. that n_1 and n_2 have no common factors), this can be described as an n_1 -day repeating orbit. If the orbit is also sun-synchronous, substituting Equation (10.16) into Equation (10.17) yields the simpler condition

$$\frac{P_n}{P'_E} = \frac{n_1}{n_2}, \quad (10.18)$$

where P'_E is one solar day of 24 hours.

It is desirable that n_1 should be as small as possible, since this determines the time interval between successive opportunities to observe a given location. On the other hand, n_2 governs the density of the sub-satellite tracks on the Earth's surface since there are n_2 ascending and n_2 descending passes. Thus we require a large value of n_2 to give a dense coverage. Because both the satellite's nodal period P_n and precession rate Ω_p depend on its semimajor axis a and inclination i (we are assuming that the orbit is circular), there is some scope for 'fine-tuning' of the ratio n_1/n_2 by adjusting these orbital parameters. However, the ratio is mainly governed by the nodal period, and as Figure 10.19 shows, this is generally between about 90 and 120 minutes for satellites in LEO. Thus in practice the ratio is constrained to roughly 0.07 ± 0.01 . For example, an orbit suitable for observing highly dynamic phenomena, for which a revisit interval of no more than one day would be acceptable, would have a maximum value of n_2 of only about 16. This would give rather coarse spatial sampling since spatially adjacent sub-satellite tracks would be about 2500 km apart. On the other hand, if we required that the spacing of the sub-satellite tracks should be at most 100 km (for example because we wish to obtain complete coverage of the Earth's surface using a narrow-swath sensor), n_2 would have to be at least 400, implying that successive revisits of a given location could not occur more frequently than once in 24 days.

Figure 10.21 shows the pattern of sub-satellite tracks for the orbit of Landsat-7. This is an exactly repeating sun-synchronous orbit with $n_1 = 16$ and $n_2 = 233$.

Spatially adjacent tracks are separated by $360/233 = 1.55$ degrees of longitude. It can also be noted that spatially adjacent tracks are followed at intervals of 7 days, in the sense that the sub-satellite track moves 1.55 degrees westward every 7 days. This can be described as a seven-day subcycle. If the swath width of the sensor is sufficiently broad, or the sensor's direction of view is steerable, a given location can be observed more than once per repeat cycle. The average time interval between these opportunities will be equal to the period of the subcycle, and this can provide a partial solution to the conflict between the requirements for frequent observation and for a dense network of sub-satellite tracks. Rees (1992) provides a detailed discussion of this technique.

10.3 Description of the satellite orbit

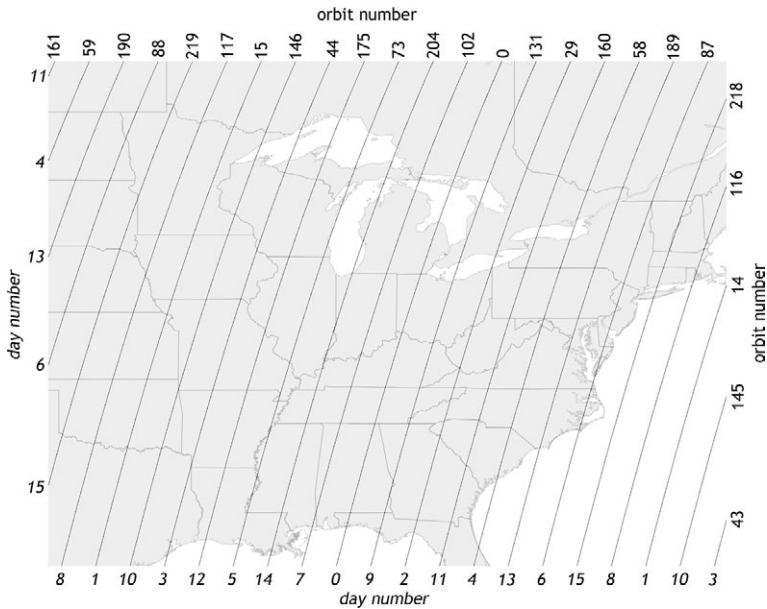


Figure 10.21. Repeat pattern for an orbit with $n_1 = 16$ and $n_2 = 233$. The figure shows part of the track of the Landsat-7 satellite. Only the descending (southbound) tracks are shown. Orbit numbers are shown in roman font, numbered from an arbitrary zero, while day numbers are shown in italic font.

Altimetric orbits

The last kind of special orbit that we consider is an orbit suitable for spaceborne altimetry. This has two desirable features. The first is that the ascending and descending sub-satellite tracks should intersect at roughly 90 degrees on the Earth's surface, so that orthogonal components of the surface slope can be determined with equal accuracy. The second feature is important when considering altimetry of the ocean surface. As discussed below, the existence of ocean tides has implications for the temporal frequency with which measurements should be made.

In order to address the first requirement, we calculate the orientation of the sub-satellite track on the Earth's surface. We assume the Earth to be spherically symmetric with radius r . For a circular orbit of inclination i , the latitude b of the sub-satellite point is given by Equation (10.11.1) as

$$\sin b = \sin \phi \sin i, \quad (10.11.1)$$

where ϕ is the angle defined in Figure 10.9. Differentiating this expression with respect to ϕ gives

$$\frac{db}{d\phi} = \frac{\sin i}{\cos b} \sqrt{1 - \frac{\sin^2 b}{\sin^2 i}} \quad (10.19)$$

for the ascending (northbound) track, so the northward component of the velocity of the sub-satellite point is given by

$$v_N = \frac{2\pi r}{P_n} \frac{db}{d\phi} = \frac{2\pi r \sin i}{P_n \cos b} \sqrt{1 - \frac{\sin^2 b}{\sin^2 i}}, \quad (10.20)$$

where P_n is the satellite's nodal period.

The simplest way to find the eastward component v_E of the velocity is first to consider the case of a non-precessing orbit around a non-rotating Earth. In this case, we can just apply Pythagoras's theorem to the components v_E and v_N to obtain

$$\pm \frac{2\pi r}{P_n} = \sqrt{1 - \left(\frac{db}{d\phi}\right)^2}$$

for v_E , where a positive sign indicates a prograde orbit and a negative sign a retrograde orbit. Finally, including the effect of the Earth's rotation at Ω_E and the orbital precession at Ω_p gives the result

$$v_E = -(\Omega_E - \Omega_p)r \cos b \pm \frac{2\pi r}{P_n} = \sqrt{1 - \left(\frac{db}{d\phi}\right)^2}, \quad (10.21)$$

where, as before, the sign is chosen according to whether the orbit is prograde or retrograde.

Equations (10.19) to (10.21) define the velocity of the sub-satellite point, and hence its direction, relative to the Earth's surface. They have been derived only for the ascending track of the orbit; for the descending track the sign of v_N is reversed. Knowing the components of the velocity of the sub-satellite point we can easily calculate the direction. This is shown in Figure 10.22, where the azimuth of the direction (the bearing measured clockwise from north) is plotted as a function of the latitude for different values of the inclination. The figure has been calculated for an orbit having a semimajor axis a of 7178 km, but will be very similar for other values of a corresponding to low Earth orbits. It is clear that when the azimuth is equal to $+45^\circ$ or -45° the ascending and descending paths will cross at 90° , and hence be optimum for altimetry.

The second criterion for altimetric orbits, of particular relevance to studies of the ocean surface topography, is the choice of repeat interval, i.e. the frequency at which a given location is observed. If the altitude of this location does not change then the problem does not arise. However, the Earth's ocean and, to a lesser extent, its solid surface are subject to tidal action which causes them to vary in height. These variations have strong periodicities; for example, what are usually the largest components of the deep-water tides have periods of 12.421, 23.934, 12.000 and 25.819 hours.

Suppose we try to observe the ocean surface with an altimeter that repeats its observation of a given location at intervals of exactly 3 days. Since exactly six cycles of the 12-hour tide will elapse between each measurement, the same point in the cycle will be measured on each occasion and we will not in fact observe any variation at all as a result of this tidal component. If our goal is to measure the mean ocean topography, rather than its mean instantaneous configuration at some point in the tidal cycle, we will fail to achieve it.

10.3 Description of the satellite orbit

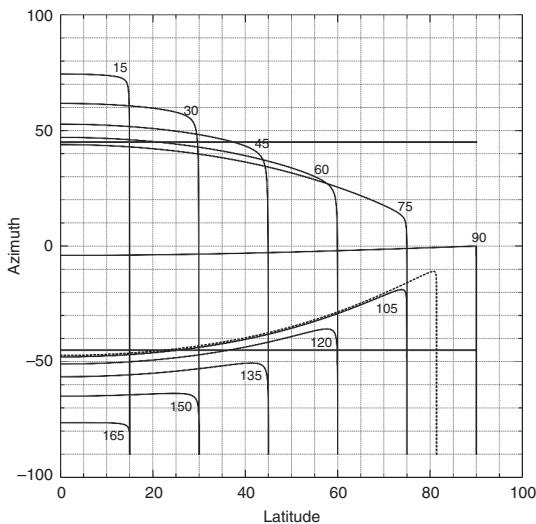


Figure 10.22. Direction of motion of the sub-satellite point for the ascending part of a satellite in a circular low Earth orbit. The solid curves are labelled with the value of the inclination, while the dotted curve corresponds to a sun-synchronous orbit. The horizontal grey lines show azimuths of $\pm 45^\circ$, where the ascending and descending paths cross perpendicularly.

Now let us suppose that we try to observe the 12-hour tidal component by increasing the sampling interval very slightly to 3.1 days. After one sample interval the 12-hour cycle will have occurred 6.2 times, so our measurement point will have advanced in effect by only 0.2 cycles. Thus at least five such samples, requiring 15.5 days, will be necessary in order to observe this component of the tide at all phases of its cycle (Figure 10.23). This is the phenomenon of *aliasing*, which arises when a periodic phenomenon of frequency f_1 is observed by means of repeated measurements at frequency f_0 . The frequency of the phenomenon is reduced to an apparent frequency f_a in the range $-f_0/2$ to $+f_0/2$ (the aliased frequency) which differs from the true frequency f_1 by an integral multiple of f_0 . This can be expressed mathematically as

$$f_a = f_1 - f_0 \left[\frac{f_1}{f_0} + \frac{1}{2} \right], \quad (10.22)$$

where $[x]$ is a function of x defined as the largest integer not exceeding x . Returning to the first of our two examples above, we have $f_1 = 2 \text{ day}^{-1}$ and $f_0 = 1/3 \text{ day}^{-1}$, so Equation (10.22) shows that the aliased frequency is zero. In the second example, f_0 is decreased to $0.322\,58 \text{ day}^{-1}$, which gives an aliased frequency of $0.064\,52 \text{ day}^{-1}$ and hence a period of 15.5 days. If an altimetric orbit is intended to sample all phases of a particular tide, it is clear that the observing frequency must not cause the tidal frequency to be aliased to too low a value. It is also clear that any exactly repeating orbit will alias the 12-hour tidal component to a frequency of zero.

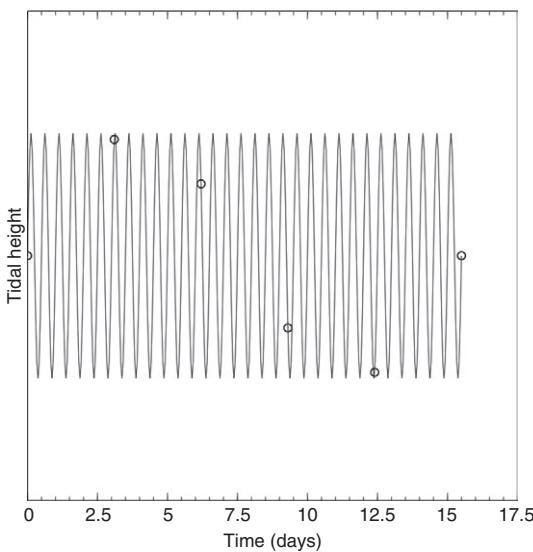


Figure 10.23. Aliasing of a tidal component. The solid line shows the variation of tidal height, which has a period of exactly 12 hours. The circles show the tidal height observed at intervals of 3.1 days. The tide is first observed at the same phase that it had at time zero after 15.5 days.

Summary

The path of a satellite's orbit around the Earth's centre of mass is an ellipse, with the Earth's centre of mass at one focus. The shape and size of the ellipse are defined by its semimajor axis a and its eccentricity e , or two equivalent variables. The period of the satellite's motion around the Earth depends only on a and not on e (if the Earth is assumed to be a sphere), and if $e=0$ (corresponding to a circular orbit) the satellite's speed is constant. If e is non-zero the speed is not constant, being greatest at perigee (the point at which the satellite is closest to the Earth's centre) and least at apogee (the farthest point).

The orientation of the orbit with respect to the plane of the Earth's equator is specified by the inclination i , a positive angle that is defined to be less than 90° if the orbit is prograde, i.e. the satellite rotates around the Earth's axis in the same sense as the Earth itself rotates (i.e. from west to east), and greater than 90° in the case of a retrograde (east to west) orbit. A true polar orbit has an inclination of 90° . Three more variables are needed to define the position of the satellite relative to the Earth's surface. One way of defining these three variables is to specify the angle φ subtended at the Earth's centre between the satellite and the perigee, the longitude of the ascending node (the point where the satellite crosses the plane of the equator in a northerly direction) and the value of φ at the ascending node. φ increases with time as the satellite travels around its orbit, and the longitude of the ascending node changes as the Earth rotates. The sub-satellite track describes a path on the Earth's surface between north and south latitudes of i (in the case of a prograde orbit) or $180^\circ - i$ (for a retrograde orbit).

The fact that the Earth is not spherically symmetric causes three important modifications to the simple description given above. First, the orbital period depends slightly on e and i as well as on a . Second, the plane of the orbit precesses around the Earth's axis, thus altering the rate of change of the longitude of the ascending node. Third, the orbit precesses in its own plane, if it is elliptical. The first two effects are useful and allow fine-tuning of orbits to give particularly desirable properties, while the third effect is undesirable although not usually important because most orbits are not elliptical. In the case of elliptical orbits, this third effect is negated if the inclination i is chosen to be 63.4° or 116.6° .

Some kinds of orbit are particularly useful for remote sensing satellites. The geostationary orbit has $e = 0$ and $i = 0$, and a period equal to the Earth's rotational period (one sidereal day) with the result that the satellite remains in a fixed position relative to the Earth's surface. This kind of orbit is much used for meteorological observations as well as for telephone, television and data relay satellites. Its disadvantages are its large semimajor axis (around 42 000 km) and the fact that it has a poor view of the polar regions. The Molniya orbit provides one possible solution to the problem of observing high latitudes, although at least three such satellites are needed to provide continuous coverage. The Navstar satellites used for the GPS system are in inclined circular orbits with periods of exactly half a sidereal day so that their sub-satellite tracks are repeated daily. The semimajor axis of these satellites is around 26 000 km.

All other satellite orbits used for remote sensing can be classed as low Earth orbits (LEOs), with orbital heights above the Earth's surface of between around 350 and 2000 km. A sun-synchronous orbit is one whose plane precesses at the same rate as the Earth orbits the Sun, which has the result that the satellite crosses a given latitude at the same solar time every day, regardless of longitude. Sun-synchronous orbits have inclinations of around 100° . This kind of orbit is widely used for remote sensing instruments that detect reflected sunlight. An exactly repeating orbit is one whose sub-satellite track forms a closed curve on the Earth's surface so that the same observing geometry is repeated at regular intervals. A dense coverage of tracks implies a long revisit interval, but this can be mitigated by the use of 'drifting subcycles' (which almost repeat after a much shorter interval) or steerable sensors. Orbit suitable for altimetric measurements have sub-satellite tracks oriented close to $\pm 45^\circ$ to lines of longitude, so that ascending and descending paths cross perpendicularly. In the case of altimetric observations of tidal phenomena, it is necessary to consider the relationship between the sampling interval imposed by the orbit and the period of the tide, since some periods will be difficult or impossible to observe as a result of aliasing.

10.4

Satellite station-keeping and orbital manoeuvres

As a final remark on special orbits, we should note that some 'station-keeping' operations will probably be necessary unless the satellite mission is a particularly brief one. Some satellite missions include deliberate changes to the orbital parameters, for example to

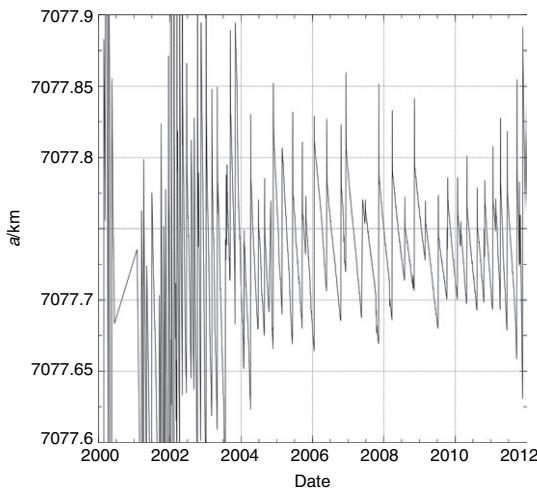


Figure 10.24. Semimajor axis of the Terra satellite from launch until the end of 2011. Note the repeated orbital adjustments of around 100 m, followed by declines in the value of the semimajor axis.

change the values of n_1 and n_2 of an exactly repeating orbit. Even if such alterations are not needed, the satellite will be perturbed in its orbit, for example by atmospheric friction and by the solar wind, and small adjustments may be necessary to bring it back into the desired orbit.

It can easily be shown that the reduction Δa in the semimajor axis of a satellite in a circular orbit, during one orbit of the Earth, is given roughly by

$$\Delta a \approx \frac{4\pi A \rho a^2}{M} \quad (10.23)$$

where A is the cross-sectional area of the satellite, M is its mass and ρ is the atmospheric density. Equation (10.23) is an approximation, since differences in drag coefficient can increase or decrease Δa by factors of up to about 2.

As an example of the application of Equation (10.23), consider the Landsat 5 satellite. This has a mass of about 1700 kg and a cross-sectional area of about 10 m^2 . It orbits at an altitude of 700 km, where the atmospheric density is of the order of $10^{-13} \text{ kg m}^{-3}$, so from Equation (10.23) we expect the satellite to descend by about 0.4 m per orbit or 5 m per day. In fact, observations in changes of the orbital parameters of satellites have been used extensively for the determination of the properties of the Earth's outermost atmosphere.

In order to compensate for the effects of atmospheric friction, and to permit other adjustments to the orbit, remote sensing (and other) satellites are often equipped with small rocket motors to enable them to make these adjustments (Figure 10.24). As discussed in Section 10.2.1, the total amount of fuel needed is proportional to the mass of the satellite and to the total velocity increment that is required. For example, the velocity increment needed to change the semimajor axis of a circular orbit from a to $a + \Delta a$ is

$$\frac{v\Delta a}{2a}$$

where v is the orbital velocity. Thus for a satellite of mass M , the mass of fuel required for such an adjustment will be

$$\frac{Mv\Delta a}{2ua}$$

where u is the speed of the exhaust gases relative to the satellite. A 5-km adjustment to a 1-tonne satellite in LEO might therefore require of the order of 1 kg of fuel.

Ultimately, uncompensated loss of energy through the action of atmospheric drag will cause a satellite in LEO to fall back to Earth, or to burn up in the atmosphere, and this provides an upper limit to the useful life of a satellite. The maximum useful lifetime would be expected, on the basis of Equation (10.23), to be proportional to M/A for a given orbital configuration. The effect of the atmosphere increases sharply with lower values of the perigee distance, and if a satellite were put into a circular orbit with an altitude of only 100 km, its lifetime would be expected to be only a couple of days. Very low orbits, below about 150 km, are thus only used for military reconnaissance missions in which mission duration can be sacrificed in favour of very high spatial resolution.

Implicit in the foregoing discussion is the need to measure the position of a satellite, sometimes to extremely high precision. For example, a satellite carrying a radar altimeter capable of yielding measurements with a precision of 10 cm will need positioning data of comparable or greater precision. This is achieved through measuring the range to the satellite from three or more known locations. The known locations may be fixed on the Earth's surface or they can be other satellites with accurately known orbits, and the range measurements can be made using laser or radio ranging. For example, the satellite Jason-2 uses three positioning systems. One of these is the GPS system using the radio signals transmitted by the constellation of Navstar satellites. The satellite also carries an array of retroreflectors that can return a strong signal from an incident laser beam, in a manner similar to Figure 3.37. A network of ground-based laser ranging stations is used to determine the position of the satellite. (In fact, laser ranging to satellites has a long history and has been used as a way of measuring the Earth's gravitational field.) The third positioning system used by Jason-2 is the Doppler Orbitography and Radio-positioning Integrated by Satellite (DORIS) system, in which a network of 60 ground-based beacons transmit signals at two frequencies (401 and 2036 MHz) (Willis *et al.* 2010). The signals received at the satellite are analysed for time delay and Doppler shift, to enable the satellite's position and velocity to be determined.

Summary

Satellites generally have rocket motors to enable the orbit to be adjusted, either to change the observing pattern or to correct for the effects of atmospheric friction. The latter are more significant for lower orbits where the atmospheric density is greater.

The position of a satellite in orbit can be determined by ranging from GPS, ground-based radio or laser beacons, to an accuracy of a few centimetres.

Review questions

- Compare aircraft and satellites as platforms for remote sensing.
- Explain what is meant by the inclination of a satellite orbit, and distinguish between prograde and retrograde orbits.
- What are the consequences of the lack of spherical symmetry of the Earth's gravitational field for the orbits of satellites?
- Distinguish between geosynchronous and sun-synchronous orbits, and discuss their uses in remote sensing.
- Describe the main types of orbits used for satellite remote sensing, and discuss their advantages and disadvantages.
- Explain why a Molniya orbit might be useful for a remote sensing observation.
- What is a 'drifting subcycle' in a satellite orbit?

Problems

1. Prove Equations (10.2) to (10.5).
2. Consider a satellite in orbit about a spherically symmetric planet. If the orbit is circular, the sub-satellite point travels uniformly along its circular track. If the orbit is not circular, however, the sub-satellite track is still circular, but the sub-satellite point moves along it at a variable rate. Show that the maximum along-track error in calculating the position of the sub-satellite point on the assumption that it moves uniformly is about eD , where e is the eccentricity of the orbit and D is the planet's diameter.
3. A satellite is in a circular sun-synchronous orbit with an inclination of 98.2° . It makes exactly 233 orbits of the Earth in exactly 16 days, and crosses the equator southbound at 0930h local time. Calculate the local time at which it crosses latitude 52° north in the southbound direction.
4. The nodal period of a satellite in a circular orbit of radius a and inclination i about the Earth is given by

$$P_N = \alpha z^{3/2} \left(1 + \frac{\beta(1 - 4 \cos^2 i)}{z^2} \right)$$

and the angular frequency of precession about the Earth's polar axis is given by

$$\Omega_P = \frac{\gamma \cos i}{z^{7/2}}$$

where $z = (a/a_e)$ and a_e is the Earth's equatorial radius, 6378.160 km. The coefficients in the equations are

$$\alpha = 5069.3 \text{ s}$$

$$\beta = 1.62395 \times 10^{-3}$$

$$\gamma = 2.01280 \times 10^{-6}$$

The orbit of the Envisat satellite is circular, with $z = 1.122\,51$ and $i = 98.52^\circ$.

- (i) Show that the orbit is sun-synchronous.
 - (ii) The orbit is an exactly repeating orbit with a revisit interval of 35 days. Calculate the number of orbits that the satellite makes around the Earth in this period.
 - (iii) If the spacing of adjacent southbound tracks is Δl in longitude, show that if the track passes through longitude zero at the equator on day zero, it will pass through longitude $+11\Delta l$ at the equator after approximately one day (the plus sign indicates that this longitude is east of the original longitude) and through longitude $-\Delta l$ on the equator on day 3.
5. The satellite Jason is in a circular orbit with a semimajor axis of 7714 km and an inclination of 66° .
- (i) Show that this orbit is suitable for altimetry.
 - (ii) Comment on the suitability of the orbit for measuring ocean tides.

11

Data processing

The general direction of this book has been to follow approximately the flow of information, from the thermal or other mechanism for the generation of electromagnetic radiation, to its interaction with the surface to be sensed, thence to its interaction with the atmosphere, and finally to its detection by the sensor. It is clear that the information has not yet reached its final destination. First, it is still at the sensor and not with the data user. Second, the 'raw' data will in general require a significant amount of processing before they can be applied to the task for which they were acquired.

In this chapter we shall discuss the more important aspects of the processes to which the raw data are subjected. For the most part, it will be assumed that the data have been obtained from an imaging sensor so that the spatial form of the data is significant. The principal processes are transmission and storage of the data, preprocessing, enhancement and classification. The last three processes are generally regarded as aspects of *image processing*, a major field of study in its own right, and we shall not be able to do much more than outline its general features. There are many books on the subject to which the interested reader may be referred, for example Campbell (2008), Schowengerdt (2007), Mather and Koch (2010), Burger and Burge (2005).

11.1

Transmission and storage of data

It is clear that the data must be brought from the sensor to the place where they are to be analysed. In the case of airborne remote sensing this presents no fundamental difficulty, since missions are comparatively short and it is easy to transport the data, whether recorded on photographic film or digitally on some medium. Similar remarks apply to short-duration missions on low altitude reusable platforms such as the Space Shuttle (or even non-reusable platforms: see Section 5.5 about Corona), but when we consider an unmanned satellite in a long-duration mission (perhaps five years or more) from which the satellite is unlikely to be recovered, the situation is rather different. Here, there are essentially three possibilities.

The simplest and most usual method of transmitting data from a satellite to the ground is to broadcast them continuously from the satellite, as they are received, to a network of receiving stations on the Earth's surface. Successful reception of the data requires that the line of sight from the receiving station to the satellite is not obscured, so that siting of the

11.1 Transmission and storage of data

receiving station is important, and also that the elevation of the line of sight above the horizon is sufficiently large, so that atmospheric degradation of the signal is not significant. Together, these requirements define a *station mask* for a given receiving station. This is the region on the Earth's surface within which the sub-satellite point must lie in order that data can be received from the satellite. If the sensor on board the satellite is a narrow swath nadir-viewing instrument, it is clear that the station mask also corresponds to the region from which data can be collected.

We can calculate the station mask quite simply if we assume that the Earth is spherical with radius R , the satellite is in orbit at an altitude h above the Earth's surface, and that there are no obstructions on the horizon as seen from the receiving station. This is shown in Figure 11.1.

The angular distance ϕ subtended at the Earth's centre between the receiving station and the satellite is given by

$$\cos(\theta + \phi) = \frac{R}{R+h} \cos\theta, \quad (11.1)$$

where θ is the elevation angle of the line of sight. For example, if we set $\theta = 5^\circ$ (a reasonable minimum value in order to avoid an excessively long atmospheric path length), Equation (11.1) shows that for a satellite at an altitude of 700 km, $\phi \approx 21^\circ$. In this case, therefore, the station mask will be a circle of radius 2400 km centred on the receiving station.

Figure 11.2 shows the typical area covered by a single station mask. Although there are now many receiving stations throughout the world, they are all situated on land and, as Figure 11.2 suggests, this cannot be sufficient to provide complete coverage over the oceans. Partly to solve this problem, but also to reduce the number of receiving stations required to give complete global coverage, an alternative approach is to store the data on board the satellite, in a tape recorder or solid-state memory. The data can then be transmitted to a receiving station when the satellite is within view. If a whole orbit's-worth of data (ascending and descending) is stored, and downlinked at the end of the

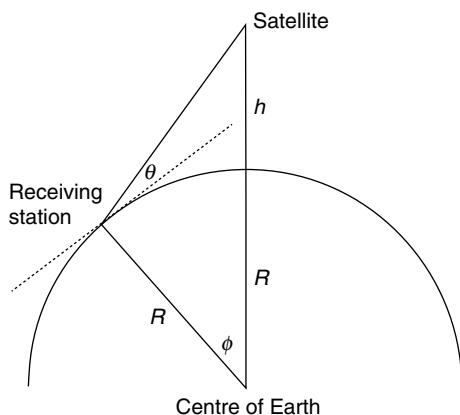


Figure 11.1. Relationship between the elevation angle θ of the line of sight to a satellite and the angle ϕ subtended at the Earth's centre between the receiving station and the satellite.

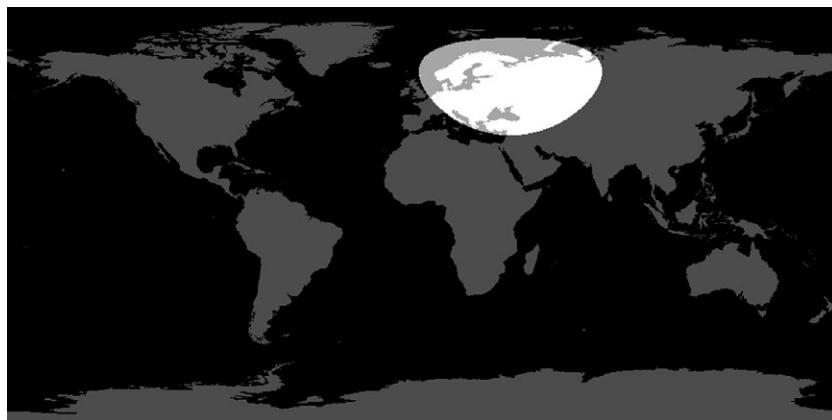


Figure 11.2. Typical idealised station mask for a satellite altitude of 700 km. The receiving station is assumed to be located in Moscow.

orbit, the data will be collected in a time of about 100 minutes and must be transmitted in a time of about 12 minutes, so the data transmission rate must be at least eight times the rate at which they are collected. This fact can limit this mode of operation such that an instrument can collect data for only a fraction of the orbit, although onboard data storage of many gigabits or even terabits can now alleviate the problem.

Finally, we mention the use of relay satellites for increasing the useful coverage of a network of receiving stations. These are satellites in geostationary orbit, and the data collected by the remote sensing satellite are collected by whichever relay satellite is in view and then re-broadcast to Earth. The later Landsat satellites use relay satellites called TDRS (tracking and data relay satellite system).

Whichever of these methods is adopted, we must give some consideration to the volumes of data involved, in order to calculate the transmission rate and data storage requirements. For an imaging sensor of swath width w and rezel area A , carried by a sensor whose ground-track speed is v , the mean rate at which rezels are viewed is vw/A . If n bits of data and k spectral bands (or equivalent, e.g. polarisation states) are recorded from each rezel, the *minimum* data rate must be

$$\frac{k n v w}{A}$$

bits per second. Inserting some typical numbers into this expression, we find data rates for satellite-borne sensors that range from 10 kb s^{-1} ($1 \text{ kb} = 1024 \text{ bits}$) for a radar altimeter to 10 Mb s^{-1} ($1 \text{ Mb} = 1024 \text{ kb}$) for an imaging radar or high resolution optical imaging system. The actual data rates will be higher than those calculated according to the formula, by a factor dependent on the instrument's design and the degree of oversampling. For example, a typical imaging radar will in fact generate data at of the order of 100 Mb s^{-1} . We can see, therefore, that spaceborne remote sensing systems have the potential to generate terabytes of data per day ($1 \text{ Tb} = 1024 \text{ Mb}$), and although at present most of these data are not permanently stored, the requirements on data storage are still large.

Table 11.1. Storage media for digital data. Approximate maximum values of the storage capacities (given in B = bytes = 8 bits) are given. For comparison, an uncompressed Landsat-7 ETM image requires about 700 MB. The approximate volumes needed to store 1 TB of data are also given. For comparison, the volumes of a typical desk drawer, a floor-to-ceiling storage rack and an entire office are of the order of 10^{-2} , 1 and 10 m^3 respectively

Medium/device	Storage capacity (GB)	Volume needed to store 1 TB (m^3)
Blu-ray disc	25	10^{-3}
USB flash drive	250	10^{-4}
Hard disc	1000	10^{-4}
Magnetic tape cartridge	5000	10^{-3}

Computer data storage comes in a variety of types. External storage media, not under the direct control of the computer's processor, can store much larger quantities of data than can be held in the computer's own memory. Access to the computer's own memory is, however, very much faster. Table 11.1 summarises the capacities of the main external data storage media, at the time of writing (in 2011). Digital data storage capacities have grown enormously in the time since the first edition of this book was published in 1990. At that time, a typical hard disc would have had a capacity of around 100 MB.

In addition to raising questions of storage capacity, the sizes of remotely sensed datasets can have implications for the transmission of data over the internet. Broadband download speeds can be as high as 100 Mb/s, although even in the developed world where access to broadband is widespread, achievable download speeds for many customers can be as low as 2 Mb/s and without broadband speed is limited to around 50 kb/s (these values are all typical for 2011). At these speeds, an uncompressed 700-MB Landsat image would take around one minute, one hour and 30 hours, respectively, to download. There is thus still considerable reason to compress image data. This topic is discussed at the end of the chapter.

Summary

Satellite data are usually transmitted to Earth by radio signal, either directly to a ground-based receiving station, or via a relay satellite. In the case of direct transmission from a satellite in low Earth orbit, the receiving station must be within around 3000 km of the sub-satellite point. Typical data volumes for a single satellite image are up to a few hundred megabytes, which are easy to store using modern computer memory technology, but which nevertheless can create some difficulties in transferring images from one computer to another.

11.2

Image processing

As was mentioned earlier, image processing is generally considered to consist of the three steps of preprocessing, image enhancement and classification. Roughly speaking, these steps involve, respectively, the calibration and removal of systematic errors in the data,

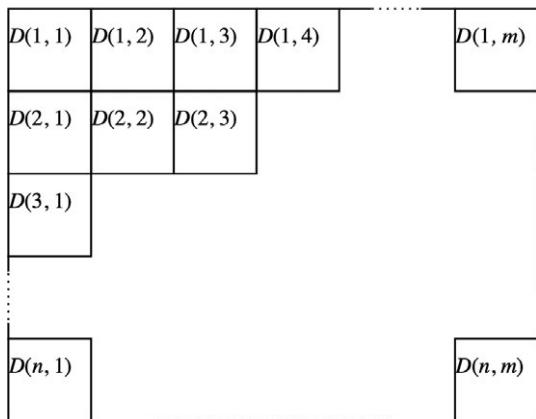


Figure 11.3. Structure of a one-band digital image as a two-dimensional array.

increasing their intelligibility (perhaps by the removal of random errors, or by redisplaying the data) as a representation of the object being sensed, and extracting meaningful patterns from the data. As will be apparent from this brief description, and more so from the more detailed descriptions that follow, the distinctions between these steps are not clear cut. This is one of the justifications for regarding image processing as a single coherent subject, whose aim is to extract meaningful, preferably quantitative, patterns from the detected data. Another justification is of course the very wide applicability of such techniques both within and outside the field of remote sensing.

Most image processing is carried out on digital data, since this is the format in which most data are supplied and since it is much easier to perform all but the simplest operations on digital data that are held in computer memory. We therefore consider a one-band image to be a two-dimensional array of numbers (Figure 11.3), each of which represents the intensity of the radiation reaching the sensor from one element (rezel) of the Earth's surface.

The figure shows an image consisting of $n \times m$ pixels, such that the location of any pixel can be specified by (i, j) , i being the row number and j the column number. The value $D(i, j)$ is an integer, variously called the pixel value, digital number (DN) or *grey level* of the pixel (i, j) . For a multi-band image, the array is three-dimensional, as shown in Figure 11.4. In this case, the pixel (i, j) is represented by the pixel values $D(i, j, k)$, where $k = 1$ corresponds to band 1, $k = 2$ to band 2 and so on.

There is no intrinsic difficulty in modifying these definitions to include continuous data such as photographic images, but it will be more convenient to continue to assume that we are dealing with digital data.

11.2.1

Preprocessing

As described at the beginning of this section, the preprocessing stage applies calibrations and removes systematic errors from the data. The most important operations are the correction of radiometric and geometrical errors, i.e. calibration of the detected signal and registration of the image data to true surface positions. We should also include under this

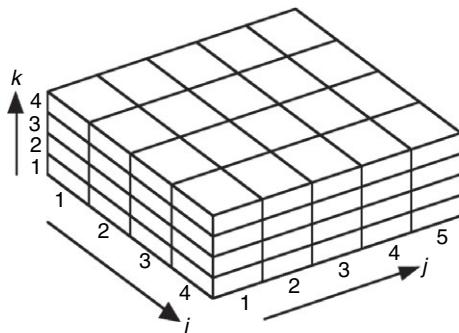


Figure 11.4. Structure of a multi-band image as a three-dimensional array. Each cell contains a single value and is indexed by the values of (i, j, k) , where i is the row number, j the column number and k the band number.

heading the initial stages of processing synthetic aperture radar data (see Chapter 9), which require substantial ‘unscrambling’ to convert them from a set of amplitudes and phases into a radar image. However, the details of SAR processing are beyond our scope here, and we shall not add to the remarks made on the subject in Chapter 9.

11.2.1.1 Radiometric correction

The data will in some cases need to be calibrated, although this is unusual and difficult for photographic images. For optical and infrared systems there are effectively two steps in this process. The first is to establish the relationship between the pixel value recorded by the sensor and the relevant physical property (normally the spectral radiance) of the radiation incident on the sensor. This aspect of calibration is usually achieved, at least in part, through the design of the instrument itself, which will often include internal calibrators or be designed to make periodic observations of, for example, the Sun. In the case of satellite imagery, nominal calibration data are often provided to users by the operating agency, usually in the form of at-satellite radiances corresponding to the minimum and maximum pixel values (for example, to pixel values of 1 and 255 for 8-bit data: see box).

Satellite sensor calibration data: example

The text below is part of the metadata file for a Landsat-7 ETM+ image. It shows that, for example, the calibration of the band 1 data is such that a digital number of 1 corresponds to an at-satellite spectral radiance of $-6.2 \text{ W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$, while a digital number of 255 corresponds to a radiance of $191.6 \text{ W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$. The radiance corresponding to a digital number D is given by linear interpolation as

$$\frac{D - 1}{255 - 1} (191.6 + 6.2) - 6.2 \text{ W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}.$$

Thus a digital number of 100 corresponds to a radiance of $70.9 \text{ W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$.

```

GROUP = MIN_MAX_RADIANCE
    MAX_DETECTING_RADIANCE_LEVEL_BAND1 = 191.600
    MIN_DETECTING_RADIANCE_LEVEL_BAND1 = -6.200
    MAX_DETECTING_RADIANCE_LEVEL_BAND2 = 196.500
    MIN_DETECTING_RADIANCE_LEVEL_BAND2 = -6.400
    MAX_DETECTING_RADIANCE_LEVEL_BAND3 = 152.900
    MIN_DETECTING_RADIANCE_LEVEL_BAND3 = -5.000
    MAX_DETECTING_RADIANCE_LEVEL_BAND4 = 157.400
    MIN_DETECTING_RADIANCE_LEVEL_BAND4 = -5.100
    MAX_DETECTING_RADIANCE_LEVEL_BAND5 = 31.060
    MIN_DETECTING_RADIANCE_LEVEL_BAND5 = -1.000
    MAX_DETECTING_RADIANCE_LEVEL_BAND61 = 17.040
    MIN_DETECTING_RADIANCE_LEVEL_BAND61 = 0.000
    MAX_DETECTING_RADIANCE_LEVEL_BAND62 = 12.650
    MIN_DETECTING_RADIANCE_LEVEL_BAND62 = 3.200
    MAX_DETECTING_RADIANCE_LEVEL_BAND7 = 10.800
    MIN_DETECTING_RADIANCE_LEVEL_BAND7 = -0.350
    MAX_DETECTING_RADIANCE_LEVEL_BAND8 = 243.100
    MIN_DETECTING_RADIANCE_LEVEL_BAND8 = -4.700
END_GROUP = MIN_MAX_RADIANCE
GROUP = MIN_MAX_PIXEL_VALUE
    MAX_PIXEL_VALUE_BAND1 = 255.0
    MIN_PIXEL_VALUE_BAND1 = 1.0
    MAX_PIXEL_VALUE_BAND2 = 255.0
    MIN_PIXEL_VALUE_BAND2 = 1.0
    MAX_PIXEL_VALUE_BAND3 = 255.0
    MIN_PIXEL_VALUE_BAND3 = 1.0
    MAX_PIXEL_VALUE_BAND4 = 255.0
    MIN_PIXEL_VALUE_BAND4 = 1.0
    MAX_PIXEL_VALUE_BAND5 = 255.0
    MIN_PIXEL_VALUE_BAND5 = 1.0
    MAX_PIXEL_VALUE_BAND61 = 255.0
    MIN_PIXEL_VALUE_BAND61 = 1.0
    MAX_PIXEL_VALUE_BAND62 = 255.0
    MIN_PIXEL_VALUE_BAND62 = 1.0
    MAX_PIXEL_VALUE_BAND7 = 255.0
    MIN_PIXEL_VALUE_BAND7 = 1.0
    MAX_PIXEL_VALUE_BAND8 = 255.0
    MIN_PIXEL_VALUE_BAND8 = 1.0
END_GROUP = MIN_MAX_PIXEL_VALUE

```

This first step in radiometric correction is to convert the pixel values into ‘at-satellite radiances’. The second step is correction for atmospheric propagation effects, to obtain ‘at-surface radiances’. The means by which this can be achieved have been discussed in Sections 6.1.5 and 6.4.5. In some cases, for example where it is necessary to compare

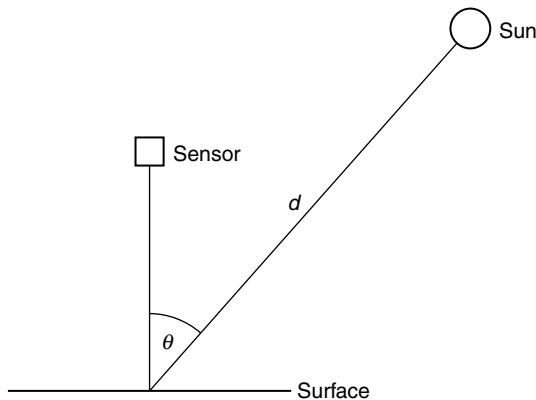


Figure 11.5. A sensor views a horizontal surface at nadir, the surface being illuminated by the Sun at a zenith angle of θ . The Sun's distance from the Earth is d astronomical units.

optical images that have been acquired under different conditions of illumination, it can be useful to correct the data for the effects of illumination geometry and even for the distance from the Sun to the Earth. This distance varies throughout the year as a result of the Earth's elliptical orbit. The length of the semimajor axis of the orbit is defined as one *astronomical unit* (AU), equal to around 1.496×10^{11} m, and the Earth's distance from the Sun varies between 0.983 AU in early January and 1.017 AU in early July (see appendix).

Figure 11.5 illustrates a typical illumination geometry for a nadir-viewing sensor. If the surface scattering can be assumed to be Lambertian (see Section 3.3.2), the reflected at-surface radiance L (i.e. the radiance at the surface heading towards the sensor) can be written as

$$L = \frac{E_{exo}}{\pi d^2} \rho \cos \theta \quad (11.2)$$

where E_{exo} is the mean exoatmospheric irradiance of the Sun at a distance of 1 AU, d is the actual distance from the Sun to the area being sensed, and ρ is the effective planetary reflectance of the surface. By using Equation (11.2) to calculate ρ from L , d and θ , we obtain a measure of reflectance that is largely independent of the viewing conditions.

Planetary reflectance: example

The text below is another part of the metadata file for the satellite image from the previous box. It shows that the image was acquired on 22 July 2001, and that the Sun's elevation angle (at the centre of the image) was 41.1° at the time. The Sun's zenith angle θ was thus 48.9° , and the distance from the Sun to the Earth on this date was 1.016 AU (this is the distance from the Sun to the Earth's centre, not its surface, but since the radius of the Earth is only 0.00004 AU the difference can be ignored). The

mean exoatmospheric spectral radiance in the ETM+ band 1 is $1969 \text{ W m}^{-2} \mu\text{m}^{-1}$, so a band 1 radiance of $70.9 \text{ W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$ corresponds to a planetary reflectance of

$$\frac{70.9 \times \pi \times 1.016^2}{1969 \times \cos 48.9^\circ} = 0.178$$

```

GROUP = PRODUCT_METADATA
...
SPACECRAFT_ID = "Landsat 7"
SENSOR_ID = "ETM+"
ACQUISITION_DATE = 2001-07-22
...
END_GROUP = PRODUCT_METADATA
GROUP = PRODUCT_PARAMETERS
...
SUN_AZIMUTH = 43.5363117
SUN_ELEVATION = 41.1244090
...
END_GROUP = PRODUCT_PARAMETERS

```

Active microwave systems are often calibrated against a target of known backscattering cross-section, such as a corner-cube reflector (Figure 3.37)

11.2.1.2 Georeferencing and geometrical correction

Georeferencing involves relating the spatial coordinates in the image (i.e. the row and column coordinates (i, j) of a pixel) to the corresponding spatial coordinates on the Earth's surface. Geometrical correction is the process of changing this relationship. If data are available on the position and direction of view of the sensor at the time the image was acquired, this may be enough to establish the necessary relationship. In most cases, however, it is necessary to use ground control points (GCPs) to determine the relationship. The use of GCPs has already been mentioned very briefly in Section 10.2. They are particularly important if several images are to be joined together in a mosaic, or if images of the same area, acquired perhaps at different times or with different sensors, need to be overlaid and compared.

The relationship between the coordinates of a pixel in the image and of the corresponding point on the Earth's surface can be expressed functionally. For example, if we denote the coordinates of a point in the image by (x_i, y_i) (these are most commonly the row and column numbers, which are integers, but for now we assume that these are continuous variables) and the coordinates of the corresponding point on the surface by (x_s, y_s) (these could be, for example, latitude and longitude), we can express the needed relationship quite generally as

$$x_s = f(x_i, y_i), \quad (11.3.1)$$

$$y_s = g(x_i, y_i). \quad (11.3.2)$$

11.2 Image processing

The functions f and g will depend on the kind of relationship that we assume exists between image coordinates and surface coordinates. In simple cases, we can assume that the same functions f and g apply everywhere within the image. In more complicated cases it may be necessary to find the appropriate forms of f and g locally.

The simplest useful model that has the form of Equations (11.3) is the *affine transformation*, represented by Equation (11.4):

$$\begin{pmatrix} x_s \\ y_s \\ 1 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix}. \quad (11.4)$$

Although much more complicated models, with many more parameters, are obviously possible, this model can allow for a shift in origin (through a_3 and a_6), rotation, different scales in the x and y directions, and some types of skew. Since it has six parameters (a_1 to a_6) and each GCP provides two pieces of information (an x coordinate and a y coordinate), in principle three GCPs are sufficient to determine this model. In practice one would use many more than three (a good general principle is to use ten times as many as the minimum), as a precaution against random errors. Determination of the appropriate values of the model parameters a_1 to a_6 is then carried out using a minimum least squares method. Figure 11.6 shows a typical result of using GCPs to define the parameters of a transformation having the form of Equations (11.3). The grid superimposed on the image has been derived from the GCPs using a least-squares fitting procedure. It represents the fact that the surface coordinates (x_s, y_s) are known for every point (x_i, y_i) in the image. The fact that the grid does not consist of straight lines in this case shows that the affine transformation (11.4) is insufficient to represent the relationship between image coordinates and geographical coordinates. In general, smaller areas are more likely than larger areas to be transformed with adequate accuracy by an affine transformation.

The example of Figure 11.6 shows an image for which the transforming functions f and g , defined in Equations (11.3), have been calculated. However, we often wish to go one step further and actually change the geometry of the image so that it conforms to the chosen coordinate system (x_s, y_s) . For example, in the case of Figure 11.6 this would correspond to distorting the image so that the superimposed grid became rectangular, with the grid lines running horizontally and vertically. In general, we can define such a transformation as follows:

$$i = F(i', j') \quad (11.5.1)$$

$$j = G(i', j') \quad (11.5.2)$$

where (i, j) are the pixel coordinates in the untransformed image, (i', j') are the pixel coordinates in the transformed image, and the functions F and G can be derived from the functions f and g in Equations (11.3). Equations (11.5) provide a ‘recipe’ for finding the pixel coordinate (i, j) in the original untransformed image that should be copied into a location (i', j') in the new image. The pixel coordinates are only defined for integer values of the row and column numbers i and j , but Equations (11.5) can yield non-integer values. If this happens, how do we choose the pixel value from the original image to copy into the new image? This is the problem of *resampling* an image.

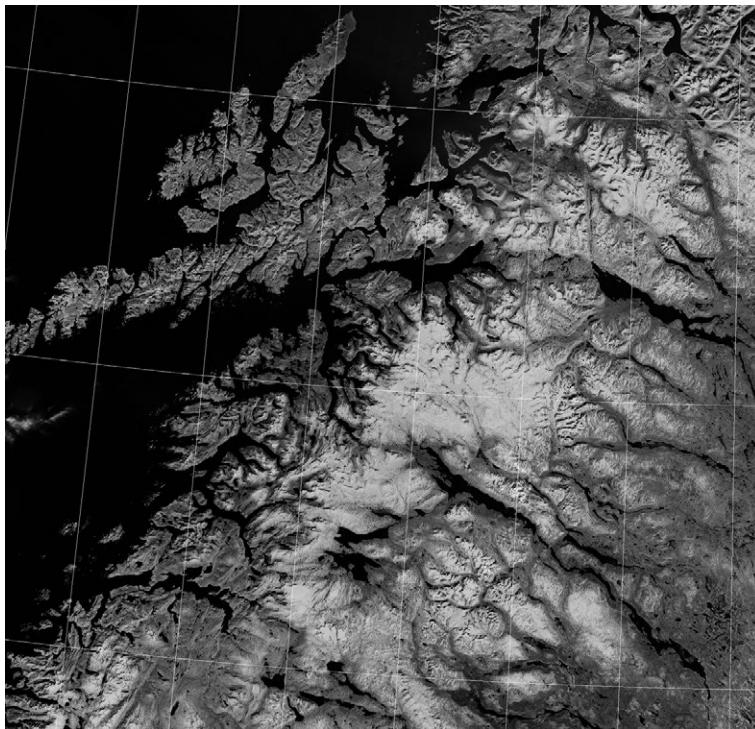


Figure 11.6. A georeferenced satellite image. The image is an extract of a mosaic of Landsat TM images showing parts of northern Norway and Sweden. It covers an area of approximately 57×60 km. The latitude–longitude grid has been superimposed after georeferencing the image. See also colour plates section.

The simplest resampling technique is *nearest-neighbour resampling*. Here, we simply identify the pixel in the original image that is spatially nearest to the calculated position (i, j) , and copy its pixel value to the location (i', j') in the new image. This approach has the merit that it does not alter any of the pixel values; on the other hand, it can produce resampled images that are visually somewhat displeasing since they tend to have jagged edges (Figure 11.7). Smoother results are produced by interpolation, in which the pixel value copied into the new location (i', j') is a suitably weighted average of the pixel values in the neighbourhood of the calculated location (i, j) . The commonest forms are *bilinear interpolation*, which uses a 2×2 pixel neighbourhood, and *bicubic interpolation*, which uses a 4×4 pixel neighbourhood (Figure 11.7). Detailed discussions of these and other interpolation techniques can be found in almost any work on digital image processing, for example (Campbell 2008). However, interpolation is generally undesirable if subsequent processing of the data will make quantitative use of the pixel values, since these are to some extent corrupted by the averaging process.

Before leaving the subject of geometrical correction, we should make a final remark about aerial photography. We saw in Section 5.5 how the relief displacement in a pair of stereophotographs can be used to deduce the surface relief of the area common to both photographs. The relief displacement is itself a form of geometrical distortion in the

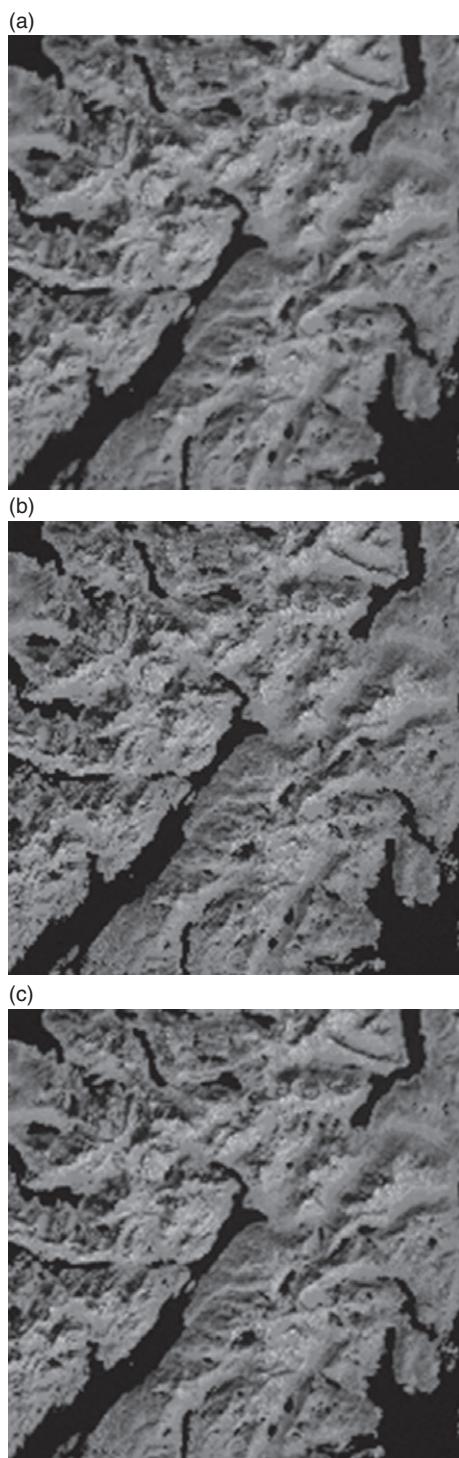


Figure 11.7. Extract of the image in Figure 11.6 reprojected to a geographical projection using (a) nearest-neighbour, (b) bilinear, and (c) bicubic resampling. See also colour plates section.

photographs. It can be removed if the relief and the viewing geometry are known, using Equations (5.8) and (5.9) to define the relationship between the coordinates of a point on the surface and the coordinates of the corresponding point in the image. A photograph to which this procedure has been applied is termed an *orthophotograph* (see for example Berlin and Avery (2003)). The analogous term *orthoimage* is used for digital imagery that has been corrected in the same way.

11.2.2 Image enhancement

Improvements to the image can be divided into those that operate on individual pixels without reference to their spatial context, and those that also make use of spatial information. The first type can generally be referred to as contrast modification, and the second as spatial filtering.

11.2.2.1 Contrast modification

In order to discuss image contrast and its modification, we first need to introduce the concept of the *image histogram*. This is simply a graph or table showing the number $h(d)$ of pixels in an image (or a sub-region of an image) having pixel value d . Contrast modifications involve changing the shape of the histogram by reassigning one pixel value to another. This can be represented by a graph of the ‘transfer function’, showing how the output pixel value is related to the input pixel value. These ideas are illustrated in Figure 11.8.

The figure shows a contrast modification in which the transfer function has a constant gradient that is greater than unity. This is referred to as a linear *contrast stretch*, and it has the effect of expanding the range of pixel values occupied by the data. In the example of Figure 11.8, the input data occupy a rather compressed range of pixel values, with almost no values at all occurring outside the range 10 to 110. The contrast stretch has therefore been chosen to expand this range to fill the whole range of pixel values (in this example, 0 to 255) that is available. (Figure 11.8c also shows a characteristic result of increasing the contrast of an image. Because the range of pixel values has been expanded, some pixel values in the new image are not represented – the histogram values are zero.) This will also obviously have the advantage of enhancing subtle tonal variations in the image (see Figure 11.9). These can be enhanced even further by increasing the gradient of the transfer function, but this would lead to some ‘clipping’ of the data. For example, if the slope of the transfer function were 5, a range of input pixel values of only $255/5 = 51$ would lead to a range of 255 in the output values.

The linear contrast stretch is merely the simplest of a range of possible contrast modifications. In general, these can be thought of as attempts to give the image histogram some desired form. We might, for example, wish to modify the contrast of one image so that its histogram matches that of another image, before attempting to make a mosaic from the two images.

Suppose we have an image with N_1 pixels and histogram $h_1(d)$, where d can take any value between 0 and n_1 , and we want to find a transfer function $f(d)$ that will transform the histogram into $h_2(d)$, where d can now take any value between 0 and n_2 . The new image has N_2 pixels. Normally, n_2 will equal n_1 (for example, if we are dealing with 8-bit data, both values will be 255), but we might, for example, wish to scale 10-bit input data into 8-bit output data. Similarly, N_2 will normally equal N_1 , although we might wish to match

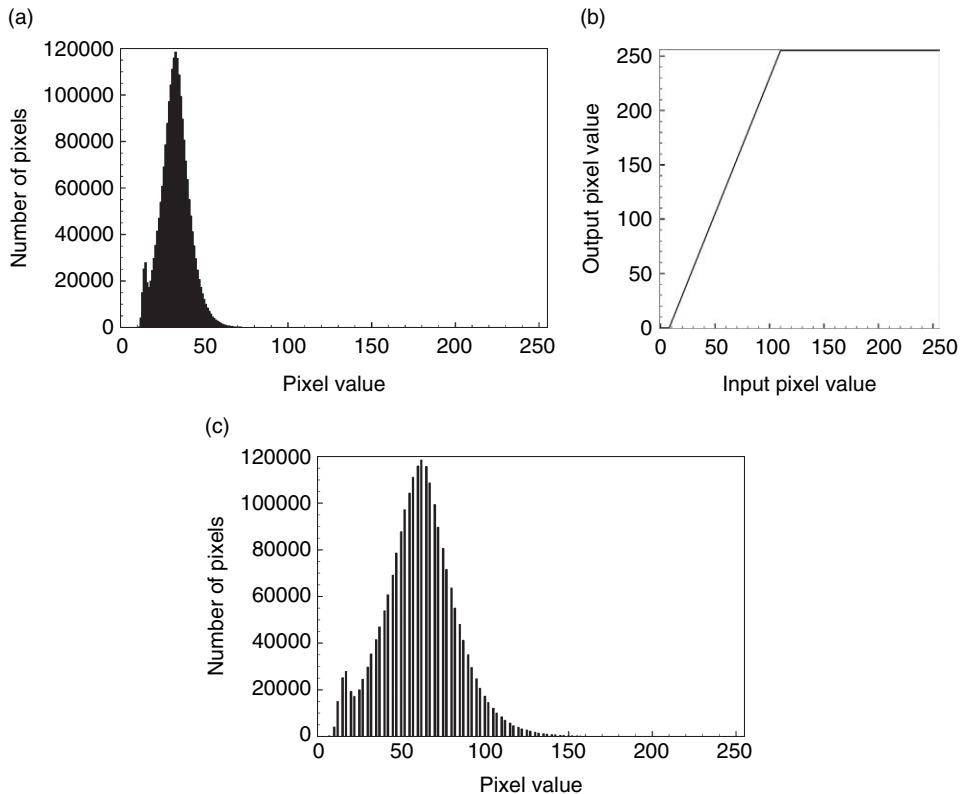


Figure 11.8. (a) An image histogram before contrast modification; (b) a transfer function relating input and output pixel values; (c) the image histogram after applying the transfer function (b) to the original image.

the histograms of two different images. Whatever the values of n_1 and n_2 , N_1 and N_2 , the required transfer function can be calculated as follows. First, scaled cumulative histograms are calculated:

$$g_1(d) = \frac{n_1}{N_1} \sum_{i=0}^d h_1(i), \quad (11.6.1)$$

$$g_2(d) = \frac{n_2}{N_2} \sum_{i=0}^d h_2(i). \quad (11.6.2)$$

$g_1(d)$ thus has a maximum value of n_1 when $d = n_1$, and similarly for $g_2(d)$. Next, we calculate the inverse function of g_2 , $g_2^{-1}(d)$. This is defined by

$$g_2^{-1}(g_2(d)) = g_2(g_2^{-1}(d)) = d. \quad (11.7)$$

Finally, the required transfer function is calculated from

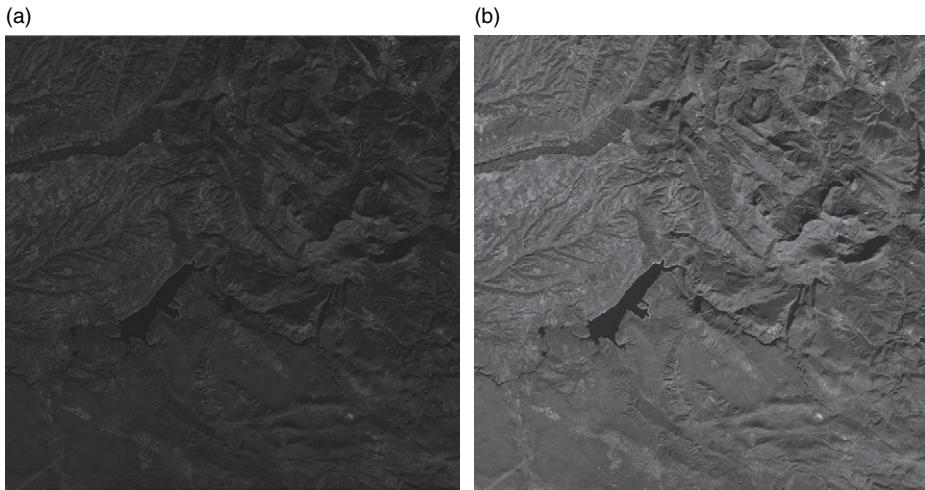


Figure 11.9. Linear contrast stretch. (a) Original image (with image histogram as shown in Figure 11.8a); (b) after application of the transfer function shown in Figure 11.8b. (Landsat-7 ETM+ image, band 8, showing Lac de Sainte-Croix and surroundings, France.)

$$f(d) = g_2^{-1}(g_1(d)). \quad (11.8)$$

As an example of this method, let us consider the very common contrast modification known as *histogram equalisation*. Here, the aim is to produce a histogram that is ‘flat’, i.e. all pixel values occur with equal frequency. For simplicity we assume that $n_1 = n_2 = n$, $N_1 = N_2 = N$. The required output histogram thus has $h_2(d) = N/(n + 1)$ for all values of d , and hence, from Equation (11.6.2),

$$g_2(d) = \frac{n(d + 1)}{n + 1}.$$

The inverse function, defined by Equation (11.7), is therefore

$$g_2^{-1}(d) = \frac{n + 1}{n}d - 1,$$

and hence the required transfer function is

$$f(d) = \frac{n + 1}{N} \sum_{i=0}^d h_1(i) - 1.$$

This is a suitably scaled version of the cumulative histogram of the original image. Figure 11.10 shows the image histogram before and after equalisation, and the transfer function.

The histogram in Figure 11.10c does not look particularly flat. This is a consequence of the fact that the transfer function defined by Equation 11.8 cannot be realised precisely using integers. However, if the histogram were recalculated in ‘bins’ of, for example, 16 pixel values, this digitisation effect would be much less apparent and the histogram would appear much flatter. Figure 11.11 shows the appearance of the image of Figure 11.9 after histogram equalisation.

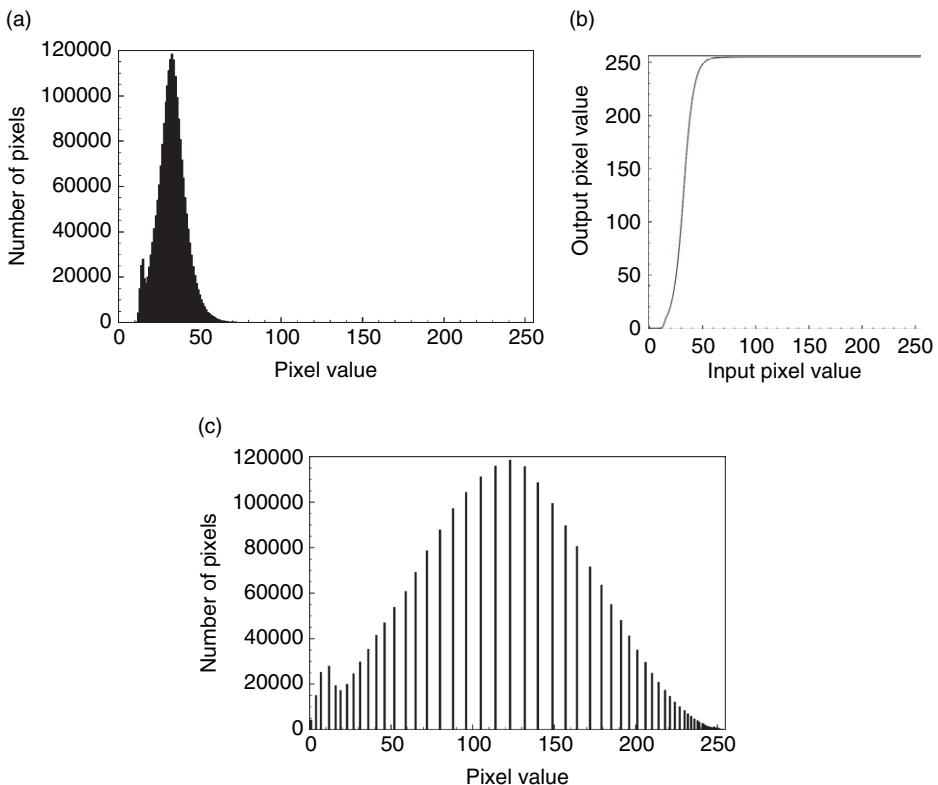


Figure 11.10. (a) Image histogram before contrast modification; (b) transfer function for histogram equalisation; (c) image histogram after equalisation.

Finally, we note that contrast modifications of the type discussed in this section are sometimes applied only to the look-up table (LUT) used to generate the display of an image, and not to the image data themselves. In this way, the operator of the image processing system is assisted in identifying features of interest in the image, but the original image data are unchanged.

11.2.2.2 Spatial filtering

The contrast modifications discussed in the previous section do not necessarily alter the image data, merely the way they are displayed. Spatial filtering, on the other hand, does change the image data. The pixel value at a given location is modified according to the pixel values in its neighbourhood.

The simplest type of spatial filtering is spatial averaging, normally applied to reduce random noise or speckle in the data. The most obvious way of accomplishing this is to replace a given pixel value with the average pixel value for that pixel and its neighbours. This can be represented diagrammatically by a grid of boxes, each box representing a pixel, the central box representing the pixel to be processed, and the number in each box representing the weight of the contribution made by that pixel to the total sum (Figure 11.12). If the total of all these weights is unity, the average pixel value



Figure 11.11. The image of Figure 11.9 after histogram equalisation.

$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

Figure 11.12. Uniform 3×3 smoothing filter.

(brightness) of the image will be unchanged. In this case the filter is said to be *normalised*. The effect of the filter defined by Figure 11.12 is shown in Figure 11.13.

Variants of this type of spatial averaging operator can have weights that decrease with distance from the centre, although the sum of the weights must still be unity if the operation is required not to change the average brightness of the image. However, all spatial averaging operations, as well as their desirable property of reducing noise, have the generally undesirable effect of blurring the image. For example, sharp edges will be smoothed out, and the contrast between a single bright or dark pixel against its neighbours will be reduced. For this reason, spatial averaging can also be described as a *smoothing* operation.

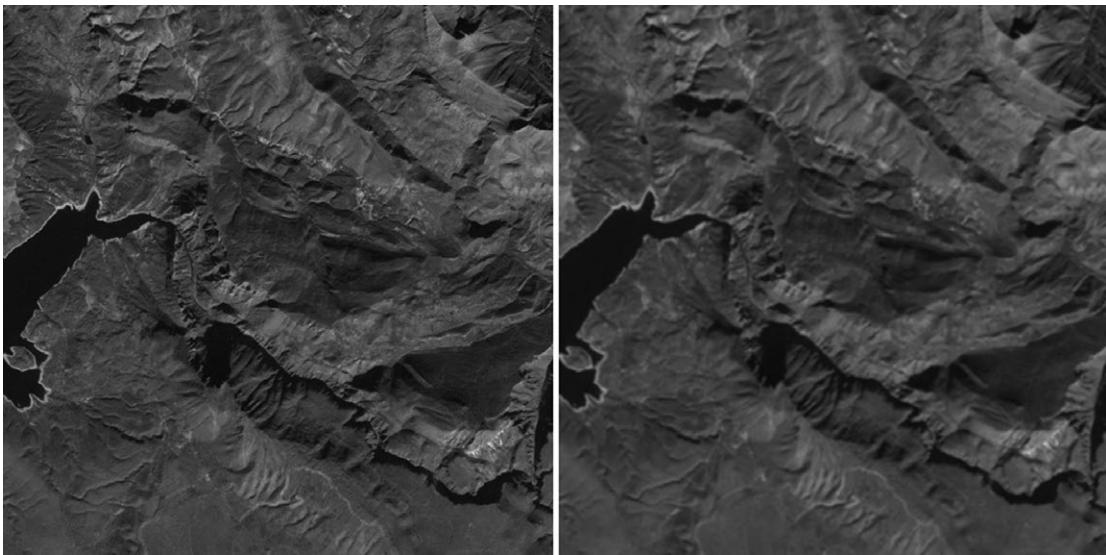


Figure 11.13. Image before and after applying the smoothing filter defined by Figure 11.12. The image is an extract of Figure 11.9.

$-\frac{1}{9}$	$-\frac{1}{9}$	$-\frac{1}{9}$
$-\frac{1}{9}$	$\frac{17}{9}$	$-\frac{1}{9}$
$-\frac{1}{9}$	$-\frac{1}{9}$	$-\frac{1}{9}$

Figure 11.14. A sharpening filter.

Many other types of spatial filter can be considered. For example, a *sharpening* filter has the opposite effect to a smoothing filter, narrowing the widths of boundaries between regions of different brightness, increasing the contrast between single pixels and their backgrounds, and (undesirably) increasing the prominence of noise in the image. One approach to the design of sharpening filters can be represented symbolically as

$$k\mathbf{I} + (1 - k)\mathbf{A},$$

where \mathbf{I} is the ‘identity operator’, i.e. the operator that, when performed on the image, leaves it unchanged (this operator has a ‘1’ in the central box and ‘0’s everywhere else), \mathbf{A} is an averaging (smoothing) operator, and k is some number greater than unity that defines the degree of sharpening. For example, if we take $k = 2$ and \mathbf{A} as in Figure 11.12, we obtain the sharpening filter shown in Figure 11.14

The effect of this filter is illustrated in Figure 11.15. Note that although edges have been enhanced, noise in the image has also been emphasised.

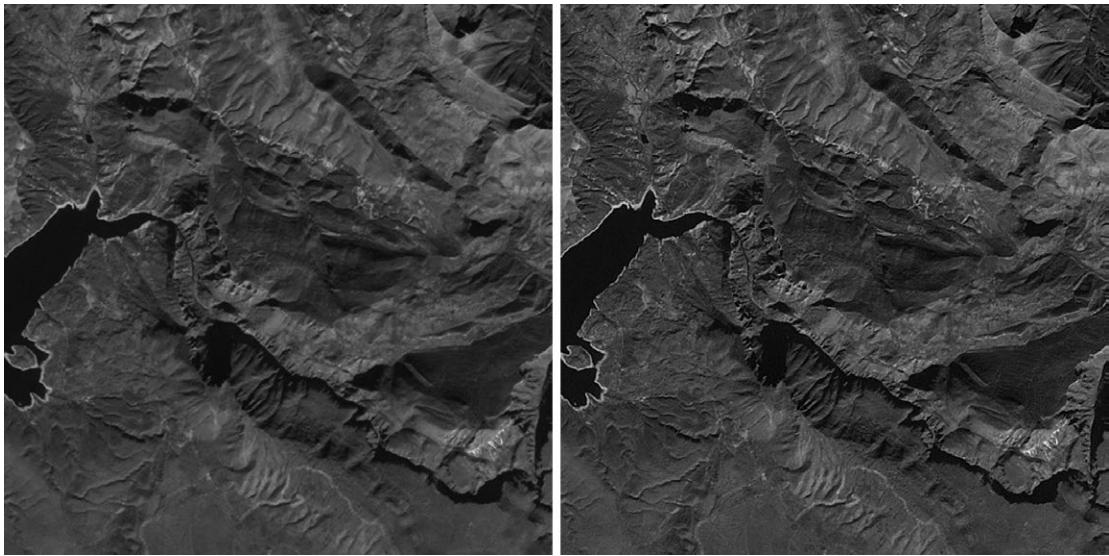


Figure 11.15. Image before and after applying the sharpening filter defined by Figure 11.14.

(a)	(b)
0 1 0	0 $-\frac{1}{4}$ 0
1 -4 1	$-\frac{1}{4}$ 2 $-\frac{1}{4}$
0 1 0	0 $-\frac{1}{4}$ 0

Figure 11.16. (a) Laplacian operator \mathbf{L} ; (b) sharpening filter $\mathbf{I} - \mathbf{L}/4$.

Another common type of sharpening filter can be represented as

$$\mathbf{I} - k \mathbf{L},$$

where \mathbf{L} is the Laplacian operator, whose 3×3 representation is shown in Figure 11.16a, and k is greater than zero. Figure 11.16b shows the corresponding sharpening filter for $k = 1/4$.

Instead of sharpening an image to enhance edges, we may sometimes wish to perform an *edge-detection* operation, in which edges are emphasised but from which uniform areas are removed. This can clearly be carried out using an operator

$$\mathbf{I} - \mathbf{S},$$

where \mathbf{S} is a sharpening operator. Thus the Laplacian operator of Figure 11.16a is an edge-detection filter (it is in fact a digital implementation of the isotropic second derivative), as is any operator $\mathbf{I} - \mathbf{A}$. These are both isotropic edge-detection filters, meaning that they can detect an edge in any orientation. The effect of the edge-detection filter of Figure 11.16a is shown in Figure 11.17.

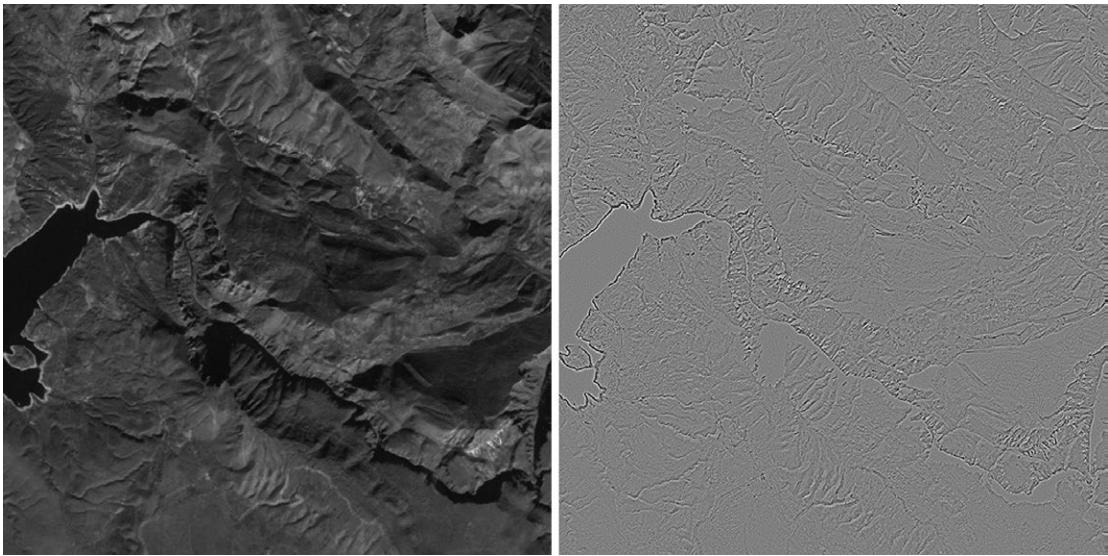


Figure 11.17. Image before and after applying the Laplace filter of Figure 11.16a. A constant value of 128 has also been added to the pixel values so that both negative and positive values can be shown. The filter is an isotropic edge-detector.

(a)	(b)
-1 0 1	1 2 1
-2 0 2	0 0 0
-1 0 1	-1 -2 -1
(c)	(d)
0 1 2	-2 -1 0
-1 0 1	-1 0 1
-2 -1 0	0 1 2

Figure 11.18. Sobel filters for directional edge-detection.

Non-isotropic filters can also be defined to detect edges in specific orientations. Common examples of these filters are the Roberts and Sobel filters. A set of Sobel filters is shown in Figure 11.18, and the effect of one of them is illustrated in Figure 11.19.

The filters that we have just been discussing are all examples of spatial *convolution operations*. Such an operation can be expressed mathematically as

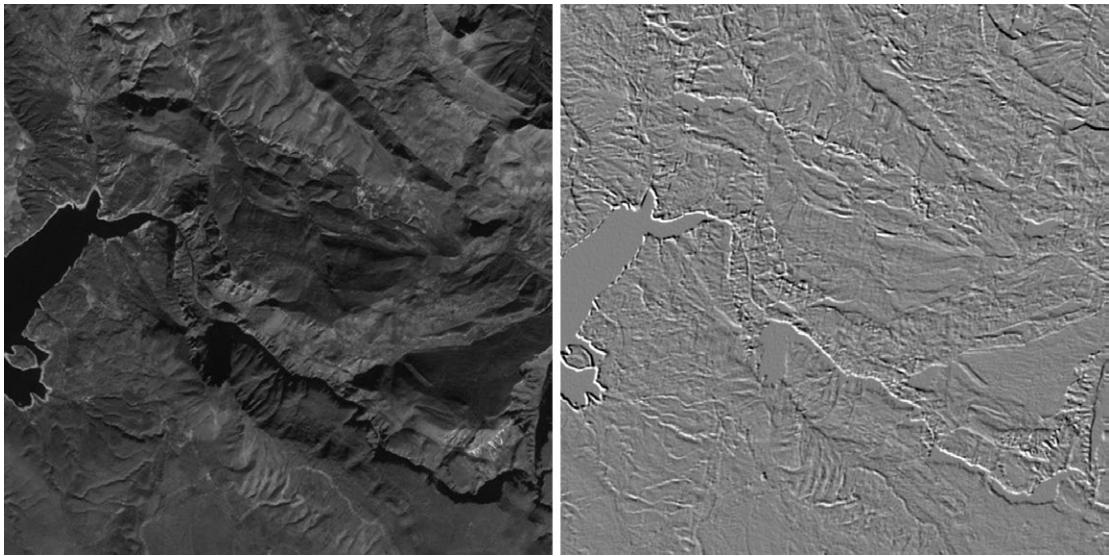


Figure 11.19. Image before and after applying the Sobel filter of Figure 11.18c. A constant value of 128 has also been added to the pixel values so that both negative and positive values can be shown. The filter is a directional edge-detector.

$$d'(i,j) = \sum_{k=a}^b \sum_{l=c}^d w(k,l) d(i+k, j+l), \quad (11.9)$$

where $d(i, j)$ is the original pixel value at (i, j) and $d'(i, j)$ is its transformed value, and $w(k, l)$ defines the array of weights and is called the *kernel*. This kernel is defined over the range of integers $k = a$ to b and $l = c$ to d (a and c are equal to -1 , b and d to $+1$, in the 3×3 example of Figure 11.12). An alternative procedure for carrying out such an operation is through the use of Fourier transforms. In Section 2.3 we introduced and discussed the concept of the Fourier transform, which relates the dependence $f(t)$ of some variable f on the time t to an equivalent description $a(\omega)$ in terms of the angular frequency ω . An analogous pair of transforms can be defined in the spatial domain. Since images are two-dimensional, it is most convenient to consider the two-dimensional Fourier transforms. These can be written as

$$a(\mathbf{q}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\mathbf{x}) \exp(-i\mathbf{q} \cdot \mathbf{x}) d\mathbf{x} \quad (11.10.1)$$

$$f(\mathbf{x}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(\mathbf{q}) \exp(i\mathbf{q} \cdot \mathbf{x}) d\mathbf{q}, \quad (11.10.2)$$

where \mathbf{x} is the vector (x, y) describing a position in the image, and $f(\mathbf{x})$ is the image brightness (or whatever variable is to have the Fourier transform applied to it) at that position; \mathbf{q} is a *spatial frequency*, expressed as a vector (qx, qy) , and $a(\mathbf{q})$ is the corresponding description of the image in terms of its spatial frequency components.

11.2 Image processing

In fact, since we are considering digital images, we use the discrete Fourier transform (DFT) rather than the spatial Fourier transforms defined by Equations (11.10). If the image is defined in a rectangular array of width m pixels, so that the column number j ranges from 0 to $m - 1$, and height n pixels, with the row number k ranging from 0 to $n - 1$, the DFT is also defined in an array of $m \times n$ pixels, with

$$a(j', k') = \frac{1}{mn} \sum_{j=0}^{m-1} \sum_{k=0}^{n-1} f(j, k) \exp\left(-2\pi i \left(\frac{jj'}{m} + \frac{kk'}{n}\right)\right). \quad (11.11.1)$$

The inverse transform is then

$$f(j', k') = \frac{1}{mn} \sum_{j=0}^{m-1} \sum_{k=0}^{n-1} a(j, k) \exp\left(2\pi i \left(\frac{jj'}{m} + \frac{kk'}{n}\right)\right). \quad (11.11.2)$$

The term at spatial frequency zero appears at coordinates $(0, 0)$ in the DFT. The frequency sampling interval in the x -direction is m^{-1} cycles per pixel (and correspondingly n^{-1} in the y -direction). However, the maximum x -component of the spatial frequency in the DFT is not 1 cycle per pixel, as might be expected, but 1/2 cycle per pixel. This occurs at $j' = m/2$, and larger values of j' correspond to negative spatial frequencies, with $j' = m - 1$ corresponding to a spatial frequency of $-m^{-1}$ cycles per pixel. This is yet another example of the phenomenon of aliasing, discussed in Section 10.3.2.4.

In fact, practically all DFTs are calculated using the fast Fourier transform (FFT) algorithm, which is most conveniently performed when the image is square, with m and n equal to one another and to a power of 2. Bracewell (1999) provides much more detail.

The potential advantage of considering spatial filtering operations in terms of Fourier transforms is that a convolution in the spatial domain is equivalent to a multiplication in the frequency domain. If the spatial extent of the convolution kernel is very large (the ‘box’ of the kind shown in Figures 11.12 and similar is large), the extent of the corresponding Fourier transform will be small (this is an extension of the idea discussed in Equation (2.18), and it may be computationally more efficient to calculate the DFT of the image, multiply by the appropriate filter function, and then retransform back into the spatial domain.

Figure 11.20 shows the Fourier transforms of some of the spatial filters described earlier in this section.

Figure 11.20a shows that the smoothing filter of Figure 11.12 retains the lower spatial frequencies (the transform values are close to 1 near $\mathbf{q} = 0$) but suppresses the higher ones. For this reason, an alternative name for such a filter is a *low-pass filter*. Similarly, the sharpening filter (Figure 11.14) can be regarded as a *high-boost filter*, since it preserves the lower spatial frequencies and increases the amplitudes of the higher frequencies. The Laplace filter (Figures 11.16a) suppresses the lowest spatial frequencies (the transform values are close to zero near $\mathbf{q} = 0$) and is hence a *high-pass filter*.

Finally, we should mention two more important classes of spatial filter. The first of these are the *non-linear* filters. These cannot be represented as convolution operations with the form of Equation 11.9. Perhaps the simplest example of a non-linear filter is the median filter, in which the central pixel of an $N \times N$ box is replaced by the median value of all the pixels in the box. This has some advantages with respect to the simple averaging

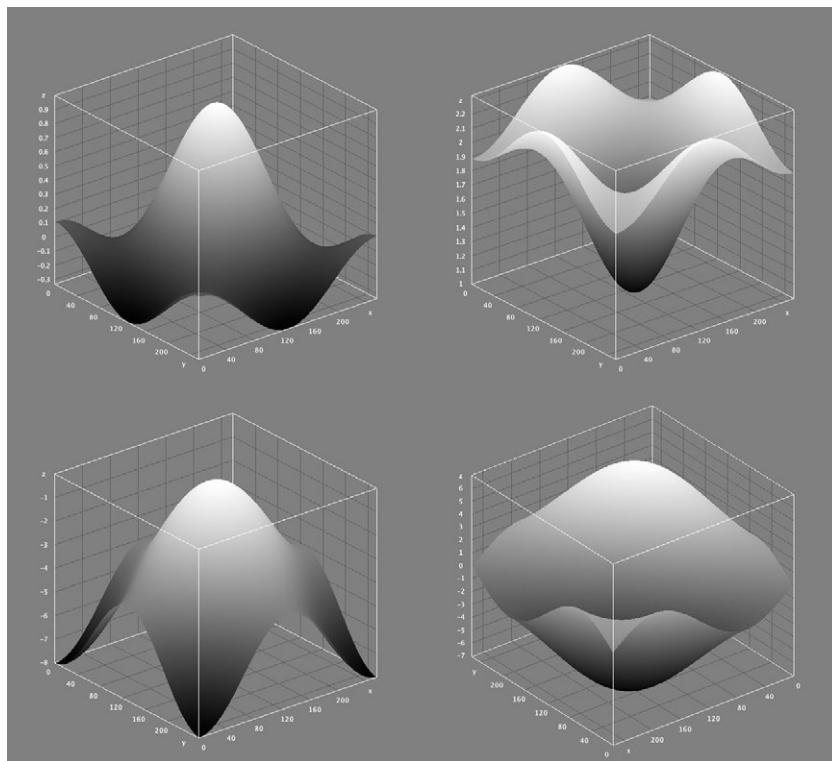


Figure 11.20. Fourier transforms of (top left) the smoothing filter of Figure 11.12; (top right) the sharpening filter of Figure 11.14; (bottom left) the Laplace filter of Figure 11.16a; (bottom right) the Sobel filter of Figure 11.18c. The surfaces show the real part of the response except in the case of the Sobel filter where the response is purely imaginary. The region plotted is from $-1/2$ to $+1/2$ cycles per pixel in the x - and y -directions.

filter of Figure 11.12, notably the fact that it will preserve sharp edges in an image. The effect of the median filter is illustrated in Figure 11.21. Many other non-linear filters have been devised to take into account specific characteristics of the image, for example filters to reduce speckle in SAR images (although in fact the median filter is also rather effective in this role).

The last type of spatial filter that we discuss is the *desstriping filter*. This is intended to correct the effects of poor row-to-row calibration in some scanning systems, notably the Landsat MSS system. We noted in Section 6.2.2 that the Landsat TM system scans 16 lines of data simultaneously. The MSS system used a similar approach, scanning six lines simultaneously using six different detectors per spectral band. Since these detectors did not have identical calibrations, the result was to produce a characteristic ‘striping’ or ‘banding’ in the images, having a period of 6 pixels. Most desstriping algorithms work by adjusting the pixel values in a single line of pixels so that their mean and standard deviation match the mean and standard deviation of some reference pixels. They can thus be thought of as a form of local contrast modification. For example, a typical MSS desstriping algorithm processes the image in blocks 100 pixels wide and 6 pixels in the

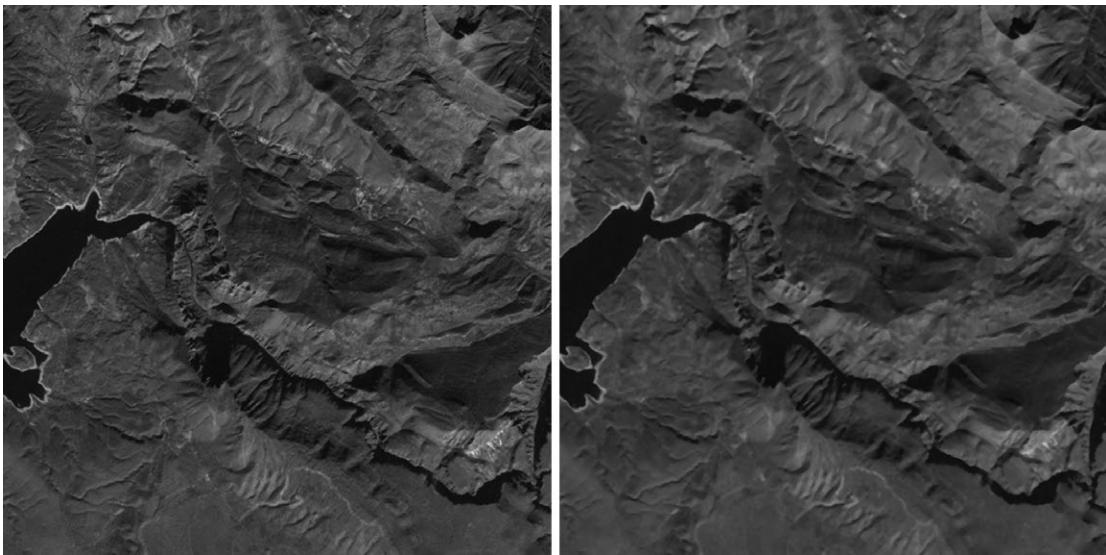


Figure 11.21. Image before and after applying a 3×3 median filter.

along-track direction. Linear contrast stretches are applied to the second to fifth rows of pixels so that their means and standard deviations match those of the first row. However, alternative approaches are also possible through the use of Fourier transforms, in which periodic noise can be detected in the frequency domain and hence reduced by suitable filtering (Figure 11.22).

11.2.3

Band transformations

Band transformations are transformations of multi-band (e.g. multispectral) images. In a normal band transformation, the same operation is performed on each pixel in the image. This operation consists of generating a new pixel value from some mathematical combination of the pixel values of the various bands of the image.

One of the simplest types of band transformation is calculation of a *vegetation index*. This is an operation performed on a red and a near-infrared band of an optical-infrared image. As we discussed in Section 3.6.1, the reflectance of green-leaved vegetation is low in the red part of the spectrum as a consequence of absorption by chlorophyll, and high in the near-infrared region, mostly as a result of multiple scattering in the mesophyll. Reflectance measurements in these regions are thus strongly correlated with the amount of *photosynthetically active radiation* (PAR) absorbed by the plant material, and vegetation indices are designed to exploit this fact.

A vegetation index is calculated from the reflectance r_r in the red band (typically 0.6 to 0.7 μm) and the reflectance r_i in the near-infrared band (typically 0.8 to 1.0 μm). In principle, these reflectances should be calibrated and corrected for atmospheric propagation effects, but in practice it is sometimes acceptable to use the original (uncorrected) pixel values. The simplest vegetation index is the *ratio vegetation index* (RVI), which is defined as

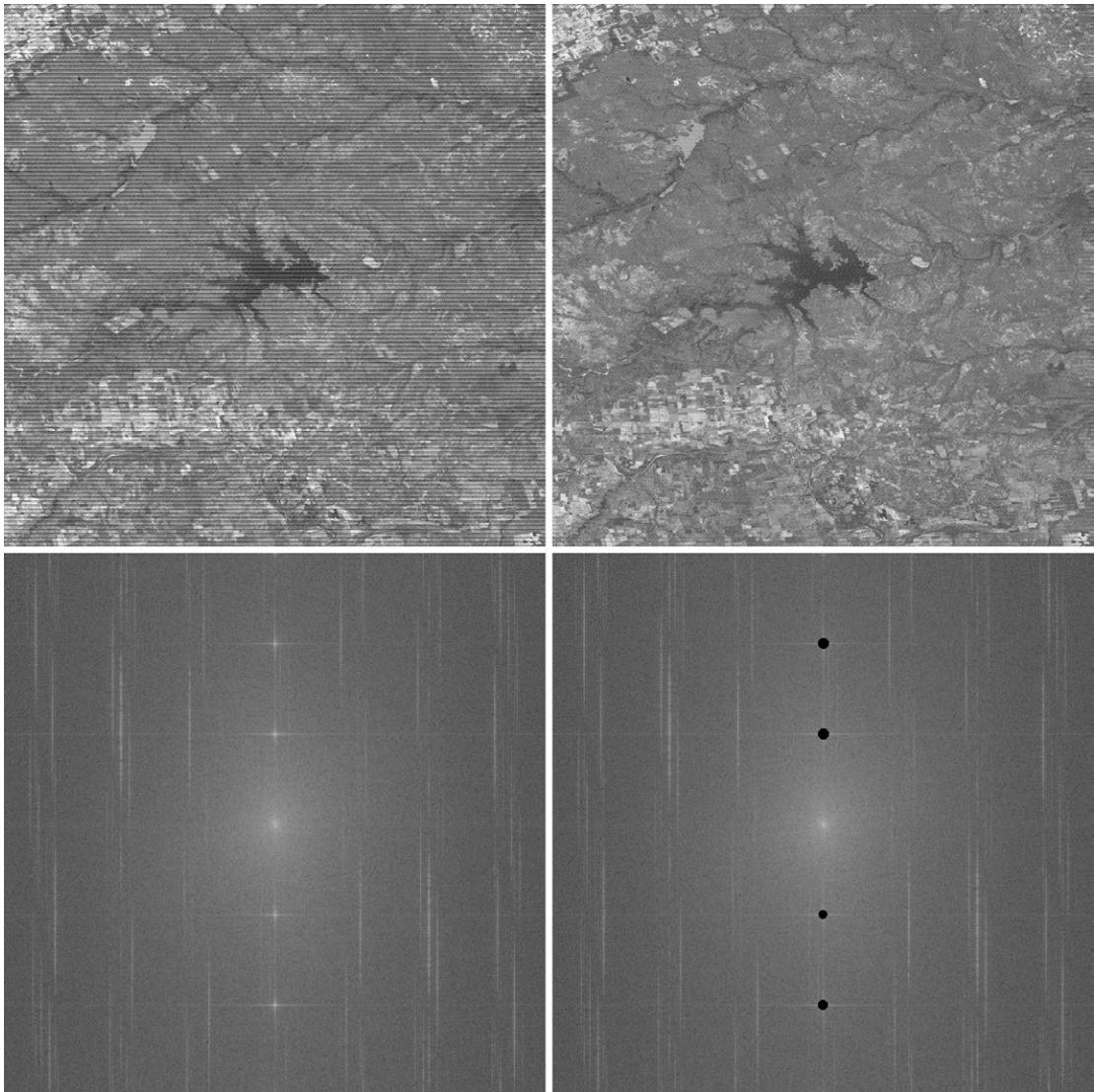


Figure 11.22. Top left: extract of a Landsat MSS image exhibiting severe six-line striping. Bottom left: Fourier transform of the striped image. Note the prominent peaks corresponding to the striping. Bottom right: Fourier transform of the image with the peaks corresponding to the striping manually removed. Top right: inverse Fourier transform of the edited transform.

$$\mathbf{RVI} = \frac{r_i}{r_r}$$

The use of a ratio is preferred to a simple difference, since it is insensitive to influences that change the calculated reflectances by equal factors in both bands. For example, Equation (11.2) is often used to convert the measured spectral radiance to the equivalent

reflectance, but it takes no account of the orientation of the surface. If the surface is tilted towards the Sun, the calculated reflectance will be greater than if the normal to the surface were vertical. However, the use of a reflectance ratio is still somewhat unsatisfactory since it diverges to infinity if the red reflectance is zero. Much more widely used is the *normalised difference vegetation index* (NDVI), defined as

$$NDVI = \frac{r_i - r_r}{r_i + r_r}. \quad (11.12)$$

This is mathematically better behaved, since it can only take values between -1 and $+1$. Many modifications of this concept have been proposed in an attempt to generate an index that is proportional to the (suitably defined) amount of plant material. For example, one common modification is to take account of the reflectance properties of bare soil. Figure 11.23 shows examples of the NDVI.

A more sophisticated approach to estimating the amount of plant material uses more than just the two spectral bands of the vegetation indices. The prototype of such transformations is the *Kauth–Thomas transformation*, which was originally applied to Landsat MSS images of agricultural areas in the USA (Kauth and Thomas 1976). Denoting the pixel values in bands 4 to 7 of an MSS image by d_4 to d_7 respectively, the Kauth–Thomas transformation can be written as

$$b = 0.433 d_4 + 0.632 d_5 + 0.586 d_6 + 0.264 d_7 + 32, \quad (11.13.1)$$

$$g = -0.290 d_4 - 0.562 d_5 + 0.600 d_6 + 0.491 d_7 + 32, \quad (11.13.2)$$

$$y = -0.829 d_4 + 0.522 d_5 - 0.039 d_6 + 0.197 d_7 + 32, \quad (11.13.3)$$

$$n = 0.233 d_4 + 0.012 d_5 - 0.543 d_6 + 0.810 d_7 + 32. \quad (11.13.3)$$

(For historical reasons, the MSS bands of Landsats 1 to 3 were numbered 4 to 7, although they were renumbered as 1 to 4 for Landsats 4 and 5. Their wavelengths are 0.5–0.6, 0.6–0.7, 0.7–0.8 and 0.8–1.1 μm respectively.)

The new variable b is called the ‘brightness’, associated mainly with variations in the soil background. As can be seen from Equation (11.13.1), it is really just a weighted average of the four pixel values. The variable g is the ‘greenness’, associated with variations in green vegetation. Since band 5 of the MSS is a red band, and bands 6 and 7 are near-infrared bands, we can see from Equation (11.13.2) that the greenness variable is roughly similar to a vegetation index. The third variable, y , is the ‘yellowness’, and is associated with the yellowing of senescent variation, and the fourth variable, n , is possibly associated with variations in atmospheric conditions. Since it appears not to be directly associated with vegetation, it has been given the name ‘nonesuch’. The Kauth–Thomas transformation is also known as the *tasseled-cap transformation*, from the shape of the surface traced out in the three-dimensional b - g - y space as vegetation ripens and then senesces (Figure 11.24).

Normalised difference indices similar to the NDVI have been defined for other pairs of wavelengths, responding to characteristic spectral reflectance functions for other materials. For example, the *normalised difference snow index* (NDSI) uses wavelengths of around 0.5 μm and 1.5 μm (Dozier 1989). This index exploits the fact that while the spectral reflectance of snow is very high at visible wavelengths (e.g. shorter than around 0.7 μm), it is very low between about 1.45 μm and 1.65 μm (see Figure 3.32). The NDSI

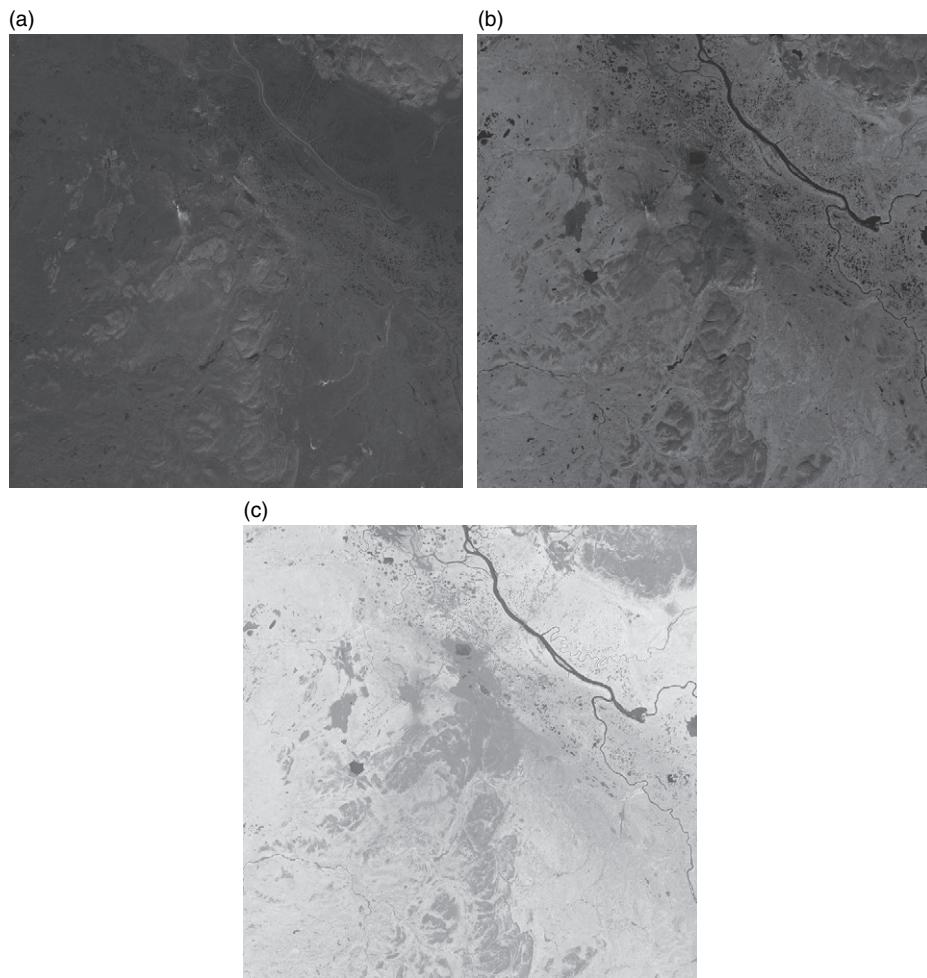


Figure 11.23. (a) Band 3 (red) of part of a Landsat ETM+ image; (b) band 4 (near-infrared); (c) NDVI. The image covers an area 57×57 km centred on the city of Noril'sk in Siberia. The city and its surroundings are heavily polluted (a plume of damaged vegetation can be seen extending to the south-east of the city).

can be used to discriminate between snow cover and cloud cover. Other examples of such indices include the *normalised burn ratio* (NBR), which uses wavelengths around $1\text{ }\mu\text{m}$ and $2\text{ }\mu\text{m}$ and is useful for assessing the severity of burn scars produced by forest fires (Key and Benson 2003), and the *normalised difference water index* (NDWI), which uses wavelengths around $0.9\text{ }\mu\text{m}$ and $1.2\text{ }\mu\text{m}$ and is useful for assessing the water content of vegetation (Gao 1996).

11.2.3.1 The principal and canonical components transformations

The band transformations that we have discussed so far have been based on knowledge of the kind of reflectance spectra expected to be represented in the image. However, it is not

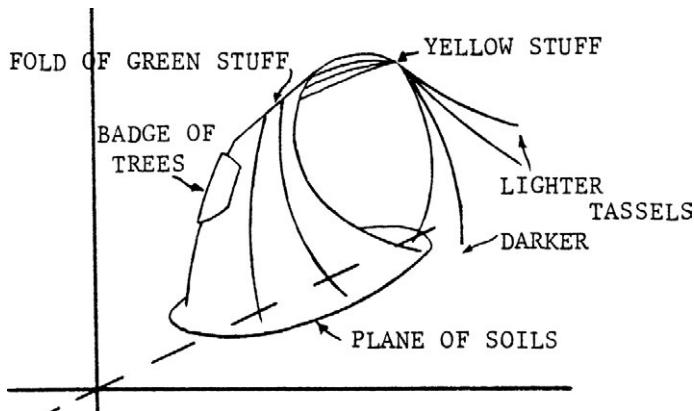


Figure 11.24. Original illustration of the ‘tasseled cap’ (Kauth and Thomas 1976). The brightness coordinate is plotted to the right, greenness is vertical and yellowness into the page. As agricultural crops ripen and then senesce the corresponding point moves from the ‘plane of soils’ towards the fold of green stuff, then to the point marked ‘yellow stuff’, and then into the ‘tassels’.

necessary to have this information available in order to perform a useful band transformation. Instead, we can be guided purely by the statistical properties of the image itself. This is the idea of the principal components transformation and the canonical components transformation.

In general, if we write d_i to represent the pixel value in band i , where i ranges from 1 to N for N -band data, we can define N linear combinations of the bands as follows:

$$\begin{aligned} d'_1 &= a_{11} d_1 + a_{12} d_2 + \cdots + a_{1N} d_N, \\ d'_2 &= a_{21} d_1 + a_{22} d_2 + \cdots + a_{2N} d_N, \end{aligned}$$

and so on. These N equations can be written more compactly in vector-matrix notation as

$$\mathbf{d}' = \mathbf{A}\mathbf{d}, \quad (11.14)$$

where \mathbf{d} is a column vector containing the N original pixel values d_1 to d_N , \mathbf{d}' is the corresponding vector after transformation, and \mathbf{A} is the matrix of coefficients a_{ij} . (In fact, the Kauth–Thomas transform described in the previous section is an example of such a set of linear combinations, apart from the fact that they all have the extra value of 32 added to them for practical convenience.)

The principal components of a multi-band image are the set of linear combinations of the bands that are both independent of, and also uncorrelated with, one another. The concept is easiest to understand in the case where there are only two bands of data. Figure 11.25 illustrates this situation. The upper part of Figure 11.25 shows the original data. It is clear that there is a strong correlation between the band 1 and band 2 values, so that knowledge of the band 1 value of a particular pixel also gives us some information about the likely band 2 value of the same pixel. The lower part of Figure 11.25 shows the same data plotted as the principal components, and now we can see that knowledge of PC1 tells us nothing about the value of PC2.

It is clear that, in order to meet the requirement that the transformed bands d'_1 and d'_2 should be independent of one another, the corresponding axes should be

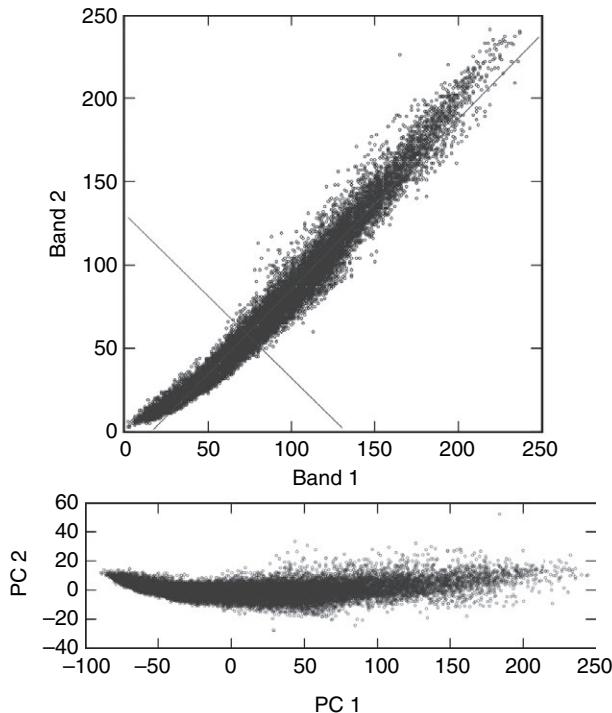


Figure 11.25. Principal components of two-band data. The scatterplot of band 2 pixel values plotted against band 1 pixel values (above) shows strong correlation between the bands. The superimposed lines, plotted through the centroid of the points, show the orientation of the principal components. The lower plot shows the two principal components.

at right angles to one another. Thus we can see that the transformation is just a rotation in the two-dimensional pixel value space. This is clearly apparent in Figure 11.25. The angle of the rotation is determined by the correlation between the values of d_1 and d_2 .

The transformation matrix \mathbf{A} in Equation (11.14) is calculated from the *covariance matrix* (or variance-covariance matrix) of the original data. For N -band data, this is an $N \times N$ matrix \mathbf{C} in which the element C_{ij} is defined as

$$\mathbf{C}_{ij} = \langle (d_i - \langle d_i \rangle)(d_j - \langle d_j \rangle) \rangle, \quad (11.15)$$

where d_i is the pixel value in band i as usual, and the angle brackets denote an average over all pixels. An element C_{ii} on the leading diagonal of the matrix is the variance σ_i^2 of the pixel values in band i , and an off-diagonal term C_{ij} (with $j \neq i$) is related to the correlation coefficient ρ_{ij} between bands i and j through

$$\mathbf{C}_{ij} = \rho_{ij} \sigma_i \sigma_j. \quad (11.16)$$

The matrix \mathbf{A} is calculated from the eigenvectors of the matrix \mathbf{C} . There are N eigenvectors, each satisfying a vector-matrix equation of the form

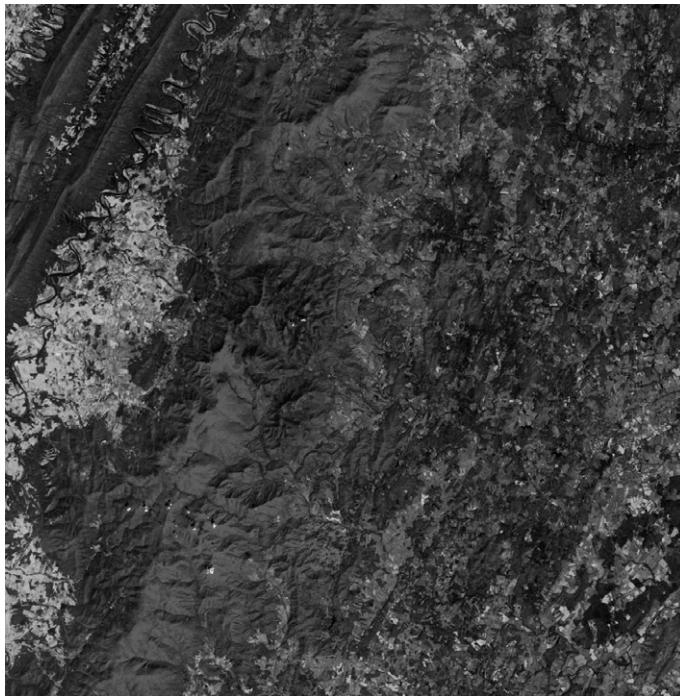


Figure 11.26. A 321 composite ASTER image showing the northern part of the state of Virginia, USA. The sinuous feature at top left is the Shenandoah River. See also colour plates section.

$$\mathbf{Cx} = \alpha \mathbf{x} \quad (11.17)$$

where \mathbf{x} is the eigenvector and α is its corresponding eigenvalue. By convention, the eigenvectors are normalised so that $\mathbf{x}^T \cdot \mathbf{x} = 1$ (\mathbf{x}^T is the transpose of \mathbf{x}) and arranged in order such that the first eigenvector has the largest eigenvalue, the second eigenvector has the next largest eigenvalue, and so on. With this definition, the matrix element a_{ij} is just the j th element of the i th eigenvector.

A principal components transformation (PCT), also known as a *Karhunen–Loëve transformation* or a *Hotelling transformation*, involves replacing the N bands d_1 to d_N of a multi-band image by the corresponding principal components d'_1 to d'_{N} . In the case of imagery having more than three bands, it will often occur that the first three principal components will contain a large percentage (e.g. 95% or more) of the total image variance. Thus the effective dimensionality of the image data has been reduced by the transformation. By assigning these three components to, say, the red, green and blue channels of a colour display, an image can be displayed in only three bands that contains most of the information present in the original multi-band image. It may also happen that one or more of the principal components will correlate with some physically meaningful variable, as is the case with VNIR images of vegetated areas.

PCT is illustrated in the following example. The original image consists of bands 1, 2 and 3 of an ASTER image showing the northern part of the state of Virginia, USA. These are shown in Figure 11.26 as a 321 RGB composite.

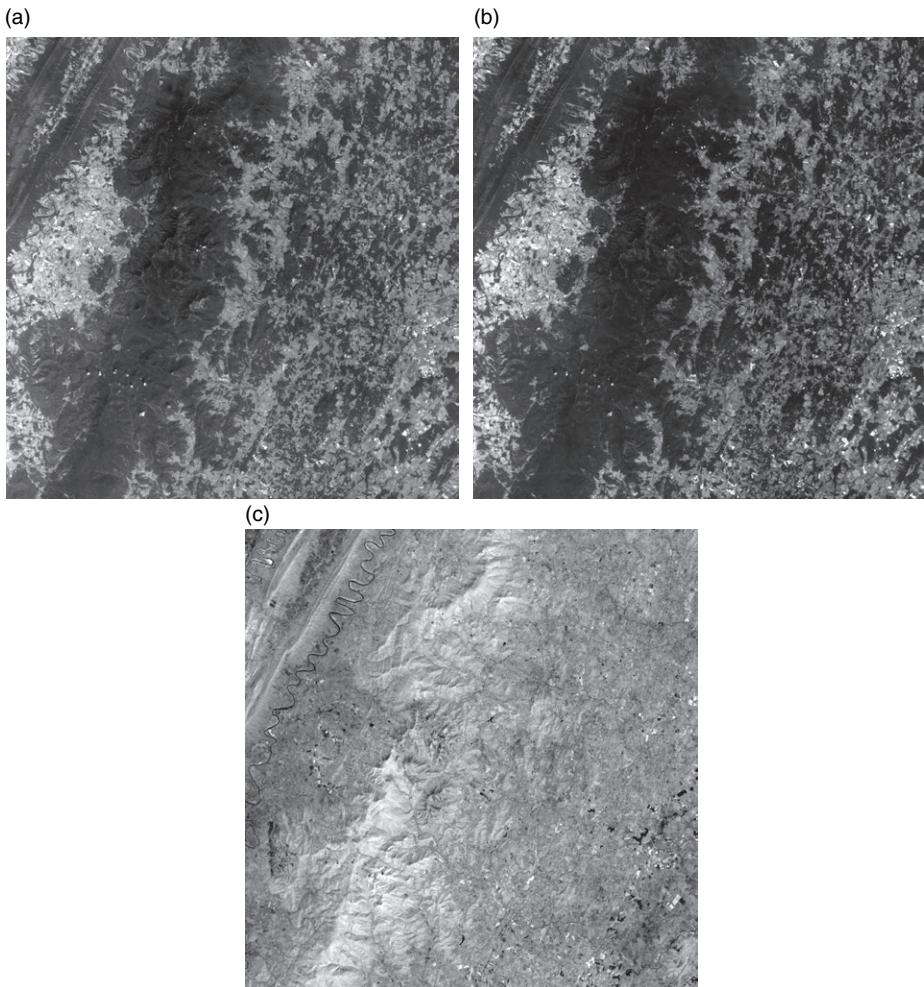


Figure 11.27. Greyscale representations of bands 1 to 3 of Figure 11.26.

Figure 11.27 shows greyscale representations of the individual bands. The variance-covariance matrix for these three bands is:

73.0552	78.8196	-19.3145
78.8196	90.8385	-31.6864
-19.3145	-31.6864	130.8638

This shows, for example, that the variance in band 2 is 90.84 units, while the covariance between bands 1 and 3 is -19.31 units (i.e. they are negatively correlated). The three eigenvalues of the variance-covariance matrix are the three columns of this matrix:

-0.5385	0.4103	0.7359
-0.6333	0.3791	-0.6747
0.5558	0.8294	-0.0557

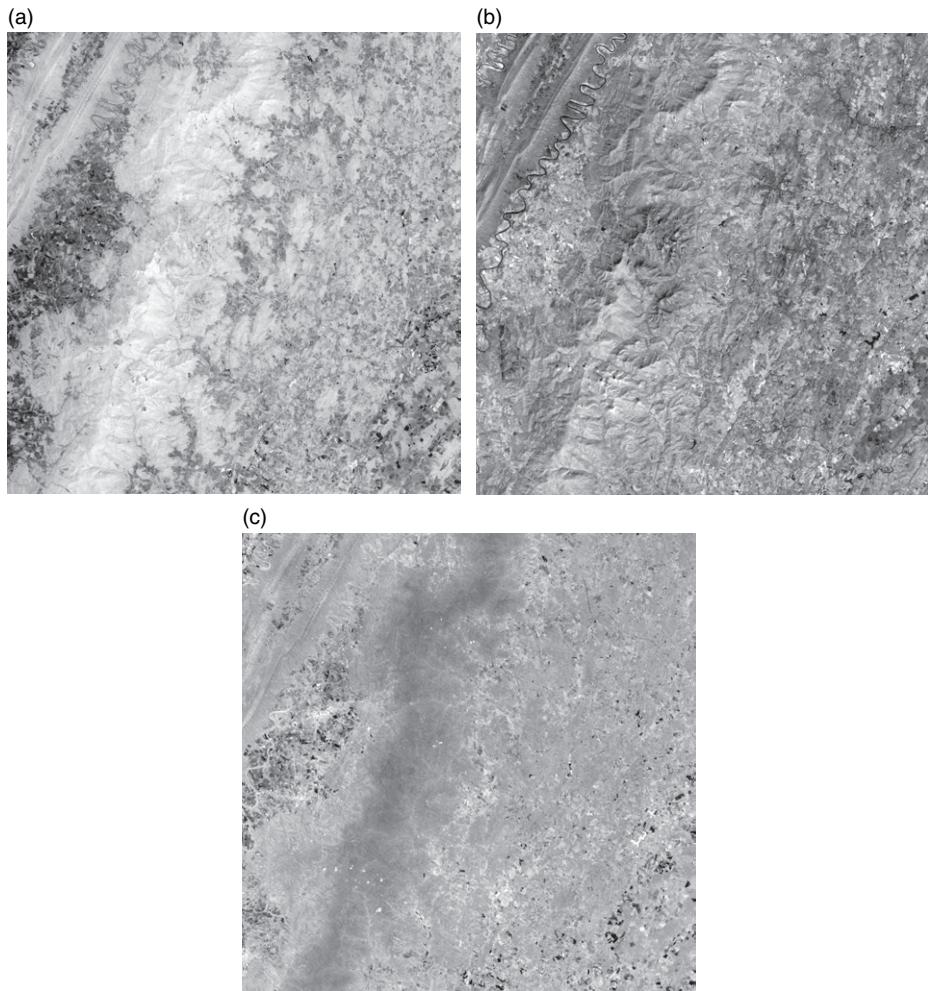


Figure 11.28. Principal components of Figure 11.26.

with eigenvalues of 185.68, 106.83 and 2.25 respectively. (These results can easily be evaluated using *Octave*, for example). Thus the first principal component is given by

$$P_1 = -0.5385 B_1 - 0.6333 B_2 + 0.5558 B_3$$

and it accounts for a fraction

$$\frac{185.68}{185.68 + 106.83 + 2.25} = 63.0\%$$

of the total variance. The first two principal components account for 99.2% of the total variance.

The three principal components are shown in Figure 11.28. They have all been contrast-stretched, otherwise almost no variation would be discernible in P_3 .

As a check, we can calculate the variance-covariance matrix for these three principal components. It is

185.68	0.00	0.00
0.00	106.83	0.00
0.00	0.00	2.25

as expected. The off-diagonal zeroes show that the principal components are uncorrelated with one another, and we say that the variance-covariance matrix has been *diagonalised*.

The *canonical components transformation* (CCT) is a similar idea to the principal components transformation. In the latter, the principal components are arranged so that the image variance is greatest in the first principal component. In the former, the transformation takes account of clustering in the data, and the components are chosen such that the separability (see Section 11.3.2) of the clusters is greatest in the first canonical component. We shall not discuss this further here. The interested reader is referred to, for example, the book by Richards and Jia (2006) for more details.

Summary

A digital image can be considered as a two- or three-dimensional array of digital numbers or pixel values. The first two dimensions are the row and column number within the image while the third dimension is the band number in the case of a multi-band image. Radiometric correction of the image involves converting the digital numbers into physically meaningful quantities such as the at-satellite radiance or the surface reflectance. Georeferencing is the process of establishing the relationship between the row and column coordinates of a pixel and the corresponding position on the Earth's surface. Once this relationship is known it can be used to reproject the image into some particular map projection. This is likely to involve resampling of the image, which may require pixel values to be spatially interpolated.

The intelligibility of an image may be enhanced by contrast modification, which changes the radiometric but not the spatial properties of the image. Pixel values are adjusted in order to change the image histogram in some desired manner. Spatial filtering, on the other hand, alters the value of a pixel depending on neighbouring pixel values. Spatial filtering can be used to reduce the noise in an image (at the expense of its spatial resolution), to sharpen it (at the expense of increased noise), to detect edges and to remove periodic artefacts.

Band transformations combine pixel values from two or more bands. The normalised difference vegetation index (NDVI) is a mathematically transformed ratio of the reflectance in the near-infrared to the reflectance in the red part of the spectrum, and is strongly correlated with the amount of green-leaved vegetation. Similar normalised difference indices have been defined using other pairs of wavebands chosen to respond to particular characteristics of reflectance spectra. The principal components transform (PCT) replaces the pixel values in n spectral bands by n linear combinations of the pixel values, such that each linear combination is uncorrelated with any of the others. This is a purely mathematical operation, requiring no decisions to be made by the data analyst, and it has the advantage of reducing the dimensionality of the data. The tasseled-cap transformation is essentially similar to the PCT but is optimised for the study of agricultural crops.

11.3

Image classification

Image classification is the process of making quantitative decisions from image data, grouping pixels or regions of the image into classes intended to represent different physical objects or types. The output of the classification process may be regarded as a

11.3 Image classification

thematic map rather than as an image. The majority of classification techniques use mainly the radiometric data (pixel values) present in the image, with little or no reference to spatial variation, and these techniques will be described first. These techniques can be thought of as follows. Suppose we have an n -band image, and the pixel values in each band can take k different values (for example, $k = 256$ for 8-bit data). The number of possible coordinates in the n -dimensional pixel-value space is k^n , a number that can very easily exceed a million (it is over 16 million for the case where $n = 3$ and $k = 256$). However, it is extremely unlikely that the image represents a million or more different classes of data, or that we could make use of the information if it did. What we require is some simplification of the data in the n -dimensional pixel value space, identifying a volume within this space as representing a single class of data.

11.3.1

Density slicing and pseudocolour display

Density slicing is a particularly simple classification technique, applied to a single-band image or to one band of a multi-band image. A range of input pixel values (a ‘slice’ of the image) is mapped to a single pixel value in the output image. Several such ranges can be defined, of course. If each of these is mapped to a different colour in the output image, the result is called a *pseudocolour image* (see Figure 11.29). Density slicing is commonly used where the pixel values have a direct relationship to a physical variable. For example, the pixel values in a calibrated thermal infrared image of an ocean area may correspond directly to the sea surface temperature, or one might wish to slice up a vegetation index image into different ranges of the vegetation index, as in Figure 11.29. In this case, values of NDVI below zero have been mapped to the same blue colour to suggest water, since in most cases the reason for the low values of NDVI is indeed the presence of water. Another common use of density slicing is to generate image masks, for example land/water masks, which can be used to define areas of the image to which subsequent processing is to be applied.

11.3.2

Multispectral classification

Multispectral classification can be approached from two fundamentally different directions. The first of these, *supervised classification*, uses information about the known distribution of classes to initiate the process. From field investigations or ancillary data (e.g. maps), the image class is already known in some areas of the image. This knowledge is used to define *training data*, i.e. a statistical description of the range of pixel values, for each of the classes of interest. The entire image is then examined, pixel by pixel, to determine to which of the classes, if any, a pixel belongs. This process obviously requires some kind of rule to decide to which class a pixel belongs. The simplest is the *box classifier* or *parallelepiped classifier*. In this case, n -dimensional (n is the number of bands in the image) boxes are defined so that they enclose all, or a high proportion, of the training data for each class. The rule is then very simple: if a pixel is within a particular box, it is assigned to the corresponding class. This is illustrated schematically for two-band data in Figure 11.30.

Although the box classifier is exceptionally easy to apply, Figure 11.30 shows two of its disadvantages. While the pixel at a should obviously be assigned to class A , the pixel at b could be assigned to either A or B , and the pixel at c cannot be assigned at all.

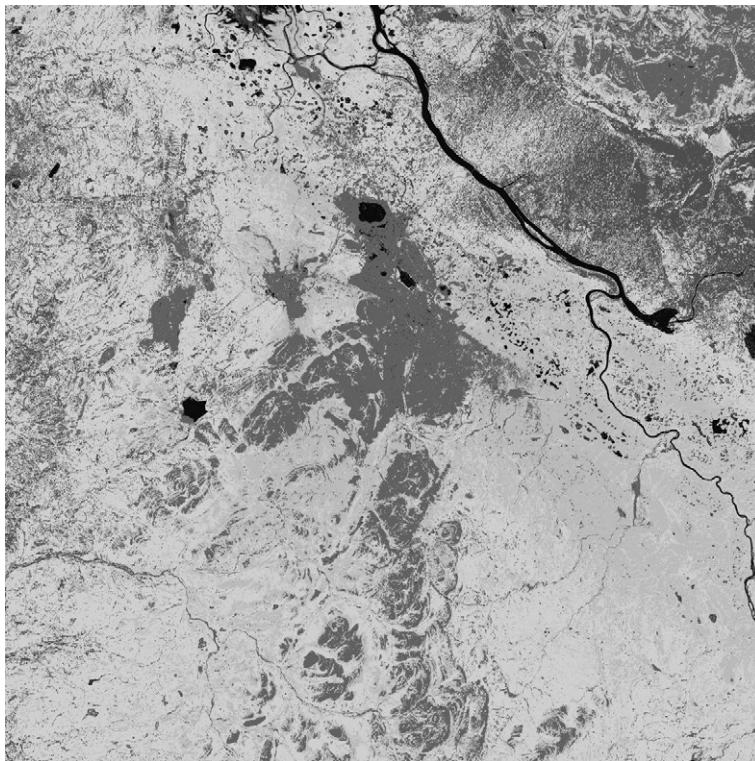


Figure 11.29. Pseudocolour representation of the NDVI image of Figure 11.23c. Blue denotes values NDVI below 0, grey between 0 and 0.25, orange between 0.25 and 0.5, light green between 0.5 and 0.7 and dark green above 0.7. The choice of colours suggests the distribution of water (blue) and pollution-damaged vegetation (orange). See also colour plates section.

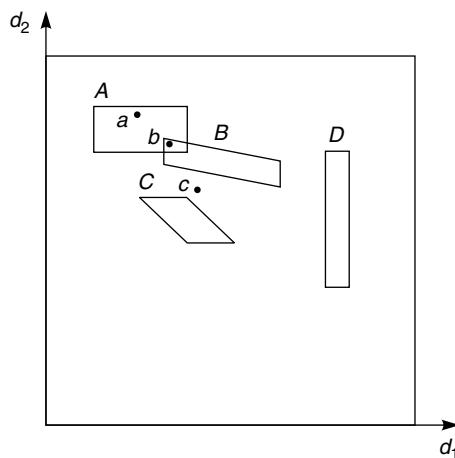


Figure 11.30. Schematic illustration of the box classifier for two-band data. d_1 and d_2 are the pixel values in bands 1 and 2 respectively, and the enclosing square shows the total ‘volume’ of pixel-value space that is available. The boxes A, B, C and D have been constructed around the training data for the corresponding classes; the points a, b and c are three pixels to be classified.

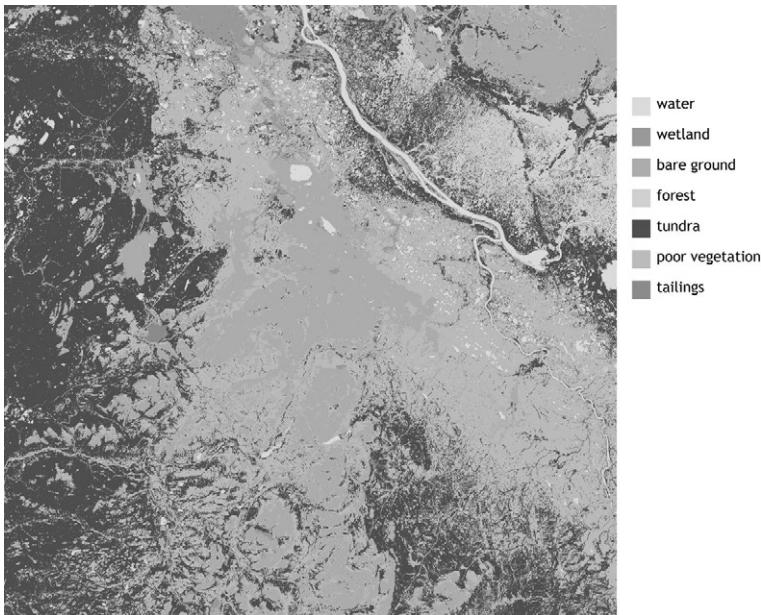


Figure 11.31. Supervised classification of the image of Figure 11.23c. See also colour plates section.

More sophisticated classification algorithms use a *discriminant function*, which quantifies how well a pixel ‘fits’ a given class. For example, the *Euclidean distance* algorithm calculates the distance, in the n -dimensional pixel-value space, from a given pixel to the mean position (centroid) of the training data for each class. The smallest distance determines the class to which the pixel will be assigned. Thus, referring again to Figure 11.30, pixels a and b will both be assigned to class A , and pixel c will be assigned to class C . The *maximum likelihood algorithm* in effect models the probability distributions for each class, using the training data, from which it is possible to estimate the likelihood that a given pixel belongs to a particular class. The most probable assignment can then be made. This approach also has the advantage that a probability threshold can be imposed, so that no classification at all will be made if the maximum likelihood of belonging to any class is below some critical value. Figure 11.31 shows a simple example of a classified image.

The opposite approach to supervised classification is *unsupervised classification*. In this case, the entire image is first analysed without reference to any training data. The aim of this analysis is to identify distinguishable clusters of data in the n -dimensional pixel-value space. Once these clusters have been identified, they can be associated with physical classes using training data.

The first step in an unsupervised classification is to cluster the data. Various clustering algorithms exist. One of the most widely used is the *isodata* algorithm (Duda and Hart 1973), also known as the *K-means* or *migrating means* algorithm. This is an iterative procedure, in which the user first specifies the number of clusters to be found. The algorithm assigns nominal centre coordinates to each of these clusters in the

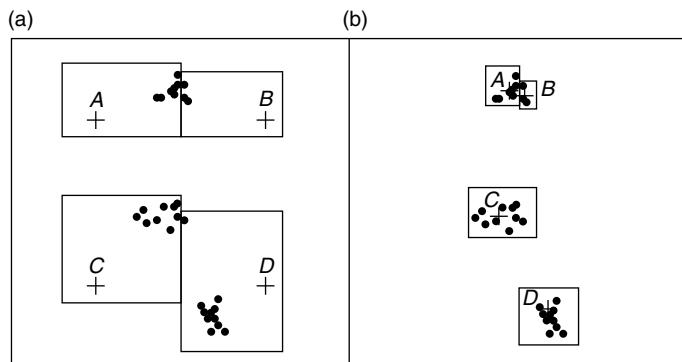


Figure 11.32. Schematic illustration of the operation of the isodata clustering algorithm. The enclosing square boxes represent the n -dimensional pixel-value space, the dots are the coordinates of individual pixels, and the crosses are the cluster means. (a) The initial cluster means, and (b) the cluster means after one iteration.

n -dimensional pixel-value space. Next, every pixel in the image is then assigned to the appropriate cluster, using a discriminant function such as the Euclidean distance. The centre coordinates of the clusters are then recalculated from the mean values of the pixels assigned to them, and the process is repeated. It is normally terminated when fewer than some specified proportion (for example, 2%) of the pixels have their assignments changed since the algorithm does not usually converge.

Figure 11.32 illustrates the operation of the isodata algorithm. In Figure 11.32a, the four cluster means A to D have been assigned arbitrarily, and the consequent assignment of the pixels to the clusters is shown by the rectangular boxes. In Figure 11.32b, the cluster means have ‘migrated’ to their new positions, giving the new assignments shown by the rectangular boxes. In fact, in this very simple example, the process has converged completely after only one iteration, since no further changes in the assignment of pixels to the clusters occurs with subsequent iterations. In addition to setting a threshold for convergence, it is necessary to specify the number of clusters and to choose their initial means. This can be done essentially manually, by specifying training areas (in which case it is a form of supervised classification), or by making an initial statistical assessment of the data. Otherwise, the process is automatic.

In essence, supervised classification forces the image classes to correspond to physical classes that are defined by the user, but without a guarantee that these classes will be statistically distinct. Unsupervised classification, on the other hand, forces the image classes to be statistically distinct but does not guarantee that they will correspond to the physical classes required by the user. *Hybrid classification* combines both approaches. In this technique, an unsupervised classification is first carried out to determine the number of distinguishable clusters present in the image. These are then compared with the training data, and clusters are split, merged or deleted as necessary. It may also be necessary to modify the definitions of the physical classes, for example by merging two physical classes that cannot be discriminated in the image. The process is normally an iterative one.

An important aspect of this process is the ability to quantify the extent to which two clusters are statistically separable. Various measures of separability can be used, all of

11.3 Image classification

which are defined from the probability distributions $p_1(\mathbf{d})$ and $p_2(\mathbf{d})$ of the two clusters in the n -dimensional pixel-value space. (The vector \mathbf{d} defines the coordinates of a pixel in this space.) For example, the *divergence* d is defined as

$$d = \int \left(p_1(\mathbf{d}) - p_2(\mathbf{d}) \right) \ln \frac{p_1(\mathbf{d})}{p_2(\mathbf{d})} d\mathbf{d}, \quad (11.18)$$

where the integral is carried out over the entire pixel-value space. This is zero for identical distributions, and infinite for distributions that do not overlap at all and hence are completely separable. The *transformed divergence*

$$d^* = 2 \left(1 - \exp(-d/8) \right) \quad (11.19)$$

is also zero for identical distributions, but has a maximum value of 2 for non-overlapping distributions. As a rough guide for real data, the divergence d should exceed about 10 for reasonable separability. A third commonly used measure of separability is the *Jeffries–Matusita distance*

$$J = \int \left(\sqrt{p_1(\mathbf{d})} - \sqrt{p_2(\mathbf{d})} \right)^2 d\mathbf{d}, \quad (11.20)$$

which, like the transformed divergence, ranges from 0 for identical distributions to 2 for non-overlapping distributions. Finally, we mention the *Bhattacharyya distance*. This is calculated from the means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ and the covariance matrices \mathbf{C}_1 and \mathbf{C}_2 of the two distributions in feature space:

$$B = \frac{1}{4} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' (\mathbf{C}_1 + \mathbf{C}_2)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + \frac{1}{2} \ln \frac{|\mathbf{C}_1 + \mathbf{C}_2|}{2\sqrt{|\mathbf{C}_1||\mathbf{C}_2|}}. \quad (11.21)$$

In this expression, \mathbf{a}' , $|\mathbf{a}|$ and \mathbf{a}^{-1} denote respectively the transpose, determinant and matrix inverse of \mathbf{a} . Like the divergence, this takes a value of zero for identical distributions and is infinite for non-overlapping distributions. For normally distributed data,

$$J = 2(1 - e^{-B}). \quad (11.22)$$

As a final remark before leaving the subject of multispectral classification, we can observe that although we have been assuming throughout this section that a pixel is characterised by its coordinates in the n -dimensional pixel-value space, this is not a fundamental necessity for image classification. Other properties can be used to characterise a pixel. These could be, for example, radar backscatter coefficients in different polarisation states, texture parameters (see Section 11.3.6), or single-band pixel values from different dates. The essential aspect is that the pixel is characterised by more than one variable, and so can be described in a multi-dimensional *feature space*.

11.3.3 Hyperspectral classification

Hyperspectral images, as noted in Chapter 6, record data in an order of magnitude more wavebands than multispectral images, and feature spaces can have numbers of dimensions as high as several hundred. This can cause some difficulties for the usual techniques of

multispectral classification and techniques for classification of hyperspectral imagery focus on the intelligent reduction of the dimensionality of the data. principal component analysis (PCA: see Section 11.2.3.1) can be used for this, as can the minimum noise fraction (MNF) transformation (Green *et al.* 1998), which is a rather similar idea. Both of these approaches can be used to estimate the effective dimensionality of the data.

Multispectral classification can be characterised as the identification of target materials by their reflectance spectra, and this is more obviously true in the case of hyperspectral imagery where the reflectance spectrum is measured with much higher spectral resolution. A number of techniques have been developed for quantifying the similarity between a spectrum measured in the imagery and a reference spectrum (which could be derived from the imagery itself, or from field or laboratory measurements). One approach is to match specific absorption features in the spectra. Another is termed *spectral angle mapping* (SAM). In SAM, the spectra are considered as vectors in the hyperdimensional feature space, and the degree of similarity between them is represented by the smallness of the angle between the vectors. Specifically, if \mathbf{u} and \mathbf{v} are the two vectors, the angle θ between them is given by

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}| |\mathbf{v}|}, \quad (11.23)$$

so that the larger the value of $\cos\theta$, the more similar are the spectra. This definition is insensitive to changes in illumination.

11.3.4 Advanced classification methods

As we noted in the previous section, an essential aspect of any classifier is a rule for making decisions about which class to assign to a given pixel. Simple examples were provided by the box classifier and the minimum distance method, while more sophisticated rules were characterised through the use of a discriminant function such as the maximum likelihood method. Bayesian statistics can also helpfully be employed (Campbell 2008). However, some particularly powerful new methods have emerged recently. One of these is the use of *artificial neural networks* (ANNs, sometimes called *perceptrons*). These are computer procedures intended to mimic the way that brains learn by reinforcing the links between particular neurons. An ANN has an *input layer*, an *output layer*, and a number of *processing layers* (or *hidden layers*) between them. This is represented schematically in Figure 11.33.

In this example, there are eight nodes in the input layer. These could represent, for example, the digital numbers representing a pixel in eight spectral bands of the image. The four nodes in the output layer could represent the probabilities that the pixel should be assigned to each of four possible classes. The value assigned to each node in the ANN is some function (this is the *transfer function*) of the values in the nodes connected to it in the layer above, and the essential feature of the ANN is that these functions are adjustable. For example, a three-input node might be represented by the function

$$\frac{a_0}{1 + \exp(-a_1 x_1 - a_2 x_2 - a_3 x_3 - a_4)},$$

where the x s are the input values and the a s are adjustable parameters. The network is trained through a process termed *back-propagation* to determine the best fitting values of

11.3 Image classification

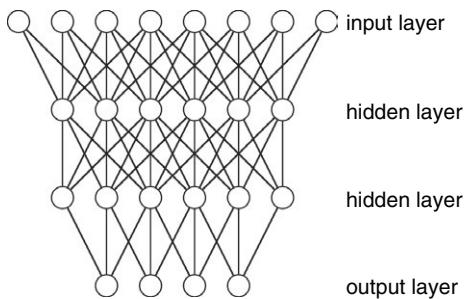


Figure 11.33. Schematic artificial neural network.

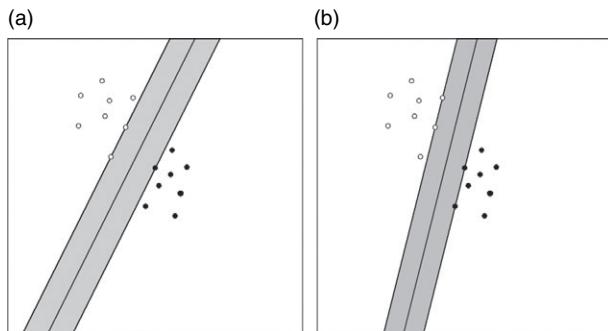


Figure 11.34. The two sets of training data are better separated by the line on the left than by the line on the right.

the parameters, usually by iteration. The power of ANNs is that the strength of the links between the nodes is determined empirically through back-propagation rather than being assumed at the outset. The disadvantages of ANNs are that they are computationally intensive to run and they need a lot of training data.

Another development, like the ANN also a learning algorithm, is the support vector machine (SVM) (Cortes and Vapnik 1995). In its simplest form this is a means of finding the optimum separation between two classes. This is illustrated schematically in Figure 11.34. Sets of training data representing two different classes are shown by black and white points plotted in feature space. In Figure 11.34a, the sets of points have been separated by a line such that all the white points are to the left of it and all the black points are to the right. In Figure 11.34b the line also separates the points, but the margin between the sets of points, shown in grey, is narrower than in Figure 11.34a. The optimum separation is the one that gives the widest margin.

The optimum separating line is found iteratively. Once it has been determined from the training data it can be used as a classifier for other data. In the case of feature spaces that have more than two dimensions, the separator is a hyperplane rather than a line, but the concept is the same. Refinements of the basic idea of the SVM allow for separation between more than two classes of data, and for non-linear surfaces between them.

11.3.5

Sub-pixel classification

The classification methods discussed in Sections 11.3.2–11.3.4 can be classed as *per-pixel classifiers*, in the sense that every pixel of the image is assigned to just one of a number of defined classes. It is implicitly assumed that the contents of each pixel are homogeneous. However, *mixed pixels* are common, especially in the case of large pixel sizes such as those from sensors such as MODIS. Sub-pixel classification is an attempt to assign more than one piece of information to the description of the contents of each pixel.

The most basic type of sub-pixel classification is *mixture modelling*, which estimates the proportions of different classes present in the pixel. In its simplest form, we assume that each class contributes independently to the reflectance of the pixel, so that the combined effect of the various classes is the sum of the contribution from each class independently. This is *linear mixture modelling*. To make this definite, we suppose that the image has n bands of data, so that the reflectance of each pixel is represented by n numbers r_i . We also suppose that there are m different classes into which we wish to classify the image, and that the reflectance in spectral band i of class j is c_{ij} . The fraction of class j present in a pixel is written as f_j . In all of these notations, i takes values between 1 and n and j takes values between 1 and m . The assumption of linearity (independence) is equivalent to assuming that

$$r_i = \sum_{j=1}^m c_{ij} f_j,$$

which can be written more compactly in matrix-vector notation as

$$\mathbf{r} = \mathbf{c}\mathbf{f}, \quad (11.24)$$

where \mathbf{r} is an n -element column vector, \mathbf{f} is an m -element column vector and \mathbf{c} is a matrix having n rows and m columns. This equation describes the linear mixing of the classes: given knowledge of \mathbf{c} and \mathbf{f} , the value of \mathbf{r} can be calculated. However, what we actually want to do is *unmixing*, in which we calculate \mathbf{f} from \mathbf{c} and \mathbf{r} .

In fact, the situation is slightly more complicated than this. First, we allow for the possibility of random errors (arising because of sensor noise etc.) by adding an error term to (11.24):

$$\mathbf{r} = \mathbf{c}\mathbf{f} + \mathbf{e}, \quad (11.25)$$

where \mathbf{e} is an n -element column vector denoting the errors, and second, we note that there is a constraint on the solution for \mathbf{f} , which is that all the fractions must sum to 1:

$$\sum_{j=1}^m f_j = 1 \quad (11.26)$$

The mathematical task is now to find, for each pixel, the value of \mathbf{f} that minimises the errors in (11.25) while satisfying (11.26). There are several methods to do this, but they are somewhat beyond the scope of this book.

The spectra that correspond to the pure classes are often referred to as *end-members*, and the matrix \mathbf{c} as the *end-member matrix*. Various techniques exist to identify the

11.3 Image classification

end-members from an image, and they can also be informed by reflectance data collected in the field. However, it is evident that unmixing can only be performed where $n \geq m$, i.e. where the number of spectral bands is at least as great as the number of classes. For this reason, unmixing methods are particularly promising when applied to hyperspectral images.

11.3.6 Texture classification

Up to this point, we have given no consideration to the spatial context of a pixel: the techniques we have discussed would lead to the same decisions being made about the class to which a pixel belonged regardless of its location within the image or its relation to its neighbours. In this section we consider the concept of image texture, while in Section 11.3.8 we consider contextual classification generally.

Although a precise definition of image texture is rather difficult to formulate, it can be loosely defined as structure in the spatial variation of the pixel values. Everyday experience, and especially the consideration of black-and-white photographs, tells us that image texture is an important aspect of the way in which the human brain interprets a scene. Our task is to define quantitative measures of image texture that correspond to its perception by the brain, or at least have some utility in differentiating between different regions of an image. Many texture measures have been proposed. The simplest is the variance of the pixel values (or similarly the range between the maximum and minimum values) within a small neighbourhood of the pixel in question. This does not, however, provide any information on the spatial scale of the variations. Scale-dependent information can be obtained in various ways, for example by calculating the Fourier transform of a small neighbourhood in the image and then extracting the coefficients for different spatial frequencies or simply by calculating the variance as a function of a varying window size. However, one of the commonest approaches to the quantification of image texture is through the use of the *grey-level cooccurrence matrix* (GLCM), also known as the *spatial dependency matrix*.

If the pixel values in the image are drawn from a set of N integers (e.g. $N = 256$ for 8-bit data), the GLCM is an $N \times N$ square matrix \mathbf{P} . It is calculated as follows. First, we choose some spatial separation to define the position of a pixel relative to some reference pixel. For example, the vector $(1, 0)$ would define the pixel immediately to the right of the reference pixel. Next we define a neighbourhood of pixels within which the GLCM is to be calculated. The matrix element P_{ij} is then the number of times that pixel value i occurs in a pixel within the neighbourhood *and* pixel value j occurs at the chosen separation from this first pixel. The elements are normally expressed as proportions of the total, i.e. the sum of all the matrix elements is 1. Once the GLCM has been calculated, various texture parameters can be derived from it. Some of the commonest are the *energy*, defined as

$$\sum \mathbf{P}_{ij}^2,$$

the *entropy*, defined as

$$\sum \mathbf{P}_{ij} \ln \mathbf{P}_{ij}$$

and the *contrast*, defined as

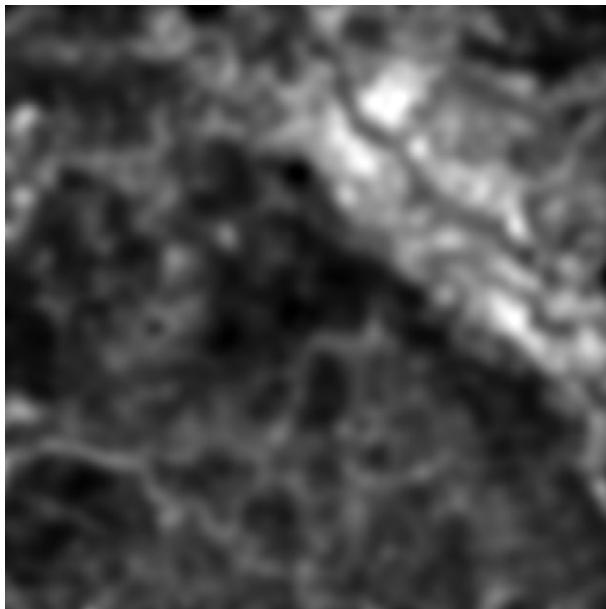


Figure 11.35. Texture image corresponding to Figure 11.23b. The area of high texture towards the top right of the image corresponds to an area of many small lakes.

$$\sum |i - j|^a \mathbf{P}_{ij}^b,$$

where a and b are often taken as 2 and 1 respectively. In all of these definitions, the sum is taken over all the matrix elements.

Since image texture is a spatial property, any texture parameter must necessarily be defined by examining variations of the pixel values within some finite neighbourhood of the image. The texture parameter associated with a particular pixel is calculated from the pixel values in a ‘window’ of the image centred on that pixel, and this has the effect of degrading the spatial resolution. The choice of window size is therefore important: if it is too small, too few pixels will be available to give a statistically meaningful measure of texture; whereas if it is too large, the resolution of the ‘texture image’ will be unnecessarily degraded. Figure 11.35 shows a typical texture image.

11.3.7

Error matrices and classification accuracy

An *error matrix*, also known as a *confusion matrix*, is a method of quantifying the performance of a classification (Foody 2002). It is a square matrix \mathbf{E} of $N \times N$ elements, where N is the number of classes in the classified image. The element E_{ij} in row i and column j is the number of pixels known to belong to class i and classified as belonging to class j . Thus the elements on the leading diagonal, E_{ii} , correspond to correctly classified pixels, and all the off-diagonal elements correspond to erroneous classifications. It is typically produced in a process similar to the use of training data to train the classification, for example by reserving some of the training data to test the accuracy. As an example, Figure 11.36 shows an error matrix for a nine-class classification.

		Classified as									Row total	Producer's accuracy (%)
		bare	urban	tailings	water	forest	tundra	poor veg	waste	wetland		
Known class	bare	6628	1115	0	1	0	1	436	1857	563	10601	62.5
	urban	470	972	1	23	0	1	19	514	286	2286	42.5
	tailings	0	0	1076	0	0	0	0	0	0	1076	100.0
	water	8	17	0	4519	0	0	1	11	26	4582	98.6
	forest	0	0	0	0	1917	176	0	0	0	2093	91.6
	tundra	4	0	0	0	1973	22420	334	0	8	24739	90.6
	poor veg	180	40	0	1	0	91	4801	0	230	5343	89.9
	waste	30	27	0	0	0	0	0	865	29	951	91.0
	wetland	125	371	0	50	0	29	1136	129	6231	8071	77.2
Column total		7445	2542	1077	4594	3890	22718	6727	3376	7373	59742	
Consumer's accuracy (%)		89.0	38.2	99.9	98.4	49.3	98.7	71.4	25.6	84.5		

Figure 11.36. An error matrix.

From the figure, we see that $E_{21} = 470$, for example, meaning that 470 pixels known to belong to class 2 (urban) were erroneously identified as belonging to class 1 (bare ground).

The error matrix can be used to calculate various performance figures for the classification. The *consumer's accuracy* (also called the *user's accuracy* or the *reliability accuracy*) for class i is defined as

$$\frac{E_{ii}}{\sum_{j=1}^N E_{ji}}.$$

Thus, for example, the fact that $E_{11} = 6628$ tells us that 6628 pixels that were known to belong to class 1 (bare ground) were correctly classified as bare ground, while the fact that $\sum_{j=1}^9 E_{j1} = 7445$ tells us that the total number of pixels classified as belonging to class 1 was 7445. Thus, provided that the pixels used to calculate the error matrix are sufficiently representative, we may conclude that the probability that a pixel that is classified as bare ground actually is bare ground is $6628/7445 = 89.0\%$. We also define the *producer's accuracy* (also called the *reference accuracy*) for class i as

$$\frac{\mathbf{E}_{ii}}{\sum_{j=1}^N \mathbf{E}_{ij}}.$$

The denominator of this fraction is now a row total rather than a column total, so that (to continue our example based on class 1) the fact that $\sum_{j=1}^9 \mathbf{E}_{1j} = 10601$ tells us that the total number of pixels actually belonging to class 1 is 10601. Thus, the probability that a pixel known to belong to class 1 will be classified as belonging to class 1 is $6628/10601 = 62.5\%$.

As we noted earlier, the off-diagonal elements of the error matrix quantify errors in the classification. They also suggest strategies for improving the accuracy. For example, Figure 11.26 shows that the consumer's accuracy of the 'urban' class is very low, but also that those pixels incorrectly classified as urban are most likely to belong to the 'bare ground' class. We may thus try to improve the discrimination between these two classes, perhaps by collecting more training data or using a different classifier, or we may simply decide to combine the two classes.

Although the performance of the classification is specified by the error matrix, it can also be convenient to derive single-parameter classification accuracies from it. The simplest and most widely used of these is just the proportion of pixels that are correctly classified, i.e. the sum of all the terms along the leading diagonal of the matrix (the *trace*) of the matrix divided by the sum of all elements:

$$A = \frac{\sum_{i=1}^N \mathbf{E}_{ii}}{\sum_{i=1}^N \sum_{j=1}^N \mathbf{E}_{ij}} \quad (11.27)$$

For our example data, this has a value of $49429/59742 = 82.7\%$. However, this is a potentially rather misleading statement of the accuracy of the classification since it ignores the fact that correct classifications may have arisen by chance. A number of techniques have been developed to deal with this problem. One of the simplest is the kappa statistic (Cohen 1960). This involves calculating the equivalent accuracy A^* for the case where the pixels are classified at random, while preserving the row and column totals. This gives (Rees 2008)

$$A^* = \frac{\sum_{i=1}^N \sum_{k=1}^N \mathbf{E}_{ik} \sum_{k=1}^N \mathbf{E}_{ki}}{\left(\sum_{i=1}^N \sum_{j=1}^N \mathbf{E}_{ij}\right)^2} \quad (11.28)$$

and the kappa statistic is then calculated as

$$\kappa = \frac{A - A^*}{1 - A^*}. \quad (11.29)$$

Thus if $A = A^*$ and the classification is no better than chance, $\kappa = 0$, while if $A = 1$, $\kappa = 1$. For the example in Figure 11.26 we have $A^* = 21.7\%$ and hence $\kappa = 77.9\%$.

It is highly unlikely that the correct classification will be known for every pixel in an image (otherwise there would have been no need to perform the classification in the first place), and some care must be exercised in choosing the pixels from which the error matrix is calculated. The proportions of pixels known to belong to the various image classes should correspond as closely as possible to the proportions for the image as a whole. A random sampling strategy is usually preferred as a way to try to achieve this.

Summary

Image classification is the process of assigning pixels or groups of pixels to membership of a defined number of classes, and hence converting the image into a map. Classification can be per-pixel, in which only the pixel values of a given pixel are used to determine its class, or it can also make use of spatial information.

Multispectral classification can be supervised or unsupervised. A supervised classification is trained using reference data, typically obtained from fieldwork, to identify the typical ranges of pixel values in each spectral band corresponding to each class. A decision rule such as the maximum likelihood classifier is then applied to determine the class membership of every pixel in the image. Advanced algorithms for determining decision rules include the use of artificial neural networks and support vector machines. The success of supervised classification depends on the similarity of the distributions in feature space of the training data for the various classes. Several mathematical measures of similarity are available to quantify this. In unsupervised classification the data are first grouped into clusters in feature space, which is a largely automatic process, and the clusters are then associated with physically meaningful classes. The K-means algorithm is a particularly common clustering algorithm. Unsupervised and supervised classifications can be combined in hybrid classification. Hyperspectral classification, where the feature space may have a dimensionality of a hundred or more, can be approached through reducing the dimensionality of the dataset (for example through PCT) or by using techniques to compare the similarity of a spectrum defined by the data to a set of reference spectra. Sub-pixel classification most commonly involves ‘spectral unmixing’, in which are estimated the proportions of different classes contributing to the spectrum measured at a particular pixel. This approach is best suited to hyperspectral data. Image texture, which is not easy to define, can also be used as a feature on the basis of which classification can be performed.

The accuracy of an image classification can be assessed using an error matrix, which compares the number of pixels known (from reference data) to belong to particular classes to the number of pixels assigned to those classes by the classification method. The error matrix provides a convenient way of identifying cases where the classifier has difficulty in discriminating between classes. It can be summarised into producer’s and consumer’s accuracies for each class, and further summarised into an overall accuracy of the classification. This accuracy can be misleadingly inflated by chance agreements, and the kappa statistic, amongst other approaches, has been developed to allow for this.

11.4

Image segmentation and detection of geometrical features

11.4.1

Segmentation

The aim of segmentation is to find homogeneous contiguous regions within an image, as a means of simplifying the geometrical description and perhaps as a precursor to recognising objects within the image. There are many approaches to image segmentation and it is an area of intense research (Neubert, Herold and Meinel 2006).

One rather obvious way to segment an image is to find edges within it, and to interpret these as the boundaries of segments. This approach can be less straightforward than it appears, however, since the boundary of an image segment must be continuous while the edges in the image may be discontinuous (Figure 11.37). Image classification or clustering provides another starting point for segmentation since it reduces the complexity of the image. At the simplest level, density-slicing (thresholding) can be used to define a region or a set of distinct regions and this shares some similarities with edge-detection (Figure 11.38). Multispectral classification can be used in a similar way, and the isodata algorithm is commonly employed for this purpose. One disadvantage of this approach is that a ‘noisy’ classification, arising from noisy image data, can result in very many small segments, perhaps only a pixel or two in size, that may not be meaningful. A simple way of reducing this problem is to apply a *majority filter* to the classified image. Here, a small neighbourhood, typically 3×3 pixels, is investigated around each pixel in the classified image. The classification of the central pixel in this neighbourhood is reassigned to whichever image class is most strongly represented in the neighbourhood. This procedure usually reduces the level of noise in the classification. Multispectral classification methods can also take spatial context into account. An example of this approach is the ECHO (extraction and classification of homogeneous objects) technique (Kettig and Landgrebe 1976), which takes into account the spectral properties of a pixel’s neighbours as well as its own spectrum (Figure 11.39).

Image segmentation can also be performed by *region-growing* (Espindola *et al.* 2006). A *seed pixel*, or small group of contiguous pixels, is chosen, and its neighbours are examined to determine whether they are sufficiently similar to the region to be added to it. This process is continued, adding sufficiently similar adjacent pixels, until all the pixels that adjoin the group are insufficiently similar to be added to it. The process then terminates and the region is complete. Another region can then be seeded and grown. The criterion of similarity might be, for example, that the Euclidean distance in feature space from the coordinates of the pixel to be added to the coordinates of the centroid of the region should be less than some specified amount.

Region-growing methods proceed from small to large regions. The opposite approach is *split-and-merge*, in which a large region (perhaps the whole image) is first tested for homogeneity. If it passes the test, the process terminates since it has already found a homogeneous region. However, if it fails the test, the region is then split into four (or some other number) of sub-regions. Each of these is tested for homogeneity. Any that pass the test of homogeneity are then also tested to identify whether they should be merged with adjacent sub-regions, while any sub-regions that fail the test of homogeneity are

11.4 Image segmentation and detection of geometrical features

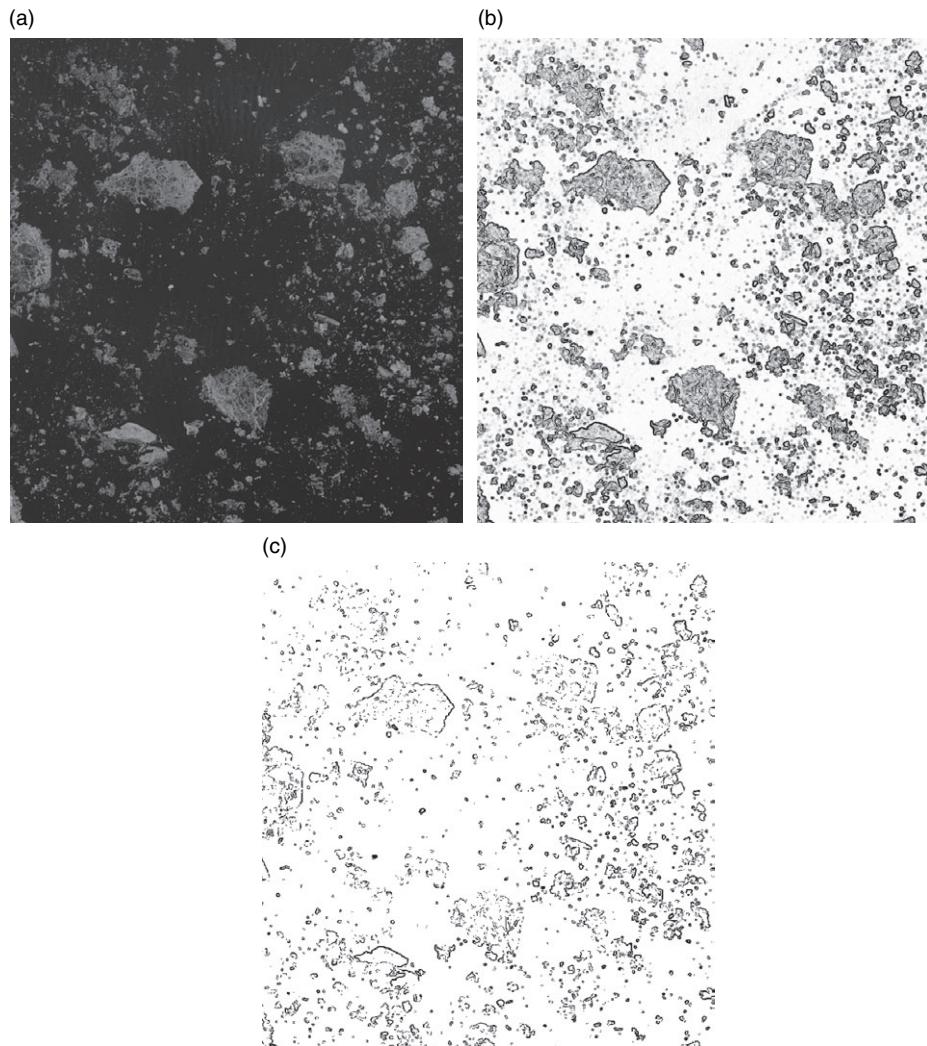


Figure 11.37. (a) Extract of band 4 of a Landsat ETM+ image showing a region of sea ice north of Alaska. (b) Effect of applying an edge-detection filter on the image. (c) Retaining only the strongest edges.

further subdivided. The advantage of this method of segmentation is that it can be terminated at some specified resolution which is larger than a single pixel.

11.4.2 Detecting shapes

The recognition and classification of objects in an image on the basis of their shape is a high-level operation in image processing. The edge-detection filters discussed in Section 11.2.2 are simple examples of shape detection, and other convolution operators can be used to detect particular shapes, in an approach usually referred to as *template matching*.

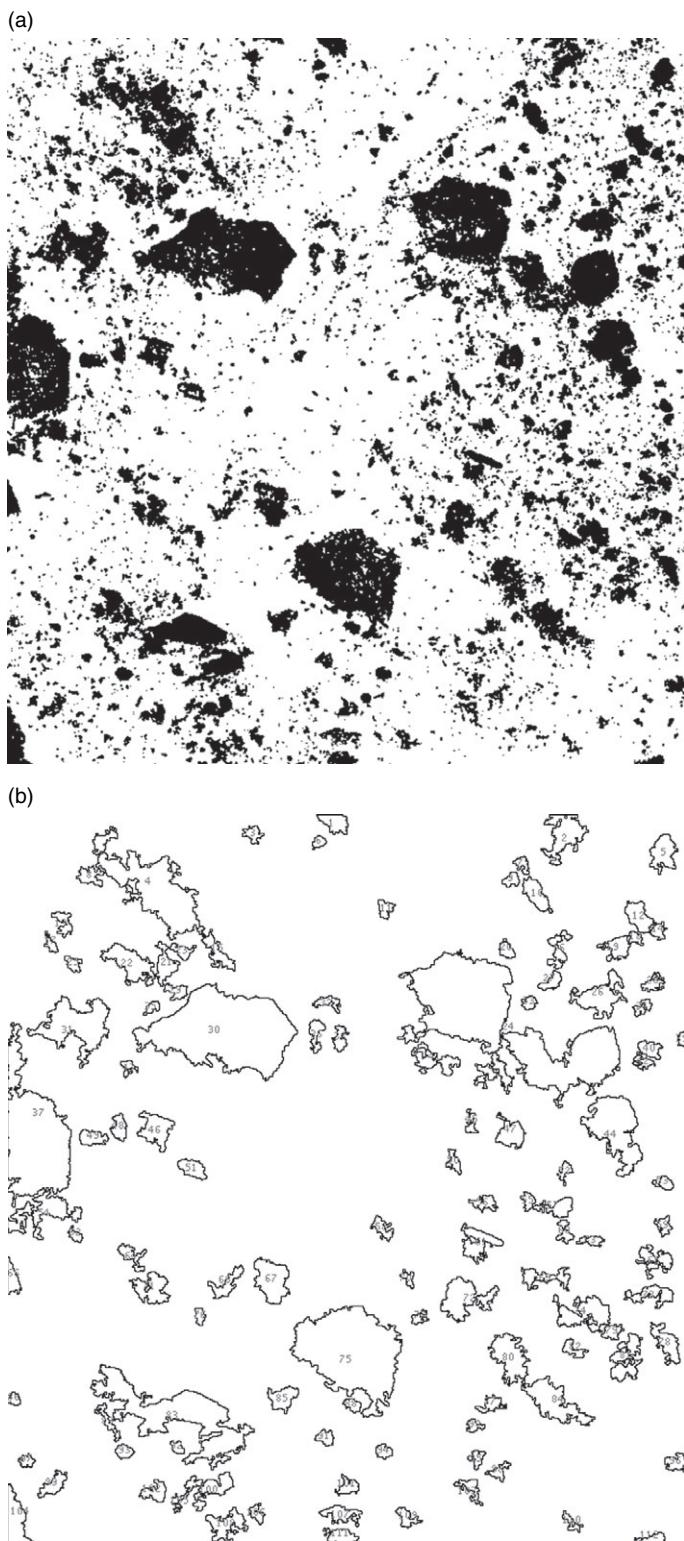
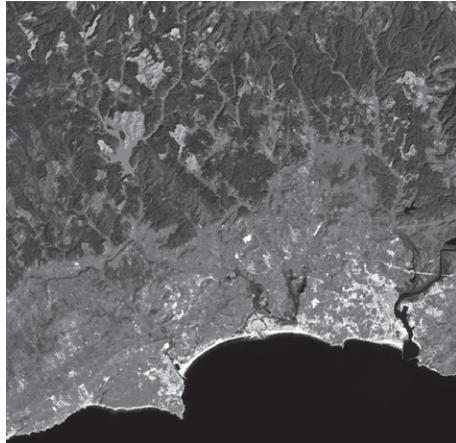


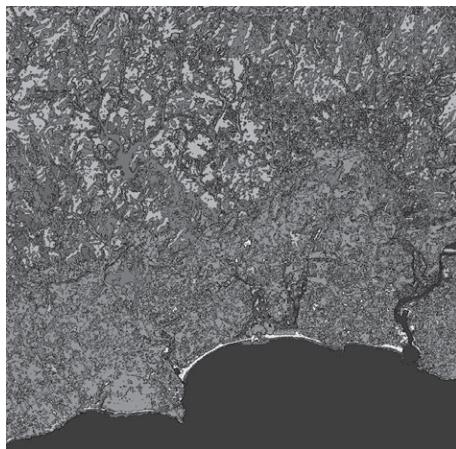
Figure 11.38. (a) The image of Figure 11.37a after thresholding and inverting the look-up table. (b) Segmentation by detection of continuous boundaries. Only ‘particles’ (ice floes in this example) larger than 100 pixels have been enumerated here.

11.4 Image segmentation and detection of geometrical features

(a)



(b)



(c)

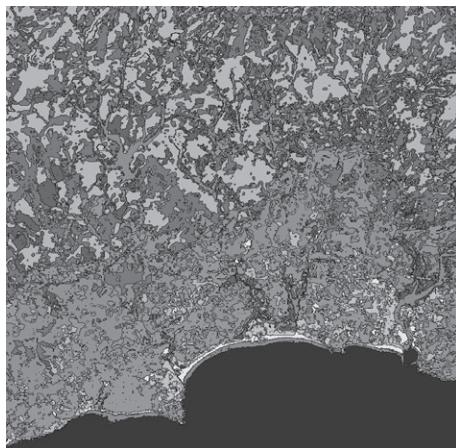


Figure 11.39. (a) 8.6 km square extract of a FCIR Landsat image centred on the Barragém de Bravura, Portugal. (b) Segmentation into ten classes using minimum-distance multispectral classification. (c) Segmentation into ten classes using ECHO. See also colour plates section.

-1	-1	-1	-1	2	-1
2	2	2	-1	2	-1
-1	-1	-1	-1	2	-1
-1	-1	2	2	-1	-1
-1	2	-1	-1	-1	2
2	-1	-1	-1	-1	-1

Figure 11.40. Templates (convolution kernels) for detection of linear features a single pixel in width.

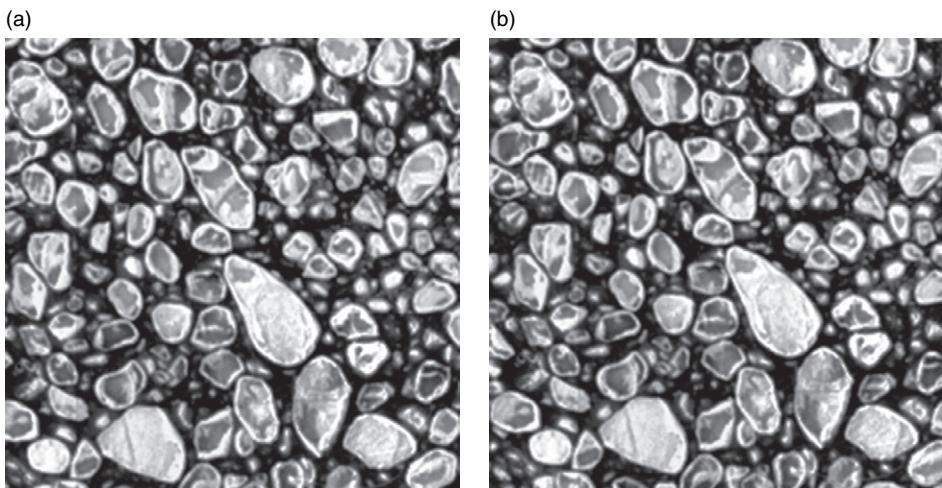


Figure 11.41. Two images differing only by a spatial displacement. The images are aerial photographs showing a form of sea-ice known as pancake ice.

The convolution kernel is designed in such a way that, if it is applied to an object having the desired shape, it will produce a maximum output. As a simple example, linear features that are a single pixel wide can be detected using the four directional templates shown in Figure 11.40. One possible method of using these templates would be to apply each of them to the original image data, then to add the squares of the outputs, and finally to threshold the result.

An alternative approach to the detection of objects having known shape, size and orientation but unknown position is through the use of Fourier transforms. This can be illustrated by considering the problem represented in Figure 11.41.

11.4 Image segmentation and detection of geometrical features

The two images 11.41a and 11.42b contain the same features with the same relative positions, but there is an unknown spatial displacement between them which we wish to determine. Clearly one could in principle define a template corresponding to Figure 11.41a and move it around over Figure 11.41b to see where it best fits, but the template would be large and the procedure therefore time-consuming. Alternatively, one simply calculates the Fourier transforms of the two images and multiplies them together. The resulting product, when retransformed back from the spatial frequency domain into the spatial domain, is just the cross-correlation function of the two images, and it will have a maximum at the image coordinates corresponding to the displacement between the two images (Figure 11.42).

A third approach to the detection of objects on the basis of their shape is through the use of *Hough transforms*. These can usefully be applied even when the size and orientation of the object are unknown, as long as they can be suitably parameterised. The simplest example of the use of the Hough transform is in the detection of straight lines, illustrated in Figure 11.43.

Figure 11.43a shows an image of width and height 100 pixels. The image contains a straight line, and P_1 to P_3 are three representative pixels on this line. The position and orientation of the line can be specified by the parameters r and θ , where r is its perpendicular distance from the origin and θ is the angle between the perpendicular and the x -axis, as shown in the figure. The coordinates (x, y) of any point on the line will satisfy the equation

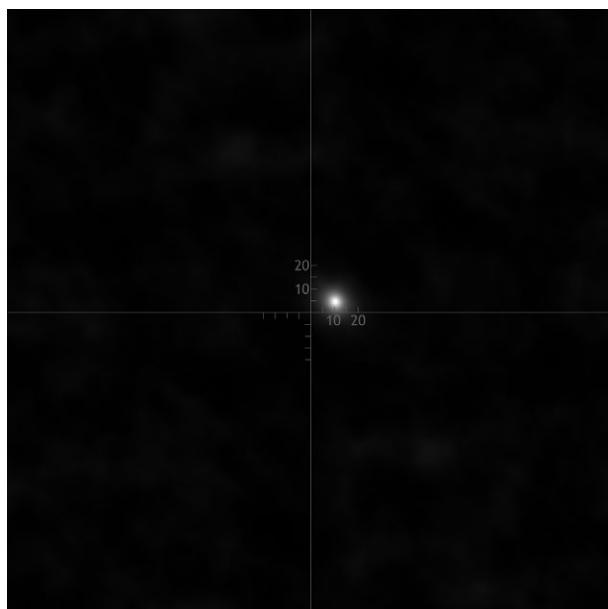


Figure 11.42. Spatial correlation between the two images of Figure 11.41. The superimposed axes show that the relative displacement between the two images is 10 pixels horizontally and 5 pixels vertically.

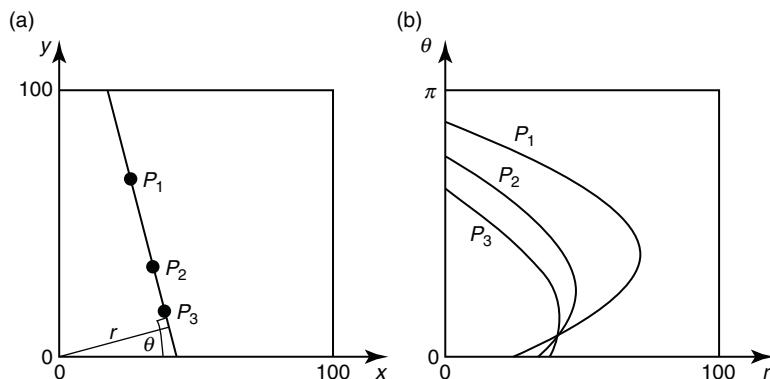


Figure 11.43. Illustration of the Hough transform for straight lines.

$$x \cos \theta + y \sin \theta = r. \quad (11.30)$$

The purpose of the Hough transform is to determine the values of r and θ most appropriate to a group of pixels such as those represented by P_1 to P_3 . Figure 11.43b shows a plot in r - θ space. The curve labelled P_1 is the set of all points in this space that satisfy Equation (11.21) for the (x, y) coordinates of the point P_1 , and similarly for the other two points. As expected, the three curves in Figure 11.43b intersect at a single point, corresponding to the values of r and θ shown in Figure 11.43a.

In practice, the Hough transform is calculated using an ‘accumulator array’. This is a two-dimensional array representing r - θ space, and initially all its members are set to zero. The procedure scans the image to identify the pixels to be transformed (these will often be derived by performing a line- or edge-detection on the original image). For the (x, y) coordinates of every such pixel, the contents of the accumulator array at (r, θ) are incremented by 1 for every combination of r and θ that satisfies Equation (11.30). Once this procedure has been completed, the accumulator array is scanned to find the maxima that correspond to straight lines in the original image (Figure 11.44).

Other Hough transforms can be defined for shapes that can be parameterised analogously to Equation (11.30). For example, a circle can be defined through the three parameters a , b and r :

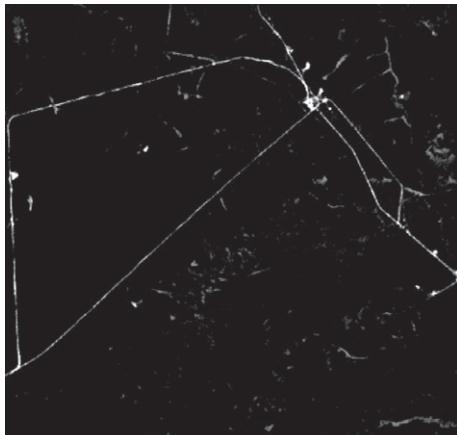
$$(x - a)^2 + (y - b)^2 = r^2,$$

where a and b are the coordinates of the centre and r is the radius. The accumulator array in this case is three-dimensional. A fuller discussion of Hough transforms can be found in Sonka, Hlavac and Boyle (2007).

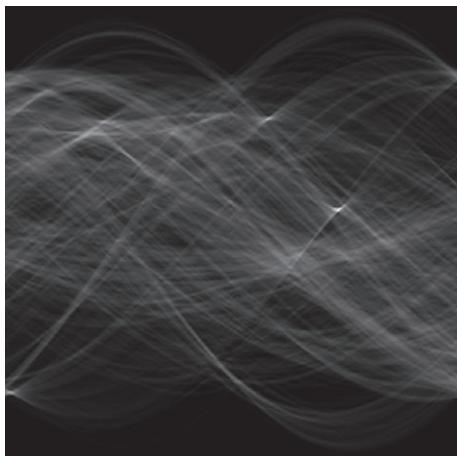
As a final approach to the detection of geometrical features, we mention the use of image segmentation (Section 11.3.8.1). Once this procedure has been carried out, shape factors (e.g. the parameters of the best-fitting ellipse, the length of the boundary, the circularity and so on) can be calculated for the region and used to characterise its shape.

11.4 Image segmentation and detection of geometrical features

(a)



(b)



(c)



Figure 11.44. (a) Original image (extract of a SPOT image showing oil pipelines in the Komi Republic, Russia). (b) Accumulator array. (c) Original image with the three most prominent detected straight lines superimposed on it. (Hough transform implemented using a Java plugin (Burger and Burge 2005) for ImageJ.) See also colour plates section.

Summary

Segmentation of an image involves finding homogeneous regions within it as a way of simplifying its geometrical description. This can be done by finding edges within the image, but noisy data mean that these edges may not form closed curves and hence define regions. The effect of noise in the data can be reduced by density-slicing, or by classifying the image. This is more effective if the spatial variability in the classification is reduced, for example by majority-filtering the output or by using a classifier, such as ECHO, that takes spatial context into account. Segmentation can also be performed by ‘growing’ homogeneous regions from seed pixels, adding new pixels if they are sufficiently similar to those already included in the region, or conversely by split-and-merge techniques that progressively subdivide a region until it is sufficiently homogeneous.

Particular shapes in images can be detected by template-matching or by the use of Hough transforms. A region of one image that is similar to a region of another image can be identified by cross-correlation of the two images, which can be performed efficiently using Fourier transforms.

11.5

Geographic information systems

A geographic information system (GIS) is a ‘geospatially aware’ database, i.e. one that makes use of georeferencing information (Longley *et al.* 2005). The term GIS may also include the hardware (computer platform) on which the database is supported (sometimes the definition includes the data, and the user), but the essential features are the ability to store, retrieve and present geospatial data, to perform quantitative operations on the data, and to recognise and make use of topological relationships between different geospatial datasets. It is, in fact, a kind of map, and the development of GIS evolved from computer cartography. Remote sensing is, as we have seen, a means of collecting and analysing geospatial data. There is clearly therefore a substantial degree of overlap between the aims of remote sensing and of GIS, but we may perhaps regard the former as a means to the latter, although a GIS may well contain data obtained from sources other than remote sensing.

There are several ways in which we can characterise how GIS represents the world. Clearly the GIS must have an unambiguous way of defining the location of a point on the Earth’s surface. We represent this generally by the coordinates (x, y) which may specify, for example, the longitude and latitude according to some datum, or the metric coordinates (i.e. specified in length units such as metres) according to some map projection.

Now that we are able to specify the location of an object, we can consider the different kinds of object that can be represented. At the most basic level, these are points, lines, polygons, surfaces and networks. A *point* is single location, representing something that has no spatial extent (like the intersection of graticule lines in a map) or whose spatial extent is of no importance. A *line* is specified by the coordinates of its two ends, and perhaps also by intermediate points if it is not straight. A line might be used to represent a

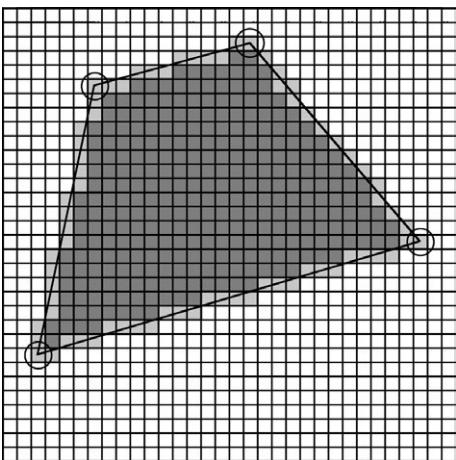


Figure 11.45. Illustration of the difference between raster and vector representation.

The polygon is defined in vector format by the locations of the four nodes, shown as circles. In raster format, the polygon's boundary is represented by the light grey cells and the whole polygon by the light and dark grey cells.

feature such as a political boundary or a road. A *polygon*, also called an *area*, is a region bounded by a closed line that does not cross itself. This could be used to define the coastline of an island, for example. A *surface* represents the value of a third coordinate z as a function of the spatial coordinates x and y . A digital elevation model (DEM) is an example of a surface, the z coordinate corresponding to elevation, but so is a representation of the spatial variation of temperature, or population density, and so also is an image of the Earth's surface (in which case z represents the image radiance). Finally, a *network* is a system of more than one connected line, the connections being termed *nodes*. This is a suitable description of a river or road system. All of these objects can have *attributes* that describe or quantify them.

As we have already seen, the fundamental spatial unit of an image is a pixel, and larger spatial structures within an image are represented as sets of pixels. This is termed a *raster* representation of location: the Earth's surface, or the part of it that concerns us, is imagined to be divided into a grid of (usually) square cells and all objects are specified in terms of these cells. The alternative approach is *vector* representation, in which position data are stored as sets of (x, y) coordinates. Vector form generally gives more compact data storage, and is more natural for points, lines and networks (Figure 11.45). It also has the advantage that it does not impose a fixed spatial resolution on the data. On the other hand, the raster format is natural for images and other surfaces and for performing topological operations on areas (e.g. finding where two areas overlap). In general, GISs more commonly use vector representation of coordinates while image-processing software uses raster representation, although there is often overlap between the two and raster-to-vector and vector-to-raster conversions will often need to be made.

Perhaps the most important way in which a GIS represents the world is through *layers* or *themes*. These allow queries to be made using algebraic operations on the layers in a

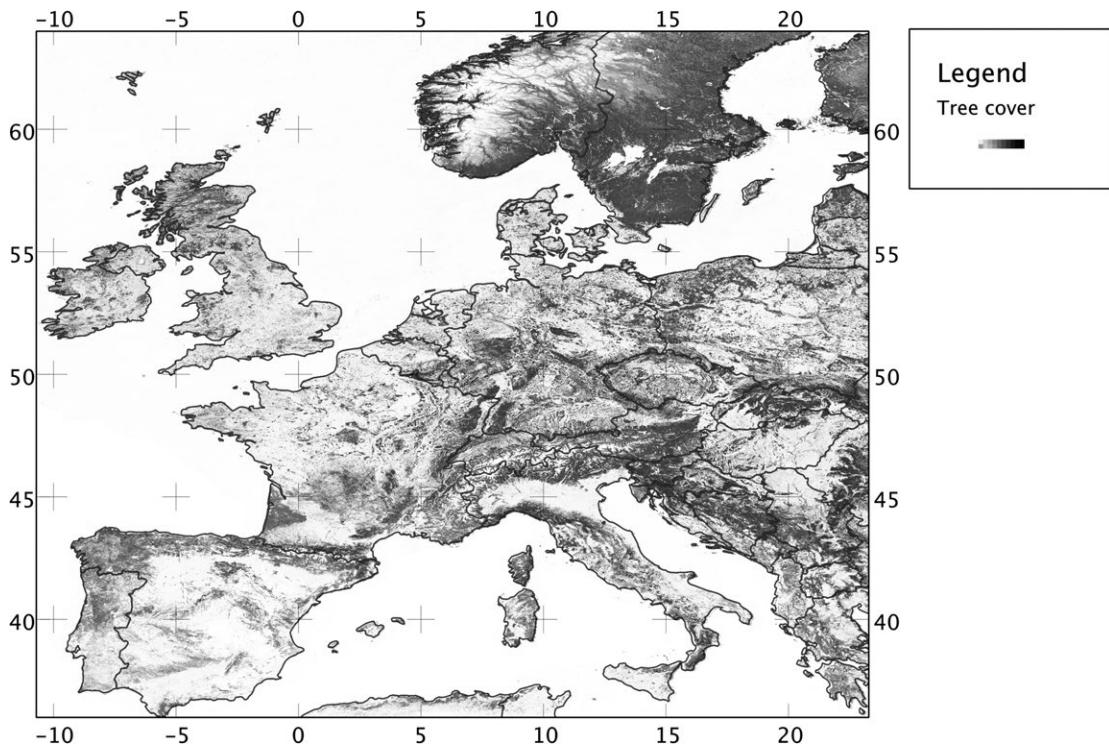


Figure 11.46. Output of a simple GIS. The GIS has two layers: polygons (areas) denoting national boundaries, and a surface (raster layer) representing the density of tree cover. Map algebra applied to this GIS would allow the total forested area to be determined for each country represented in the GIS.

process termed *map algebra*. As a very simple example, we could imagine a GIS consisting of two layers, one being a map of forested areas (quite possibly derived from classification of a remotely sensed image) and the other being a set of national boundaries represented as areas. A query to this GIS would allow us to evaluate the total forested area in each country represented in the GIS (Figure 11.46).

Summary

A geographic information system (GIS) is a georeferenced database that can store, retrieve, manipulate and output geospatial data. Characteristics of a GIS are its ability to represent geometrically defined objects such as points, lines, polygons, networks and surfaces, and to represent several layers of data and be able to combine information from the layers using map algebra. Most GISs make use of vector representation of data where it is meaningful to do so.

11.6

Image formats and data compression

11.6.1

Image compression

Digital images are stored in a wide range of formats generally, and this is also true of images used in remote sensing. In Section 11.2 we discussed the data volumes necessary to specify images, and noted that these can easily be hundreds of megabytes. This can raise problems of storage and transmission, and methods by which the data can be compressed are of particular interest. The subject of data compression might more logically have been presented in Section 11.2, but it is more convenient to discuss it here, following on from the discussions of image classification, segmentation and GIS. The classification process itself is a form of data compression, since fewer bits are required to specify a classified image than the original image from which it was derived.

In general, methods of data compression can be divided into *reversible* (or lossless, or loss-free) processes from which all the original information can be retrieved, and *irreversible* (lossy) processes which entail some loss of information. Clearly a multispectral classification is an irreversible compression. Reversible data compression methods can be divided into two types: those that operate on a pixel-by-pixel basis, and those that make use of the spatial uniformity present in an image. Pixel-by-pixel compression exploits the fact that some pixel values occur more frequently than others. By encoding the data such that frequently occurring values are represented by short sequences of 0s and 1s (not strictly binary numbers since leading zeroes are also significant) and infrequently occurring values are given longer codes, it is possible to reduce the overall data volume. The theoretical limit to this approach is given by the Shannon information. If each pixel value can take one of M possible values, such that the value i occurs n_i times, the information content I can be defined as

$$I = N \log_2 M - \sum_{i=1}^M n_i \log_2 n_i \quad (11.31)$$

bits, where

$$N = \sum_{i=1}^M n_i$$

is the total number of pixels. To make this more definite, suppose a single-band image consists of 1024×1024 pixels, specified as 8-bit integers. Thus, $N = 1.05 \times 10^6$ and $M = 256$, and the uncompressed image occupies exactly one megabyte. We suppose also that the image histogram is uniform from pixel value a to pixel value b (inclusive of both values) and zero outside this range. Thus

$$n_i = \frac{N}{b - a + 1}$$

for $a \leq i \leq b$, and 0 otherwise. Substituting this expression into Equation (11.22) gives

$$I = N \log_2(b - a + 1),$$

so if we suppose that the range of non-zero pixel values is, for example, $b - a + 1 = 32$, we find that the Shannon information content of the image is just 640 kilobytes, or 5 bits per pixel. *Huffman coding* and the *Lempel-Ziv* (LZ) and *Lempel-Ziv-Welch* (LZW) algorithms are widely implemented forms of reversible pixel-by-pixel compression that can approach the Shannon information limit quite closely.

The other main type of reversible data compression employs the spatial information in the image. It is evident that if every pixel in a 1024×1024 pixel image has a pixel value that, while specified as an 8-bit number, actually has the same value (say 42), the whole image can be precisely specified in very few bits – as indeed we have just done in the preceding statement. *Run-length encoding* consists of specifying a pixel value, and then the number of places within the sequence of data (the length of the run) for which this value is repeated. Its scope for compressing the data volume obviously depends on the prevalence of homogeneous areas within the image, and it therefore tends to be more suitable for compressing classified images than original data unless the latter have been smoothed to reduce the level of noise. A more sophisticated version of run-length encoding uses *tesseral addressing* to specify the area and location of square regions within the image. However, whatever spatially based scheme is used, it is apparent that little or no saving in data volume can be made if the image is noisy, since in this case pixel values will change frequently, resulting in short ‘runs’ of data.

Irreversible (lossy) data compression can take a large number of forms, many of which are obvious. Images can be cropped to remove uninteresting areas, or sub-sampled to reduce their spatial resolution and volume at the same time. An alternative to sub-sampling is to form the Fourier transform of the image and then crop it so that it contains only the lower spatial frequencies. As an example, if the spatial resolution is halved using this technique, the dimensions of the Fourier transform array will also be halved so that the number of pixels in the Fourier transform is one quarter of the number in the original image. On the other hand, the Fourier transform actually needs two numbers for each pixel, to hold the real and imaginary parts respectively, so the data volume has in fact only been halved. More efficient schemes make use of transforms that are similar in function to the Fourier transform, but recognise that the input (spatial domain) data will always be real. Examples of such transforms include the *Hadamard transform*, the *Hartley transform* and the *discrete cosine transform* (DCT). Wavelet transforms are also used for image compression. Irreversible compression of images is clearly not desirable if they are to be used as input to quantitative image processing.

11.6.2

Image formats for remote sensing

In this section we describe the characteristics of some of the most important formats used for remotely sensed images.

11.6.2.1

Flat binary format

The simplest image format is a flat binary file. This contains nothing but the image data, so the necessary *metadata* (that is, data describing various attributes of the image such as its dimensions, georeferencing data and so on) are usually supplied as a separate text file.

11.6 Image formats and data compression

One important consideration for a flat binary file is the order in which the pixel values are stored. A pixel value has up to three coordinates: its row number, column number and (in the case of multi-band imagery) band number. Single-band images are usually stored row by row, i.e. all the pixel values of the first row are stored first, followed by all the pixels of the second row and so on. This is called *row-major order*, and we can represent it symbolically as

row > column

meaning that the row number increases more slowly than the column number as we proceed from one storage location to the next. With an analogous notation, we can represent three conventions for storing multi-band data as follows:

row > band > column,

which is called *bands interleaved by line* (BIL) format,

row > column > band,

which is called *bands interleaved by pixel* (BIP) format, and

band > row > column,

which is called *band sequential* (BSQ) format.

Another important consideration is the number of bytes used to represent each pixel value. A single byte can represent integers from 0 to 255, and this is still very common for remotely sensed images but other possibilities exist, including two-byte integers which can represent integers from 0 to 65 535 (unsigned integers) or from –32 768 to 32 767 (signed integers). Floating-point numbers (i.e. those that can be expressed as some decimal fraction multiplied by 10 raised to the power of some positive or negative integer exponent) are generally stored in four or eight bytes. If the data are represented by more than one byte, it is also necessary to specify the order in which the bytes are stored. The two possibilities are ‘*big-endian*’, in which the more significant bytes precede the less significant bytes, and ‘*little-endian*’.

The flat binary format cannot itself be compressed, but an image in this format can be stored in a compressed archive in a format such as zip, rar or 7z. These algorithms generally use some form of LZ or LZW compression.

11.6.2.2 Text format

Text format, or ASCII format, is a simple but rather inflexible way of storing image data. The pixel values are stored in a text file, usually in row-major order, with successive values separated by a space, comma or other delimiting symbol and the end of each row denoted by a ‘new line’ character. The advantages of this format are that it is not necessary to supply metadata indicating the number of rows and columns in the image since these can be inferred from the file itself, and that it is capable of representing larger integers than the range possible using a single byte, or indeed floating-point data. However, the format does not easily lend itself to the storage of multi-band data. The same remarks about compression apply to this format as to the flat binary format, and in fact the text format is usually rather wasteful of space: A typical text-format image might occupy two to three times as much space as the equivalent flat binary image.

11.6.2.3 GeoTIFF format

The TIFF (tagged image file format) format is in wide general use for digital images. It is a very flexible format, and many variants of it exist. TIFF files include a header that

contains metadata about the image in the form of ‘tags’. The GeoTIFF format exploits this to include georeferencing information in the header. The format allows for multiple (more than three) bands of data and floating-point values, and can use various forms of reversible compression, although not all image processing programs are able to recognise images incorporating these features. One major advantage of the GeoTIFF format is that, at least in the case of 8-bit, 3-band, uncompressed images, they can be read by a very wide range of image viewing software (although most of these programs make no use of the georeferencing data). This means that images used in scientific analysis can also be shared with non-specialist users. The header is small compared with the image so that an uncompressed GeoTIFF image occupies only very slightly more space than the equivalent flat binary image.

11.6.2.4 Other raster formats

There are other more specialised formats used for remotely sensed data. The *hierarchical data format* (HDF) is a flexible file format designed to store large amounts of numerical data, not just images. It is a ‘self-describing’ format, in the sense that it includes its own metadata, and tools for creating and reading HDF data are freely available. On the other hand, at the time of writing, the number of image processing applications that can read the format is rather limited. *Erdas Imagine* is commercial image processing software and, as always in this book, the inclusion of the name of a commercial product carries neither positive nor negative implications about its usefulness. However, the native image format, denoted by the file extension .img, is recognised by (and can be written by) a number of other image processing programs and is sometimes used for storage and transmission of images. The *CEOS* (Committee on Earth Observation Satellites) format is often used for radar images. None of these uses image compression. The *MrSID* (the acronym stands for multiresolution seamless image database, and is pronounced ‘Mister Sid’) format uses wavelet compression either reversibly or, with larger compression ratios, irreversibly.

11.6.2.5 Vector formats

One of the commonest formats for vector data is the *shapefile*. This format was developed by the commercial company ESRI, but it is largely an open format so that shapefiles can be exchanged between a range of programs. In fact, a shapefile consists of a number of files: at minimum, the shapefile itself (with the file extension .shp), which represents the geometry of points, lines and polygons, a database (extension .dbf) which contains attribute data, and the shape index file (extension .shx). Other optional files may be included, of which the commonest is the projection file (extension .prj) which contains information on the map projection.

11.6.2.6 Non-specialist image formats

The image formats that we have discussed above are more or less specialised, but formats used for storing computer images in general can also sometimes be used to store remotely sensed data. Apart from the TIFF format, which has already been mentioned, other common formats include GIF (graphics interchange format), PNG (portable network graphics) and JFIF (joint photographic expert group file interchange format). The GIF format can store only a single byte per pixel but uses reversible (LZW) compression. It has been subject to some licensing issues, and partly as a result of these the rather more

Review questions

versatile *portable network graphics* (PNG) format was developed in the early 2000s. This is also compressed, but unlike GIF it can store three bands of data. An image stored in PNG format might typically occupy around a quarter as much space as in flat binary format.

The very widespread JFIF format, usually rather imprecisely called the *JPEG format*, uses DCT to achieve irreversible compression. In the usual implementation of this format, the image data are processed in blocks of 8×8 pixels. The DCT is calculated for each block, and sufficient terms retained so that the regenerated image will appear reasonably accurate to the human eye. Very large compression factors can be achieved using this method, and it is often used for the storage and transmission of ‘quick-look’ images. It is obviously not suitable for images that will be subjected to quantitative image processing. However, the more recently developed JPEG2000 format, which is based on wavelet compression, can achieve significant reversible compressions as well as irreversible compression.

Summary

Images can be compressed reversibly (losslessly) or irreversibly (lossily). Irreversible compression, such as is provided by the JFIF format, gives greater compression factors but damages the quantitative integrity of the image so that it is less useful, or not useful at all, for quantitative analysis. Reversible compression is however useful for the storage and transmission of remotely sensed images, which can be large. Compression techniques are generally based on run-length encoding or on pixel-by-pixel recoding of the pixel values of the image, such as in Huffman coding. There is a theoretical limit to the extent to which pixel-by-pixel recoding can compress an image, set by the properties of the image histogram.

A wide range of image formats is in use for the storage of remotely sensed data. Some of these include metadata, describing the characteristics of the image, its georeferencing data and so on, in a header within the image file itself, while others make use of an external metadata file. The simplest file format is the ‘flat binary’ image file, which does not include a header. In this case the metadata must include information on the order in which the bytes are stored in the image. The GeoTIFF format, which has a header that includes georeferencing information, has become widely used and is capable of being compressed.

Review questions

Explain what is meant by the *station mask* of a satellite receiving station.

What steps are needed to convert the pixel value recorded by a satellite sensor to the corresponding at-satellite spectral radiance? What further steps are needed to calculate the planetary and surface reflectances?

What is meant by georeferencing of a satellite image?

Why is spatial interpolation generally necessary in reprojecting an image?

Explain the concept of histogram equalisation. Why might it be useful?

Explain what is meant by a spatial convolution filter in digital image processing.

What is the difference between edge-detection and edge-enhancement? Discuss applications of both.

Discuss the use of non-linear spatial filters in digital image processing.

Describe the nature and use of band transformations in multispectral imagery.

Distinguish between supervised and unsupervised image classification.

Explain what is meant by mixture modelling and spectral unmixing.

How can the accuracy of an image classification be assessed?

What is meant by ‘image texture’? How can it be quantified?

How can particular shapes be identified in an image?

Give a brief description of a geographic information system.

Discuss the methods available for image compression. What determines the amount of compression that is achievable?

Problems

1. How many uniformly spaced geostationary satellites would be needed to ensure that all points on the Earth’s surface with latitudes less than 66.5° can be seen at an elevation angle of at least 10° ?

2. (i) Estimate appropriate values, corresponding to those in Table 11.1, for printed paper as a data storage medium. (ii) Estimate the potential data storage requirements from a 5-year spaceborne remote sensing mission.

3. Identify the nature of the spatial filters whose kernels are given below:

(i) (ii) (iii)

1 3 1 1 3 1 −1 −3 −1

3 −8 3 3 6 3 −3 16 −3

1 3 1 1 3 1 −1 −3 −1

4. A single band of an image has a histogram with a Gaussian distribution of standard deviation σ . Show that the theoretical limit to which the data can be compressed using Huffman coding is

$$\log_2 \sigma + 2.05$$

bits per pixel. You may assume that σ is large enough for the digitisation noise to be ignored. You may find the following integrals useful:

$$\int_{-\infty}^{\infty} \exp(-x^2/2\sigma^2) dx = \sigma\sqrt{2\pi},$$

$$\int_{-\infty}^{\infty} x^2 \exp(-x^2/2\sigma^2) dx = \sigma^3 \sqrt{2\pi}.$$

5. (i) Show that if \mathbf{A} , \mathbf{S} and \mathbf{E} are, respectively, isotropic smoothing (averaging), sharpening (high-boost) and edge-detection (high-pass) filters described in the spatial domain, and \mathbf{I} is the identity operator, the following are possible relationships:

$$(a) \mathbf{S} = a\mathbf{I} + (1 - a)\mathbf{A},$$

$$(b) \mathbf{E} = b(\mathbf{I} - \mathbf{S}).$$

Discuss the range of values that a and b can take, and any further modification of (b) that may be needed in a practical implementation.

- (ii) The kernel of a 3×3 Laplace filter is shown below:

$$\begin{matrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{matrix}$$

Explain why this is an edge-detection filter, and derive two sharpening and two smoothing filters from it using the equations in (i). Explain briefly the differences between the two filters in each case.

6. The relationship between spectral radiance L_λ and brightness temperature T in a thermal infrared band of a satellite sensor can be expressed as

$$T = \frac{K_2}{\ln\left(\frac{K_1}{L_\lambda} + 1\right)}.$$

For band 6 of the Landsat-7 ETM+ instrument, the constants have values $K_1 = 666.1 \text{ W m}^{-2} \text{ sr}^{-1} \mu\text{m}^{-1}$ and $K_2 = 1282.7 \text{ K}$. Use the satellite metadata given in Section 11.2.1.1 to find the at-satellite brightness temperature corresponding to a digital number of 132 recorded in band 6.1.

Appendix Data tables

This appendix collects together some of the more commonly needed general data in remote sensing. Some of them also appear in the main text; however, it is felt that it will be a convenience also to present them here.

A.1

Physical constants

c	speed of light <i>in vacuo</i>	$2.9979 \times 10^8 \text{ m s}^{-1}$ (defined as $299\,792\,458 \text{ m s}^{-1}$)
h	Planck constant	$6.6261 \times 10^{-34} \text{ J s}$
e	charge on the proton	$1.6022 \times 10^{-19} \text{ C}$
m_e	mass of the electron	$9.1094 \times 10^{-31} \text{ kg}$
u	atomic mass unit	$1.6605 \times 10^{-27} \text{ kg}$
m_p	mass of the proton	1.0073 u
m_n	mass of the neutron	1.0087 u
μ_0	permeability of free space	$1.2566 \times 10^{-6} \text{ H m}^{-1}$ (defined as $4\pi \times 10^{-7} \text{ H m}^{-1}$)
ϵ_0	permittivity of free space	$8.8542 \times 10^{-12} \text{ F m}^{-1}$
Z_0	impedance of free space	$3.7673 \times 10^2 \Omega$
G	gravitational constant	$6.6726 \times 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$
k	Boltzmann constant	$1.3807 \times 10^{-23} \text{ J K}^{-1}$
σ	Stefan–Boltzmann constant	$5.6705 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$

A.2

Units

The table lists some units in common use in remote sensing, together with their SI equivalents.

1 degree	$\pi/180 \approx 1.7453 \times 10^{-2}$ radian
1 minute of arc (1/60 degree)	2.9089×10^{-4} radian
1 second of arc (1/60 minute)	4.8481×10^{-6} radian

(cont.)

1 angstrom (\AA)	10^{-10} m (exactly)
1 inch	25.4 mm (exactly)
1 foot	304.8 mm (exactly)
1 statute mile	1.6093 km
1 nautical mile	1.852 km (exactly)
1 astronomical unit (AU)	$1.4960 \times 10^{11} \text{ m}$
1 hectare	10^4 m^2 (exactly)
1 litre	10^{-3} m^3 (exactly)
1 acre	$4.0469 \times 10^3 \text{ m}^2$
1 pound	$4.5359 \times 10^{-1} \text{ kg}$
1 knot	$5.144 \times 10^{-1} \text{ m s}^{-1}$
1 bar	10^5 Pa (exactly)
1 atmosphere	$1.0133 \times 10^5 \text{ Pa}$
1 torr = 1 mm Hg	$1.333 \times 10^2 \text{ Pa}$
1 electronvolt (eV)	$1.6022 \times 10^{-19} \text{ J}$
1 calorie	4.1840 J
1 gauss	10^{-4} T
1 gamma	10^{-9} T

A.3**Illuminance at the Earth's surface**

Illuminance is the photometric quantity that is analogous to the radiometric quantity called irradiance. Its definition is discussed in Section 5.3. The following table gives typical values of the illuminance at the Earth's surface.

Condition	Illuminance (lux)
Noon, clear day	10^5
Noon, overcast	10^4
Sunrise and sunset	500
Sun 5° below horizon	5
Full Moon, clear sky	0.5
Full Moon, overcast	0.05
Clear starlit sky	0.005

A.4

Properties of the Sun and Earth

Sun's radius	6.955×10^8 m
Sun's mass	1.989×10^{30} kg
Total radiated solar power	3.839×10^{26} W
Sun's effective black-body temperature	5780 K
Earth's equatorial radius	6 378 137 m
Earth's polar radius	6 356 752 m
Earth's mean (authalic) radius	6 371 007 m
Semimajor axis of Earth's orbit around the Sun	1 AU (exactly)
Earth's mass	5.974×10^{24} kg
Earth's angular velocity about its polar axis	7.292×10^{-5} s ⁻¹
Sidereal day	86 164.09 s
Tropical year (equinox to equinox)	31 556 926 s
Sidereal year (fixed star to fixed star)	31 558 150 s
Geocentric gravitational constant (GM)	3.986×10^{14} m ³ s ⁻¹
Standard gravitational acceleration at Earth's surface (g)	9.807 m s ⁻²
Earth's dynamical form factor (J_2)	1.0826×10^{-3}
Mean global albedo	0.35
Land area	1.489×10^{14} m ²
Ocean area	3.614×10^{14} m ²
Mean land elevation	860 m
Mean ocean depth	3.9 km
Mean annual atmospheric temperature at sea level:	
at the equator	27 °C
at 30° N or S latitude	20 °C
at 60° N or S latitude	-2 °C
at the poles	-25 °C

A.5

Position of the Sun

Since much environmental remote sensing is dependent on solar illumination, it is important to be able to calculate the Sun's position in the sky at a given date, time and position on the Earth's surface. This will permit, for example, the hours of darkness and solar elevation angles to be calculated.

The relevant variables describing the Sun's position relative to the Earth are the *solar declination* δ and the *equation of time* E . It will sometimes also be necessary to know the

A.5 Position of the Sun

distance D between the centres of the Earth and the Sun. The declination is the Sun's angle above or below the Earth's equatorial plane. The equation of time is the difference in position between the true Sun and a fictitious Sun (the 'mean Sun') that appears to travel uniformly across the sky. It is normally expressed in minutes of time. For a position on the Earth's surface having longitude λ (with longitudes east of the Greenwich meridian being positive), the Sun's *hour angle* H is given, in degrees, by

$$H = 15T - 180 + \lambda + \frac{E}{4}, \quad (\text{A.1})$$

where T is the Universal Time (Greenwich Mean Time) in hours, λ is measured in degrees and E in minutes of time. It may be necessary to add or subtract 360° from Equation (A.1) to generate a value that is in the range -180° to $+180^\circ$.

For a position on the Earth's surface having latitude φ (defined such that north latitudes are positive) and from which the solar hour angle is H , the Sun's elevation angle a is given by

$$\sin a = \sin \delta \sin \phi + \cos \delta \cos \phi \cos H \quad (\text{A.2})$$

and its azimuth angle A (north = 0° , east = 90° etc.) is given by

$$\cos A = \frac{\sin \delta - \sin \phi \sin a}{\cos \phi \cos a}. \quad (\text{A.3})$$

The ambiguity in solving Equation (A.3) for A is resolved by noting that if the hour angle H is negative, the azimuth A must be less than 180° , while if H is positive, A must be greater than 180° .

The values of δ , E and D are tabulated in a number of official publications, they can be found on the Internet, or they can be calculated from standard astronomical formulae (Figure A.1). The following formulae are simple approximations that use only the day number d within the year (i.e. January 1 = 1, February 1 = 32 etc.). In all cases, the formulae assume that trigonometrical quantities are expressed in degrees. Equation (A.4) gives the value of E in minutes, to an accuracy of about 1 minute. Equation (A.5) is accurate to about 1 degree. Equation (A.6) gives the value of D in astronomical units, accurate to about 0.001 AU.

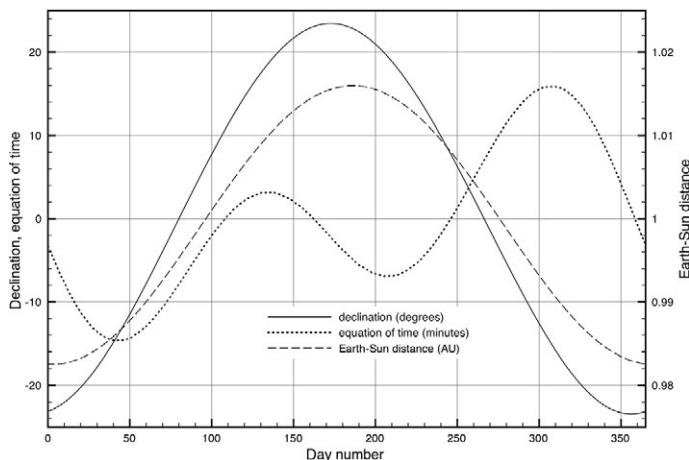


Figure A.1. The Sun's declination, the equation of time, and the Earth–Sun distance, calculated for the year 2011. The graphs for any other year are very similar.

$$E \approx 9.9 \sin 1.97(d - 80) - 7.7 \sin 0.986(d - 3) \quad (\text{A.4})$$

$$\sin \delta \approx 0.3987 \sin 0.986(d - 80) \quad (\text{A.5})$$

$$D \approx 1 - 0.0167 \cos 0.986(d - 3) \quad (\text{A.6})$$

Example

Suppose we wish to estimate the Sun's position from 2°N, 104°E at 00:00 UT on 1 November 2011. An exact calculation gives $\delta = -14.25^\circ$ and $E = +15.84$ minutes for this date and time. Equation (A.1) gives the Sun's hour angle $H = -72.04^\circ$, and substitution of this value into Equation (A.2) gives the solar elevation as 16.86° . From Equation (A.3), $\cos A = -0.2679$. Taking the inverse cosine gives $A = 105.54^\circ$ or 254.46° , and since H is negative, the correct value is $A = 105.54^\circ$. The value of the Earth–Sun distance according to an accurate calculation is 0.992 AU.

Using the approximations of Equations (A.4) and (A.5), we obtain $\delta = -15.43^\circ$ and $E = +16.64$ minutes, and hence $H = -71.84^\circ$, $a = 16.92^\circ$ and $A = 106.79^\circ$. The solar distance is given by Equation (A.6) as 0.992 AU.

REFERENCES

- Abdalati, W., and K. Steffen. 1995. "Passive microwave-derived snow melt regions on the Greenland Ice Sheet." *Geophysical Research Letters* **22**: 787–790.
- Anon. MODIS UCSB Emissivity Library. <http://www.icesis.ucsb.edu/modis/EMIS/html/em.html>.
- Baltsavias, E. P. 1999. "Airborne laser scanning: basic relations and formulas." *ISPRS Journal of Photogrammetry and Remote Sensing* **54** (2–3): 199–214.
- Baltsavias, E. P., and A. Gruen. 2003. Resolution convergence: a comparison of aerial photos, LiDAR and IKONOS for monitoring cities. In V. Mesev (ed.), *Remotely Sensed Cities*, 47–82. London: Taylor & Francis.
- Bamber, J. L., R. L. Layberry, and S. P. Gogineni. 2001. "A new ice thickness and bed data set for the Greenland ice sheet. 1: Measurement, data reduction, and errors." *Journal of Geophysical Research* **106** (D24): 33773–33780.
- Barale, V., J. F. R. Gower, and L. Alberotanza. 2010. *Oceanography from Space: Revisited*. Springer.
- Bartsch, A., R. A. Kidd, W. Wagner, and Z. Bartalis. 2007. "Temporal and spatial variability of the beginning and end of daily spring freeze/thaw cycles derived from scatterometer data." *Remote Sensing of Environment* **106**: 360–374.
- Bass, F. G., and I. M. Fuks. 1979. *Wave Scattering from Statistically Rough Surfaces*. Oxford: Pergamon Press.
- Batut, A. 1890. *La Photographie Aérienne par Cerf-volant*. Paris: Gauthier-Villars et fils.
- Beckman, P., and A. Spizzichino. 1987. *The Scattering of Electromagnetic Waves from Rough Surfaces*. Norwood, Massachusetts: Artech House.
- Berk, A., G. P. Anderson, P. K. Acharya, et al. 2005. "MODTRAN 5: a reformulated atmospheric band model with auxiliary species and practical multiple scattering options: update." *Proceedings of SPIE* **5806** (1): 662–667. doi:10.1117/12.606026.
- Berlin, G. L. L., and T. E. Avery. 2003. *Fundamentals of Remote Sensing and Airphoto Interpretation*. 6th edn. Upper Saddle River, New Jersey: Prentice Hall.
- Bingham, R. G., and M. J. Siegert. 2007. "Radio-echo sounding over polar ice masses." *Journal of Environmental and Engineering Geophysics* **12** (1): 47–62. doi:10.2113/JEEG12.1.47.
- Bohren, C. F., and D. R. Huffman. 1983. *Absorption and Scattering of Light by Small Particles*. New York: John Wiley.
- Bracewell, R. N. 1999. *The Fourier Transform and Its Applications*. 3rd edn. New York: McGraw-Hill.
- Brown, G. 1977. "The average impulse response of a rough surface and its applications." *IEEE Transactions on Antennas and Propagation* **25** (1): 67–74. doi:10.1109/TAP.1977.1141536.
- Burger, W., and M. J. Burger. 2005. *Digital Image Processing: An Algorithmic Introduction Using Java*. Berlin: Springer-Verlag.

References

- Burns, R. W. 2000. *John Logie Baird: TV Pioneer*. London: The Institution of Engineering and Technology.
- Campbell, J. B. 2008. *Introduction to Remote Sensing*. 4th edn. New York: Guilford Press.
- Carsey, F. D. 1992. *Microwave Remote Sensing of Sea Ice*. Washington, DC: American Geophysical Union.
- Chahine, M. T., L. Chen, P. Dimotakis, et al. 2008. "Satellite remote sounding of mid-tropospheric CO₂." *Geophysical Research Letters* **35** (17): L17807.
- Chander, G., X. Xiong, T. Choi, and A. Angal. 2010. "Monitoring on-orbit calibration stability of the Terra MODIS and Landsat 7 ETM+ sensors using pseudo-invariant test sites." *Remote Sensing of Environment* **114** (4): 925–939. doi:10.1016/j.rse.2009.12.003.
- Charalambos, P., and Y. Suya. 2010. "Delineation and geometric modeling of road networks." *ISPRS Journal of Photogrammetry and Remote Sensing* **65** (2): 165–181. doi:10.1016/j.isprsjprs.2009.10.004.
- Chase, A., D. Z. Chase, J. F. Weishampel, et al. 2011. "Airborne LiDAR, archaeology, and the ancient Maya landscape at caracol, Belize." *Journal of Archaeological Science* **38** (2): 387–398.
- Chavez, P. S. 1988. "An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data." *Remote Sensing of Environment* **24** (3): 459–479. doi:10.1016/0034-4257(88)90019-3.
- Cohen, J. 1960. "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement* **20**: 37–40.
- Cohen, W. B., and S. N. Goward. 2004. "Landsat's role in ecological applications of remote sensing." *BioScience* **54** (6): 535–545. doi:10.1641/0006-3568(2004)054[0535:LRIEAO]2.0.CO;2.
- Comiso, J. C. 2009. "Enhanced sea ice concentrations and ice extents from AMSR-E data." *Journal of the Remote Sensing Society of Japan* **29** (1): 199–215.
- Cortes, C., and V. Vapnik. 1995. "Support-vector networks." *Machine Learning* **20** (3): 273–297.
- Crosetto, M., O. Monserrat, M. Cuevas, and B. Crippa. 2011. "Spaceborne differential SAR interferometry: data analysis tools for deformation measurement." *Remote Sensing* **3** (2): 305–318. doi:10.3390/rs3020305.
- Dong, S., and L. Huang. 2011. Mapping surface displacement based on D-InSAR technique. In *2011 International Conference on Remote Sensing, Environment and Transportation Engineering (RSETE)*, 3259–3262. IEEE. doi:10.1109/RSETE.2011.5965008.
- Donlon, C. J., P. J. Minnett, C. Gentemann, et al. 2002. "Toward improved validation of satellite sea surface skin temperature measurements for climate research." *Journal of Climate* **15** (4): 353–369. doi:10.1175/1520-0442(2002)015<0353:TIVOSS>2.0.CO;2.
- Dozier, J. 1989. "Spectral signature of Alpine snow cover from the Landsat Thematic Mapper." *Remote Sensing of Environment* **28**: 9–22.
- Drinkwater, M. R., D. G. Long, and A. W. Bingham. 2001. "Greenland snow accumulation estimates from satellite radar scatterometer data." *Journal of Geophysical Research* **106**: 33935–33950.
- Duda, R. O., and P. E. Hart. 1973. *Pattern Recognition and Scene Analysis*. New York: Wiley.
- Elachi, C., and J. van Zyl. 2006. *Introduction to the Physics and Techniques of Remote Sensing*. 2nd edn. Hoboken, New Jersey: Wiley-Interscience.
- Eppler, D. T., L. D. Farmer, A. W. Lohanick, et al. 1992. Passive microwave signatures of sea ice. In F. D. Carsey (ed.), *Microwave Remote Sensing of Sea Ice*, 47–71. Washington, DC: American Geophysical Union.

References

- Espindola, G. M., G. Camara, I. A. Reis, L. S. Bins, and A. M. Monteiro. 2006. "Parameter selection for region-growing image segmentation algorithms using spatial autocorrelation." *International Journal of Remote Sensing* **27** (14): 3035–3040. doi:10.1080/01431160600617194.
- Fan, K., W. Huang, H. Lin, et al. 2011. "Shallow water depth retrieval from space-borne SAR imagery." *Journal of Oceanography* **67**(4): 405–413.
- Farr, T. G., P. A. Rosen, E. Caro, et al. 2007. "The Shuttle Radar Topography Mission." *Reviews of Geophysics* **45**: RG2004. doi:10.1029/2005RG000183.
- Feynman, R. P., R. B. Leighton, and M. Sands. 2005. *The Feynman Lectures on Physics*. 2nd edn. 3 vols. Reading, Massachusetts: Addison-Wesley.
- Flood, M. 2001. "Laser altimetry: from science to commercial LiDAR mapping." *Photogrammetric Engineering and Remote Sensing* **67** (11): 1209–1217.
- Flynn, L. E., D. McNamara, C. T. Beck, et al. 2009. "Measurements and products from the Solar Backscatter Ultraviolet (SBUV/2) and Ozone Mapping and Profiler Suite (OMPS) instruments." *International Journal of Remote Sensing* **30** (15–16): 4259–4272. doi:10.1080/01431160902825040.
- Foody, G. M. 2002. "Status of land cover classification accuracy assessment." *Remote Sensing of Environment* **80**: 185–201.
- Forshaw, M. R. B., A. Haskell, P. F. Miller, D. J. Stanley, and J. R. G. Townshend. 1983. "Spatial resolution of remotely sensed images." *International Journal of Remote Sensing* **4** (3): 497–520.
- Frison, P. L., and E. Mougin. 1996. "Monitoring global vegetation dynamics with ERS-1 wind scatterometer data." *International Journal of Remote Sensing* **17**: 3201–3218.
- Fu, L.-L. 2010. Determining ocean circulation and sea level from satellite altimetry: progress and challenges. In V. Barale, J. F. R. Gower, and L. Alberotanza (eds.). *Oceanography from Space: Revisited*, 147–164. Springer.
- Gao, B. 1996. "NDWI – a normalized difference water index for remote sensing of vegetation liquid water from space." *Remote Sensing of Environment* **58** (3): 257–266.
- Gao, J. 2009. "Bathymetric mapping by means of remote sensing: methods, accuracy and limitations." *Progress in Physical Geography* **33** (1): 103–116. doi:10.1177/0309133309105657.
- Gens, R., and J. L. van Genderen. 1996. "Review article: SAR interferometry – issues, techniques, applications." *International Journal of Remote Sensing* **17** (10): 1803–1835. doi:10.1080/01431169608948741.
- Goetz, A. F. H. 1979. *Preliminary Stereosat mission description*. JPL Laboartory Report. Pasadena, California: Jet Propulsion Laboratory.
- Green, A. A., M. Berman, P. Switzer, and M. D. Craig. 1998. "A transformation for ordering multispectral data in terms of image quality with implications for noise removal." *IEEE Transactions on Image Processing* **26** (1): 65–74.
- Gueymard, C. 2004. "The sun's total and spectral irradiance for solar energy applications and solar radiation models." *Solar Energy* **76** (4): 423–453.
- Hahn, C. J., W. B. Rossow, and S. G. Warren. 2001. "ISCCP cloud properties associated with standard cloud types identified in individual surface observations." *Journal of Climate* **14**: 11–28.
- Hapke, B. 2005. *Theory of Reflectance and Emittance Spectroscopy*. Cambridge: Cambridge University Press.
- Harris, R. 2009. Remote sensing policy. In T. A. Warner, M. D. Nellis, and G. M. Foody (eds.), *The SAGE Handbook of Remote Sensing*, 18–30. London: SAGE Publications Ltd.

References

- Hausman, J., and V. Zlotnicki. 2010. "Sea state bias in radar altimetry revisited." *Marine Geodesy* **33**: 336–347. doi:10.1080/01490419.2010.487804.
- Hecht, E. 2003. *Optics*. 4th edn. Harlow, UK: Pearson Education.
- Henyey, L. C., and J. L. Greenstein. 1941. "Diffuse radiation in the galaxy." *Astrophysical Journal* **93**: 70–83.
- Holt, B., P. Kanagaratnam, S. P. Gogineni, et al. 2009. "Sea ice thickness measurements by ultrawideband penetrating radar: first results." *Cold Regions Science and Technology* **55** (1): 33–46. doi:10.1016/j.coldregions.2008.04.007.
- Hunt, G. R. 1980. Electromagnetic radiation: the communication link in remote sensing. In B. S. Siegel and A. R. Gillespie (eds.) *Remote Sensing in Geology*, 5–45. New York: John Wiley & Sons.
- Höfle, B., and M. Rutzinger. 2011. "Topographic airborne LiDAR in geomorphology: a technological perspective." *Zeitschrift für Geomorphologie, Supplementary Issues* **55** (2): 1–29. doi:10.1127/0372-8854/2011/0055S2–0043.
- Høgda, K. A., R. Storvold, and T. R. Lauknes. 2010. SAR imaging of glaciers. In P. Pellikka and W. G. Rees (eds.), *Remote Sensing of Glaciers*, 153–178. London: Taylor & Francis/CRC.
- Irons, J. 2011. *Landsat 7 Science Data Users Handbook*. <http://landsathandbook.gsfc.nasa.gov/>.
- Jensen, J. R. 2006. *Remote Sensing of the Environment: An Earth Resource Perspective*. 2nd edn. Upper Saddle River, New Jersey: Prentice Hall.
- Jiang, Liming, Hui Lin, Jianwei Ma, Bing Kong, and Yao Wang. 2011. "Potential of small-baseline SAR interferometry for monitoring land subsidence related to underground coal fires: Wuda (Northern China) case study." *Remote Sensing of Environment* **115** (2): 257–268. doi:10.1016/j.rse.2010.08.008.
- Jimenez-Munoz, J. C., J. Cristobal, J. A. Sobrino, et al. 2009. "Revision of the single-channel algorithm for land surface temperature retrieval from Landsat thermal-infrared data." *IEEE Transactions on Geoscience and Remote Sensing* **47** (1): 339–349. doi:10.1109/TGRS.2008.2007125.
- Johnson, W. 1995. "Contents and commentary on William Moore's 'a treatise on the motion of rockets and an essay on naval gunnery'." *International Journal of Impact Engineering* **16** (3): 499–521.
- Jol, H. M. 2009. *Ground Penetrating Radar: Theory and Applications*. Amsterdam: Elsevier.
- Jones, H. G., and R. A. Vaughan. 2010. *Remote Sensing of Vegetation*. Oxford: Oxford University Press.
- Jones, M. O., L. A. Jones, J. S. Kimball, and K. C. McDonald. 2011. "Satellite passive microwave remote sensing for monitoring global land surface phenology." *Remote Sensing of Environment* **115** (4): 1102–1114. doi:10.1016/j.rse.2010.12.015.
- Joss, J., and E. G. Gori. 1978. "Shapes of raindrop size distributions." *Journal of Applied Meteorology* **17**: 1054–1061.
- Justice, C. O., J. R. G. Townshend, B. N. Holben, and C. J. Tucker. 1985. "Analysis of the phenology of global vegetation using meteorological satellite data." *International Journal of Remote Sensing* **6**: 1271–1318.
- Kahle, A. B., A. R. Gillespie, and A. F. H. Goetz. 1976. "Thermal inertia imaging: a new geologic mapping tool." *Geophysical Research Letters* **3**: 22–28.
- Kasischke, E. S., M. A. Tanase, L. L. Bourgeau-Chavez, and M. Borr. 2011. "Soil moisture limitations on monitoring boreal forest regrowth using spaceborne L-band SAR data." *Remote Sensing of Environment* **115** (1): 227–232. doi:10.1016/j.rse.2010.08.022.

References

- Kauth, R. J., and G. S. Thomas. 1976. The tasseled cap – a graphic description of the spectral-temporal development of agricultural crops as seen by LANDSAT. In *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data*, 4B41–4B51. West Lafayette, Indiana: Purdue University.
- Kettig, R. L., and D. A. Landgrebe. 1976. “Classification of multispectral image data by extraction and classification of homogeneous objects.” *IEEE Transactions on Geoscience Electronics* GE14: 19–26.
- Key, C. H., and N. C. Benson. 2003. *The normalized burn ratio (NBR): a Landsat TM radiometric measure of burn severity*. US Geological Survey Northern Rocky Mountain Science Center.
- Kimball, J. S., K. C. McDonald, A. R. Keyser, S. Frolking, and S. W. Running. 2001. “Application of the NASA scatterometer (NSCAT) for determining daily frozen and nonfrozen landscape of Alaska.” *Remote Sensing of Environment* 75: 113–126.
- Kneizys, F. X., E. P. Shettle, L. W. Abreu, *et al.* 1988. *User’s Guide to LOWTRAN 7*. Hanscom Air Force Base, Massachusetts: Air Force Geophysics Laboratory.
- Kraus, K. 2007. *Photogrammetry: Geometry from Images and Laser Scans*. 2nd edn. Berlin: Walter de Gruyter GmbH & Co.
- Laws, J. O., and D. A. Parsons. 1943. “The relation of raindrop size to intensity.” *Transactions of the American Geophysical Union* 24: 452–460.
- Liang, S. 2004. *Quantitative Remote Sensing of Land Surfaces*. Hoboken, New Jersey: John Wiley & Sons.
- Lillesand, T. M., R. W. Kiefer, and J. Chipman. 2008. *Remote Sensing and Image Interpretation*. 6th edn. Hoboken, New Jersey: John Wiley & Sons.
- Lin, I., L. K. Kwoh, Y.-C. Lin, and V. Khoo. 1997. Ship and ship wake detection in the ERS SAR imagery using computer-based algorithm. In *Remote Sensing: A Scientific Vision for Sustainable Development*, 1:151–153. Florence: IEEE Publications.
- Liu, W. T. 2002. “Progress in scatterometer application.” *Journal of Oceanography* 58: 121–136.
- Long, M. W. 2001. *Radar Reflectivity of Land and Sea*. Boston: Artech House.
- Longair, M. S. 2003. *Theoretical Concepts in Physics*. 2nd edn. Cambridge: Cambridge University Press.
- Longley, P., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2005. *Geographic Information Systems and Science*. 2nd edn. Chichester, UK: John Wiley & Sons.
- Macdonald, R. A. 1995. “CORONA: Success for space reconnaissance, a look into the Cold War, and a revolution for intelligence.” *Photogrammetric Engineering and Remote Sensing* 61: 689–720.
- Maddy, E. S., and C. D. Barnet. 2008. “Vertical resolution estimates in Version 5 of AIRS operational retrievals.” *IEEE Transactions on Geoscience and Remote Sensing* 46 (8): 2375–2384. doi:10.1109/TGRS.2008.917498.
- Maini, A. K., and V. Agrawal. 2010. *Satellite Technology: Principles and Applications*. Chichester, UK: John Wiley & Sons.
- Mallet, C., and F. Bretar. 2009. “Full-waveform topographic lidar: state-of-the-art.” *ISPRS Journal of Photogrammetry and Remote Sensing* 64 (1): 1–16. doi:10.1016/j.isprsjprs.2008.09.007.
- Marshall, G. J., J. A. Dowdeswell, and W. G. Rees. 1994. “The spatial and temporal effect of cloud cover on the acquisition of high quality Landsat imagery in the European Arctic sector.” *Remote Sensing of Environment* 50: 149–160.

References

- Marshall, J. S., and W. M. K. Palmer. 1948. "The distribution of raindrops with size." *Journal of Meteorology* **5**: 165–166.
- Martinez, J.-M., and T. Le Toan. 2007. "Mapping of flood dynamics and spatial distribution of vegetation in the Amazon floodplain using multitemporal SAR data." *Remote Sensing of Environment* **108** (3): 209–223. doi:10.1016/j.rse.2006.11.012.
- Mather, P. M., and M. Koch. 2010. *Computer Processing of Remotely-Sensed Images*. 4th edn. Chichester, UK: John Wiley & Sons.
- Mätzler, C. 2002. *MATLAB Functions for Mie Scattering and Absorption*. Bern: Institut für Angewandte Physik, University of Bern.
- Mayaux, P., and K. Bossard. 2000. The Land Cover Map for Australia in the Year 2000. *GLC2000 Database*. <http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php>.
- Mayer, B., and A. Kylling. 2005. "Technical note: The libRadtran software package for radiative transfer calculations – description and examples of use." *Atmospheric Chemistry and Physics* **5** (7): 1855–1877. doi:10.5194/acp-5-1855-2005.
- McClain, C. R. 2009. "A decade of satellite ocean color observations." *Annual Review of Marine Science* **1** (1): 19–42. doi:10.1146/annurev.marine.010908.163650.
- Meier, W. N., and J. Stroeve. 2008. "Comparison of sea-ice extent and ice-edge location estimates from passive microwave and enhanced-resolution scatterometer data." *Annals of Glaciology* **48** (1): 65–70. doi:10.3189/172756408784700743.
- Merchant, C. J., A. R. Harris, E. Maturi, et al. 2009. "Sea surface temperature estimation from the Geostationary Operational Environmental Satellite-12 (GOES-12)." *Journal of Atmospheric and Oceanic Technology* **26** (3): 570–581. doi:10.1175/2008JTECHO596.1.
- Mitchell, T. D., and P. D. Jones. 2005. "An improved method of constructing a database of monthly climate observations and associated high-resolution grids." *International Journal of Climatology* **25** (6): 693–712.
- Naesset, E. 1997. "Determination of mean tree height of forest stands using airborne laser scanner data." *ISPRS Journal of Photogrammetry and Remote Sensing* **52**: 49–56.
- Nerem, R. S., D. P. Chambers, C. Choe, and G. T. Mitchum. 2010. "Estimating mean sea level change from the TOPEX and Jason altimeter missions." *Marine Geodesy* **33**: 435–446. doi:10.1080/01490419.2010.491031.
- Neubert, M., H. Herold, and G. Meinel. 2006. "Evaluation of remote sensing image segmentation quality – further results and concepts." *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **36** (4/C42).
- Njoku, E. G., M. Moghaddam, D. Moller, and N. Molotch. 2010. "Microwave remote sensing for land hydrology research and applications: Introduction to the Special Issue." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **3** (1): 3–5. doi:10.1109/JSTARS.2010.2042748.
- Njoku, E. G., T. L. Jackson, V. Lakshmi, T. Chan, and S. V. Nghiem. 2003. "Soil moisture retrieval from AMSR-E." *IEEE Transactions on Geoscience and Remote Sensing* **41** (2): 215–229.
- Oliver, C., and S. Quegan. 2004. *Understanding Synthetic Aperture Radar Images*. Raleigh, North Carolina: SciTech Publishing.
- Parcak, S. H. 2009. *Satellite Remote Sensing for Archaeology*. Abingdon, UK: Taylor & Francis.
- Pavlis, N. K., S. A. Holmes, S. C. Kenyon, and J. K. Factor. 2008. An Earth gravitational model to degree 2160: EGM2008. EGU General Assembly, Vienna.

- Platnick, S., M. D. King, S. A. Ackerman, *et al.* 2003. "The MODIS cloud products: algorithms and examples from Terra." *IEEE Transactions on Geoscience and Remote Sensing* **41** (2): 459–473.
- Purdy, R. 1999. "Legal and privacy implications of 'spy in the sky' satellites." *Mountbatten Journal of Legal Studies* **3** (1): 63–79.
- Purkis, S., and V. Klemas. 2011. *Remote Sensing and Global Environmental Change*. Chichester, UK: Wiley-Blackwell.
- Quattrochi, D. A., and J. C. Luval. 2009. Thermal remote sensing in Earth science research. In T. A. Warner, M. D. Nellis, and G. M. Foody (eds.), *The SAGE Handbook of Remote Sensing*, 64–78. London: SAGE Publications Ltd.
- Rebelo, L.-M., C. M. Finlayson, and N. Nagabhatla. 2009. "Remote sensing and GIS for wetland inventory, mapping and change analysis." *Journal of Environmental Management* **90** (7): 2144–2153. doi:10.1016/j.jenvman.2007.06.027.
- Rees, W. G. 1992. "Orbital subcycles for Earth remote sensing satellites." *International Journal of Remote Sensing* **13**: 825–834.
- Rees, W. G. 2006. *Remote Sensing of Snow and Ice*. Boca Raton, Florida: Taylor & Francis/CRC.
- Rees, W. G. 2008. "Comparing the spatial content of thematic maps." *International Journal of Remote Sensing* **29** (13): 3833–3844.
- Rees, W. G. 2012. "Automated spaceborne detection of degraded vegetation around Monchegorsk, Kola Peninsula, Russia." *Polar Record* **48**: 107–112.
- Rees, W. G. and Satchell, M. J. F. 1997. The effect of median filtering on synthetic aperture radar images. *International Journal of Remote Sensing*, **18**, 2887–2893.
- Remund, Q. P., D. G. Long, and M. R. Drinkwater. 2000. "An iterative approach to multisensor sea ice classification." *IEEE Transactions on Geoscience and Remote Sensing* **38**: 1843–1856.
- Renard, J.-B., G. Berthet, C. Brogniez, *et al.* 2008. "Validation of GOMOS-Envisat vertical profiles of O₃, NO₂, NO₃, and aerosol extinction using balloon-borne instruments and analysis of the retrievals." *Journal of Geophysical Research* **113**: A02302. doi:10.1029/2007JA012345.
- Richards, J. A., and X. Jia. 2006. *Remote Sensing Digital Image Analysis*. 4th edn. Berlin: Springer.
- Ries, J. C., R. J. Eanes, C. K. Shum, and M. M. Watkins. 1992. "Progress in the determination of the gravitational coefficient of the Earth." *Geophysical Research Letters* **19** (6): 529–531. doi: 10.1029/92GL00259.
- Robinson, I. S. 2010. *Measuring the Oceans from Space: The Principles and Methods of Satellite Oceanography*. Chichester, UK: Springer/Praxis Publishing.
- Rott, H., C. Mätzler, D. Strobl, S. Bruzzi, and K. G. Lennart. 1988. *Study on SAR Land Applications for Snow and Glacier Monitoring*. European Space Agency Contract Report. Innsbruck: Universität Innsbruck.
- Schanda, E. 1986. *Physical Fundamentals of Remote Sensing*. Berlin: Springer-Verlag.
- Schowengerdt, R. A. 2007. *Remote Sensing: Models and Methods for Image Processing*. 3rd edn. Academic Press.
- Segelstein, D. 1981. *The complex refractive index of water*. M.S. thesis, Kansas City: University of Missouri.
- Sharkov, E. 2003. *Passive Microwave Remote Sensing of the Earth: Physical Foundations*. Berlin: Springer/Praxis.

References

- Smith, D., C. Mutlow, J. Delderfield, B. Watkins, and G. Mason. 2012. "ATSR infrared radiometric calibration and in-orbit performance." *Remote Sensing of Environment* **116**: 4–16. doi:10.1016/j.rse.2011.01.027.
- Smith, M.J., and C.F. Pain. 2009. "Applications of remote sensing in geomorphology." *Progress in Physical Geography* **33** (4): 568–582. doi:10.1177/0309133309346648.
- Sobrino, J.A., and M.H. El Kharraz. 1999. "Combining afternoon and morning NOAA satellites for thermal inertia estimation. 2. Methodology and application." *Journal of Geophysical Research* **104** (D8): 9455–9465.
- Solberg, A.H.S., C. Brekke, and P.O. Husoy. 2007. "Oil spill detection in Radarsat and Envisat SAR images." *IEEE Transactions on Geoscience and Remote Sensing* **45** (3): 746–755. doi:10.1109/TGRS.2006.887019.
- Sonka, M., V. Hlavac, and R. Boyle. 2007. *Image Processing, Analysis, and Machine Vision*. 3rd edn. CL-Engineering.
- Stewart, R.H. 1992. *Methods of Satellite Oceanography*. Berkeley, California: University of California Press.
- Stiles, W.H., and F.T. Ulaby. 1982. "The active and passive microwave response to snow parameters. 1: wetness." *Journal of Geophysical Research* **85**: 1037–1044.
- Sun, J., and H. Kawamura. 2009. "Retrieval of surface wave parameters from SAR images and their validation in the coastal seas around Japan." *Journal of Oceanography* **65** (4): 567–577. doi:10.1007/s10872-009-0048-2.
- Tsiolkovsky, K.E. 1903. *Issledovanie Mirovych Prostranstv Reaktivnymi Priborami (The Exploration of Cosmic Space by Means of Reaction Devices)*. Kaluga.
- Ulaby, F.T., R.K. Moore, and A.K. Fung. 1981. *Microwave Remote Sensing: Active and Passive. Volume 1. Microwave Remote Sensing Fundamentals and Radiometry*. Reading, Massachusetts: Addison-Wesley.
- Ulaby, F.T., R.K. Moore, and A.K. Fung. 1982. *Microwave Remote Sensing: Active and Passive. Volume 2. Radar Remote Sensing and Surface Scattering and Emission Theory*. Reading, Massachusetts: Addison-Wesley.
- van de Hulst, H.C. 1981. *Light Scattering by Small Particles*. New York: Dover.
- Wagner, W., G. Lemoine, and H. Rott. 1999. "A method for estimating soil moisture from ERS scatterometer and soil data." *Remote Sensing of Environment* **70**: 191–207.
- Wan, Z., and J.A. Dozier. 1996. "Generalized split-window algorithm for retrieving land-surface temperature from space." *IEEE Transactions on Geoscience and Remote Sensing* **34**: 892–905.
- Waquet, F., B. Cairns, K. Knobelspiesse, et al. 2009. "Polarimetric remote sensing of aerosols over land." *Journal of Geophysical Research* **114** (D1). doi:10.1029/2008JD010619.
- Warner, T.A., M.D. Nellis, and G.M. Foody. 2009. *The SAGE Handbook of Remote Sensing*. London: SAGE Publications Ltd.
- Weissmann, M., and C. Cardinali. 2007. "Impact of airborne Doppler lidar observations on ECMWF forecasts." *Quarterly Journal of the Royal Meteorological Society* **133** (622): 107–116. doi:10.1002/qj.16.
- Willis, P., H. Fagard, P. Ferrage, et al. 2010. "The International DORIS Service (IDS): toward maturity." *Advances in Space Research* **45** (12): 1408–1420. doi:10.1016/j.asr.2009.11.018.
- Wooster, M.J., B. Zhukov, and D. Oertel. 2003. "Fire radiative energy for quantitative study of biomass burning: derivation from the BIRD experimental satellite and comparison to MODIS fire products." *Remote Sensing of Environment* **86** (1): 83–107.

References

- Wu, Y., and J. Wang. 2008. Detection of ship wakes in SAR images based on Radon transform. In *9th International Conference on Signal Processing, ICSP 2008*, 969–972. IEEE, October 26. doi:10.1109/ICOSP.2008.4697289.
- Wunsch, C., and D. Stammer. 1998. “Satellite altimetry, the marine geoid, and the oceanic general circulation.” *Annual Review of Earth and Planetary Sciences* **26** (1): 219–253. doi:10.1146/annurev.earth.26.1.219.
- Yang, H., and F. Weng. 2011. “Error sources in remote sensing of microwave land surface emissivity.” *IEEE Transactions on Geoscience and Remote Sensing* **49** (9): 3437–3442. doi:10.1109/TGRS.2011.2125794.
- Yorks, J. E., D. L. Hlavka, W. D. Hart, and M. J. McGill. 2011. “Statistics of cloud optical properties from airborne lidar measurements.” *Journal of Atmospheric and Oceanic Technology* **28** (7): 869–883. doi:10.1175/2011JTECHA1507.1.
- Young, I. R., S. Zieger, and A. V. Babanin. 2011. “Global trends in wind speed and wave height.” *Science* **332** (6028): 451–455. doi:10.1126/science.1197219.
- Zhang, Y. 2011. “Mean global and regional distributions of MOPITT carbon monoxide during 2000–2009 and during ENSO.” *Atmospheric Environment* **45** (6): 1347–1358. doi:10.1016/j.atmosenv.2010.11.044.

INDEX

- AATSR instrument 6.35
absolute temperature scale, absolute zero 1.7,
2.21, 2.23, 3.57
absorption 2.30, 3.3, 3.45, 3.59, 3.65, 3.80, 4.6,
4.13, 4.19, 5.27, 6.10, 6.22, 6.55, 6.60, 6.64,
7.21, 7.22, 11.27, 11.43
coefficient 3.19, 3.60, 4.17, 4.21, 4.29, 6.57,
6.59
cross-section 3.43, 3.50, 3.51, 3.60, 4.17
efficiency 3.44, 3.47,
length 3.6, 3.11, 3.43, 3.61, 3.88, 6.45, 6.55,
7.11
lines 3.54, 3.75, 4.6, 6.32, 6.55, 6.59, 7.20, 7.27,
9.2
resonant 3.59
accumulator array 11.58
accuracy
classification 11.49
consumer's (user's, reliability) 11.50
producer's (reference) 11.50
across-track direction 9.6, 9.13, 9.15, 9.30
active systems 1.7, 1.8, 8.1, 9.1, 9.36, 10.4,
ADEOS satellites 9.11
Advanced
Along-Track Scanning Radiometer (AATSR)
6.35
Land Observation Satellite (ALOS) 9.37
Microwave Scanning Radiometer for EOS
(AMSR-E) 7.11, 7.16, 7.17, 7.18
Microwave Sounding Unit – A (AMSU-A) 7.27
Scatterometer (ASCAT) 9.12, 9.14
Spaceborne Thermal Emission and Reflection
Radiometer (ASTER) 6.14, 6.27, 6.28, 11.32
Very High Resolution Radiometer
(AVHRR) 6.39
aerosols, atmospheric 4.12, 4.29, 5.28, 6.10, 6.19,
6.73, 7.21
measurement of 6.30, 6.69, 9.1
affine transformation 11.11
Africa 6.23
AFS index 5.4
agriculture, applications of remote sensing in 1.6,
5.32, 6.22, 6.24, 9.11, 9.37
air
mean molar mass 4.4
polarisability 3.9
refractive index 3.14
Airborne Thematic Mapper 10.4
aircraft 10.2
airlight 4.7
AIRS 4.9, 6.64
Alaska 6.39, 11.54
albedo 3.24, 3.55, 3.89, 6.49
diffuse 3.25, 3.88, 3.89, 6.72
measurement of 1.6, 3.71, 6.30, 9.1
albedometer 3.71
aliasing 2.17, 8.4, 8.22, 10.27, 11.23
Alissa LiDAR 9.2
along-track direction 7.26, 8.5, 8.10, 8.27, 8.35,
9.6, 9.13, 9.14, 9.17, 9.19, 9.22, 9.27, 9.43,
10.4, 10.35, 11.25
ALOS satellite 9.37
altimetric orbits 10.24, 10.36
ambiguity
distance 9.33
height 9.34
phase 9.34
range 8.4, 9.29,
9.43, 10.2

- amplitude
 of electromagnetic wave 2.5, 2.7, 2.32, 3.2, 3.5, 3.17, 3.33
 transmittance function 2.33
- AMSR-E 7.11, 7.16, 7.17, 7.18
- anaglyph 5.24
- Ångström relation, Ångström exponent 4.14
- angular frequency 2.3, 2.12, 3.7, 3.12, 3.53, 4.17, 4.23, 6.48,
- ANN 11.44
- anode 6.2
- Antarctica 8.11, 8.27, 8.32
- antenna 2.36, 2.38, 7.1, 8.15, 9.3
 dipole 7.4
 dish 7.1, 7.4, 7.6
 efficiency 7.5, 9.5, 9.43
 gain 7.5
 mechanical scanning of 7.8
 monopole 7.4
 phased 7.9
 rectangular 7.4
 temperature 7.2, 7.7
 Yagi 7.4
- anthocyanins 3.76
- Antoine's equation 4.1
- apogee 10.9
- Apollo 6 1.4
- apsis, apsides 10.10
- Aqua satellite 6.17, 6.64, 7.27
- archaeology, applications of remote sensing in 1.6, 5.34, 6.24, 6.52, 8.11, 8.33
- area
 (polygon), concept in GIS 11.60
 effective, of antenna 7.5, 7.35, 9.4
- argon, atmospheric 4.1
- Argon satellite programme 5.32
- Ariane-5 rocket 10.9
- Aristotle 1.1
- array, phased 7.9
- artificial neural network (ANN) 11.44
- ASA number 5.4
- ASCAT 9.12, 9.14
- ascending node 10.12, 10.21
- ASCII (text) image format 11.66
- ash, atmospheric 4.13
- asphericity, effect of Earth's on satellite orbits 10.14
- ASTER 6.14, 6.27, 6.28, 11.32
- astronomical unit (AU) 2.29, 11.9, A1
- astronomy, radio 7.1
- atan2 function 10.13
- Atlantic Ocean 6.29
- atmosphere
 absorption and scattering of radiation 4.6, 6.9, 6.22, 6.55, 6.60, 7.20, 8.3, 9.1, 9.17
 chemistry, measurement of 1.8, 6.59, 7.30, 7.32
 composition and structure 4.1, 4.12, 6.33, 6.36, 7.18
 convection 4.9
 density 4.2
 dispersion of VNIR radiation in 3.14
 downwelling radiation 7.21
 drag on satellite 10.5, 10.19, 10.30
 dry 4.1, 8.8, 8.30, 8.35, 8.36
 isothermal 4.5
 microwave brightness temperature 7.15
 Moon, absence of on 4.7
 optical thickness 4.7, 4.8, 4.10, 4.17, 5.27, 6.35, 6.56, 7.21, 7.30
 pressure 4.2, 4.4, 6.55, 6.57, 6.58, 7.30, 8.8, 8.22
 refractive index 3.14
 scale height 4.5, 4.26, 4.29, 6.57
 temperature profile 4.3, 7.20
 temperature sounding 1.8, 3.64, 6.55, 6.65, 6.69, 7.27, 7.32
 transparency of 1.7, 4.6, 4.7
 turbulence 4.26, 5.26, 5.28
 upwelling radiation 5.27, 7.19
 windows 1.7, 4.8, 1.9, 4.8, 6.55
 atmospheric correction 1.9, 4.1, 6.9, 6.19, 6.29, 6.34, 7.9, 7.11, 7.19, 8.7, 8.29, 11.7, 11.27

- Atmospheric Infrared Sounder (AIRS) 4.9, 6.64
 atom, hydrogen 3.54
 attenuation (extinction) 3.64, 3.50, 3.67, 3.83, 4.9,
 4.13, 4.18, 4.29, 5.27, 6.28, 6.55, 7.19, 8.33
 attribute, concept in GIS 11.60
 AU 2.29, 11.9, A1
 Aura satellite 7.32
 Australia 6.26, 6.42
 autocorrelation function 3.35, 3.37, 3.39
 averaging filter 11.19, 11.24, 11.72
 AVHRR 6.39
 azimuth
 defocussing 9.28
 direction 9.14, 9.27
 resolution 9.14, 9.17, 9.19, 9.21, 9.26, 9.30
 shift 9.27, 9.38
 back-propagation 11.45
 backscatter lidar 9.1
 backscatter ultraviolet observations 6.60, 6.68
 backscattering coefficient (microwave)
 3.23, 3.35, 3.38, 3.84, 9.4, 9.5, 9.9, 9.16,
 9.23, 9.36
 real materials 3.81, 9.43,
 Baikonur cosmodrome, Kazakhstan 10.7
 Baird, John Logie 1.3
 Balkan-1 8.10
 balloons, as platforms for remote sensing 1.2, 10.1
 band-gap 6.3, 6.31, 7.1
 banding, image 11.25
 bands interleaved by line (BIL) 11.66
 bands interleaved by pixel (BIP) 11.66
 band sequential (BSQ) format 11.66
 band transformation 11.26
 bandwidth 2.17, 6.8, 6.19, 7.8, 7.30, 8.19, 8.31,
 9.21
 barium sulphate, used as optical reflectance
 calibrator 3.28
 Barragém de Bravura, Portugal 11.55
 barrel distortion 5.13
 base-height ratio 5.24
 baseline 5.22, 5.22, 5.23, 5.36, 9.33, 9.34, 9.35
 bathymetry 1.6, 6.28, 9.37, 9.38
 Batut, Arthur 1.2
 Bayesian statistics 11.44
 beam-limited operation 8.18, 8.30
 beam solid angle (beamwidth) 7.2, 7.3,
 7.4, 7.23, 7.32, 8.3, 8.10, 8.18, 8.27, 9.1, 9.6,
 9.14, 9.20
 Beer-Lambert law 3.5
 Bessel function 2.34
 Betula nana 3.75
 Bhattacharyya distance 11.42
 bias, electromagnetic or sea-state 8.26
 bicubic interpolation 11.13
 bidirectional reflectance distribution
 function (BRDF) 3.23, 3.77, 3.89x2,
 6.9, 8.19
 bidirectional reflectance factor 3.28
 big-endian 11.66
 BIL 11.66
 bilinear interpolation 11.13
 binocular vision 5.24
 biomass, determination of 6.25, 7.15
 BIP 11.66
 bistatic
 radar equation 9.3
 scattering coefficient 3.23, 9.3
 Biwa, Lake (Japan) 9.37
 black body 2.23, 2.25, 2.29, 3.39, 3.62, 3.76, 6.32,
 6.33, 6.51, 6.72
 blue sky 4.7
 bolometer 6.31
 Boltzmann
 constant 2.24, 7.2, 7.6
 distribution 4.3
 boundary layer, atmospheric 4.12
 box classifier 11.38
 Bragg scattering 3.34
 Brewster angle 3.19, 3.78
 BRDF *see* bidirectional reflectance distribution
 function
 brightness
 in Kauth-Thomas transformation 11.28

- temperature [2.28, 2.38, 3.62, 6.33, 6.41, 6.43, 6.53, 6.55, 6.59, 6.72, 6.73, 7.3, 7.11, 7.12, 7.15, 7.16, 7.19, 7.25, 7.30, 7.35, 11.73]
 broadening of absorption lines [3.57]
 Brown model of radar altimeter waveform [8.15]
 BSQ format [11.66]
- C band [2.4, 9.12, 9.39]
 calibration of imagery [6.10, 11.6]
 CALIOP instrument [9.2]
 CALIPSO satellite [9.2]
 Cambridge, UK [6.14]
 camera obscura [1.1]
 candela [5.3]
 canonical components transformation (CCT) [11.35]
 Cape York, Australia [6.20]
 carbon dioxide
 absorption lines (infrared) [4.6, 6.55, 6.59]
 atmospheric [4.2, 4.5, 4.6, 6.58, 8.8]
 polarisability [3.9]
 profiling [1.6, 6.65, 6.67]
 carbon monoxide
 atmospheric [4.2]
 molecule [3.56]
 profiling [6.65, 6.67]
 carotene [3.76]
 cartography and surveying, applications of remote sensing in [1.3, 1.6, 5.34, 6.24, 8.11, 8.27, 9.37]
 charge-coupled device (CCD) [6.4, 6.13, 6.17, 6.69]
 chirp [8.31, 8.37]
 chlorophyll [3.73, 3.75, 6.18, 6.29, 11.27]
 circular polarisation [2.7, 7.26]
 civil engineering, application of remote sensing to [1.6]
 classification, image [6.25, 11.36]
 climatology, application of remote sensing to [1.6, 6.24, 6.30, 6.43]
 cloud
 albedo [3.73]
 cumulus [3.73]
 detection and monitoring of [1.6, 6.25, 6.30, 6.39, 6.41, 6.52, 9.1]
 discrimination from snow [11.29]
 effect on microwave radiation [3.80, 4.17, 4.29, 7.21, 9.36]
 effect on surface temperature [6.49]
 multiple (volume) scattering by [3.68, 3.73, 3.90]
 optical properties of [4.16]
 physical characteristics [3.73, 3.81, 4.16]
 profiling [6.53, 9.1]
 statistics [4.14]
 top altitude [8.10]
 top temperature [6.53, 7.30]
 VNIR reflectance spectrum of [3.74]
 clustering algorithms [11.40]
 coastal zone, applications of remote sensing in [1.6, 6.19, 6.24, 6.28, 9.38]
 coefficient
 absorption [3.60, 3.65, 3.68, 3.73, 3.81, 3.89, 4.17, 6.55, 6.59, 6.60, 6.72, 7.35]
 extinction (attenuation) [3.67, 4.9, 4.13, 4.29, 6.28, 6.55]
 Fresnel [3.18, 3.26, 3.35, 3.39, 3.72, 3.78, 7.11, 9.30]
 Mie [3.50]
 scattering and backscattering [3.23, 3.32, 3.65, 3.68, 3.69, 3.73, 3.81, 3.90, 4.16, 4.21, 6.60, 9.1, 9.4, 9.5, 9.8, 9.9, 9.16, 9.23, 9.36, 9.43]
 coherence
 between SAR images [9.35]
 length [8.19]
 time [9.21, 9.43]
 width [8.19]
 coherent radiation [2.9, 8.19, 9.20, 9.23]
 collimated radiation [2.19, 3.23, 3.60]
 collision broadening [3.58]
 colour shift [5.2]
 column integral [4.8]
 Committee on Earth Observation Satellites, image format [11.67]
 complex exponential notation [2.12, 3.5, 3.45, 6.51]
 compound lens [5.9]
 compression, compressed

- compression, compressed (cont.)
 archive 11.66
 image 11.5, 11.63
 pulse 8.31, 9.13
- concrete
 microwave backscattering from 3.82
 thermal infrared emissivity of 3.77
- conductivity
 electrical 3.11, 3.79, 3.88, 6.3, 7.11, 8.33
 thermal 6.45
- confusion matrix 11.49
- conical scan 7.8, 7.22, 7.26
- consumer's accuracy 11.50
- contrast
 as texture measure 11.48
 film 5.4
 modification, image 11.14
 scene 5.26
 stretch 11.15
- convolution
 spatial 11.22, 11.25, 11.56
 temporal 8.21
- Coriolis force 8.24
- corner-cube reflector 3.82, 10.32, 11.10
- Corona satellite programme 5.32, 11.1
- correlation length 3.35, 3.39
- cost
 of satellite imagery 1.11
 of satellite launch 1.6
- covariance matrix 11.31
- cross-correlation, spatial 11.57
- cross-polarised gradient ration 7.15
- cross-section
 absorption 3.43
 scattering 3.44
- cross-track direction 9.6, 9.13, 9.15, 9.30
- cumulus cloud 3.73
- current, ocean 1.6, 6.29, 8.24
- curvature, Earth 4.10, 8.18
- dark-pixel subtraction 6.10
- data volume 11.4, 11.72
- DCT 11.65, 11.68
- Debye equation 3.9, 4.17
- decibel 3.81, 7.4, 9.5, 9.8
- declination, Sun A3
- Defense Meteorological Satellite Program (DMSP) 6.1, 7.15, 7.22
- delta-function 2.14, 3.26, 3.33, 3.34, 9.7, 9.23
- density slicing 11.37
- depletion region 6.2
- depression angle 3.23
- desert, microwave backscattering from 3.82
- desriping (debanding) filter 11.25
- DFT 11.23
- DIAL 9.2
- dielectric constant 2.27, 3.2, 3.15, 3.36, 3.42, 3.44, 3.78, 4.17, 4.23, 7.11, 7.15, 9.7, 9.36
 contrast 3.22, 3.47
- differential absorption lidar (DIAL) 9.2
- diffraction 2.31, 3.41, 4.26, 5.12, 6.6, 7.1, 8.18, 8.35, 9.14, 9.38
- Fraunhofer 2.32, 3.32, 7.3, 7.6, 9.22
- Fresnel 2.35
- grating 6.8, 6.21, 6.62, 6.64, 6.68, 6.69
- diffuse albedo 3.25, 3.88, 3.89, 6.72
- diffusivity, thermal 6.48
- digital elevation model (DEM) 5.24, 6.24, 8.12, 9.37, 11.60
- digital photography 5.6
- digital number 11.6
- dihedral reflector 3.82
- dimensionality of image data 11.43
- DIN number 5.4
- dipole antenna 7.4
- Dirac delta-function *see* delta-function
- directivity 7.5
- disaster assessment 1.6, 5.34, 6.24, 9.37, 6.24
- discrete cosine transform (DCT) 11.65, 11.68
- discrete Fourier transform (DFT) 11.23
- discriminant function 11.39
- dish antenna 7.1, 7.6
- dispersion, wave 3.12, 4.23, 4.27, 6.8, 6.11, 8.5

- distortion, lens 5.13
 distortion in imaging radar 9.15, 9.26, 9.30, 9.38
 divergence, as measure of class separability 11.42
 DMSP 6.1, 7.15, 7.22
 DN 11.6
 Dnepr rocket 10.7, 10.9
 doping, semiconductor 6.2
 Doppler
 broadening 3.57, 3.90
 effect 2.18, 9.20, 9.26
 lidar 9.2
 Orbitography and Radio-positioning Integrated by Satellite (DORIS) 10.32
 scatterometer 9.6
 Dornier 228 5.30, 10.3
 downwelling 7.21
 dry atmosphere 4.1, 8.8, 8.30, 8.35, 8.36
 duration (oceanography) 8.25
 dust, atmospheric 4.13
 dye, sensitising 5.2
 dynamical form factor 10.14
 dynode 6.2

 Earth Observation 1.1
 East Anglia, UK 6.21
 eccentricity, satellite orbit 10.10, 10.15, 10.17, 10.35
 ECHO 11.55
 edge-detection 11.21, 11.54, 11.72
 effective area, antenna 7.5, 7.35, 9.4
 efficiency
 absorption 3.44, 3.47, 3.51
 antenna 7.5, 9.5, 9.43
 attenuation (extinction) 3.44
 scattering 3.44, 3.51
 EGM2008 geoid 8.23
 EHF band 2.4
 eigenvalues and eigenvectors of covariance matrix 11.32
 Einstein, Albert 2.5, 2.18, 3.15, 4.23
 elastic scattering 3.65
 electrically scanned array 7.9
 Electrically Scanned Microwave Radiometer (ESMR) 7.9
 electrical conductivity 3.11, 3.79, 3.88, 6.3, 7.11, 8.33
 electric permittivity 2.2, 3.1, 3.2, 8.36
 electromagnetic bias 8.26
 electromagnetic spectrum 2.3
 electronic transition 3.54
 electronvolt 2.5, 3.54, A2
 ellipse, orbital 10.9
 ellipsoid 8.22
 WGS84 8.23
 elliptical polarisation 2.8
 emissivity 1.8, 2.27, 2.30, 3.22, 3.24, 3.78, 6.35, 6.43, 6.47, 6.49, 6.51, 6.73, 7.15, 7.19, 7.21
 of real materials in the microwave region 3.77, 7.11, 7.13, 7.15,
 of real materials in the TIR region 3.76, 6.45, 6.52
 emulsion, photographic 5.1
 endianness 11.66
 end-member 11.47
 energy, as texture measure 11.48
 energy levels, quantum-mechanical 3.53
 enhancement, image 11.14
 entropy, as texture measure 11.48
 Envisat satellite 6.35, 6.69, 8.30, 10.35
 EO-1 satellite 6.21
 equation of time A3
 equatorial radius, Earth 10.14, 10.35, A2
 error
 matrix 11.49
 slope-induced 8.27
 ERS satellites 8.25, 9.17, 9.35
 ERTS satellite 1.4, 1.5
 ESMR 7.9
 ETM+ imager 6.5, 6.6, 6.32, 6.38, 6.44, 6.73, 11.16, 11.72
 Euclidean distance 11.39
 exactly repeating orbits 10.22, 10.35x2
 exitance
 luminous 5.4, 5.9, 5.27

- exitance (cont.)

 radiant 2.21, 2.25, 2.29, 3.25, 5.4

 exoatmospheric solar irradiance and radiance

 2.30, 11.9

 exp function 2.12

 exterior orientation 5.15

 extinction (attenuation) 3.64, 3.50, 3.67, 3.83, 4.9,
 4.13, 4.18, 4.29, 5.27, 6.28, 6.55, 7.19, 8.33

 extraction and classification of homogeneous
 objects (ECHO) 11.55

 false-colour infrared (FCIR) film 1.3, 5.2

 fan-beam 9.12, 9.13

 far field 2.35

 fast Fourier transform (FFT) 11.23

 feature space 11.43

 fetch, oceanographic 8.25

 film, photographic 5.1

 colour 5.2, 5.34

 contrast 5.4

 false-colour infrared 1.3, 5.2

 infrared 5.2

 mapping 5.4, 5.5, 5.7

 panchromatic 3.71, 5.2,

 construction of 5.1

 reconnaissance 5.5,

 spatial resolution 5.4, 5.7, 5.11,

 speed 5.4

 filtering, spatial 11.18

 fire, remote sensing of 6.24, 6.43?, 11.29

 flat binary image format 11.65

 flooding, monitoring of 6.24, 9.37

 fluorescence 3.59, 3.65

 flux

 density 2.5, 2.9, 2.38, 3.3, 3.5, 3.23, 3.43, 3.60,
 3.66, 7.6, 9.3,

 luminous 5.3

 f/number 5.9

 focal length 5.8, 6.7, 6.17, 6.38, 6.39, 6.42

 fog 4.14, 4.29, 7.21

 footprint 7.8, 7.11, 7.13, 7.23, 8.3, 8.6, 8.10, 8.17,
 8.22, 8.30

 forestry, applications of remote sensing in 1.2,
 1.6, 5.34, 6.24, 6.28, 8.7, 9.12, 11.29

 forward bias 6.3

 forward gain, antenna 7.5

 Fourier transform 2.12, 2.21, 2.38, 3.33, 6.63, 7.3,
 8.37, 9.24, 11.22, 11.26, 11.47, 11.57, 11.65

 discrete 11.23

 fast 11.23

 spectrometer 6.62

 Fraunhofer diffraction 2.32, 3.32, 7.3, 7.6, 9.22

 free space 2.1, 3.3

 impedance of 2.5

 wavelength 3.7, 3.88

 frequency

 angular 2.3, 2.12, 3.7, 3.12, 3.53, 4.17, 4.23,
 6.48

 Nyquist 2.17

 Fresnel, Augustin-Jean 2.35

 Fresnel,

 coefficients 3.18, 3.26, 3.35, 3.39, 3.72, 3.78,
 7.11

 diffraction 2.35

 distance 2.35, 7.3

 fringe, interference 6.62, 9.32, 9.34

 fully developed sea 8.25

 gain, antenna 7.5, 9.3

 Galaxy, microwave emission by 7.1, 7.21

 gallium arsenide 8.1

 gases

 dielectric constant of 3.7

 ideal 4.4

 well mixed 4.5, 4.10, 6.58, 7.31

 GCP 10.4, 11.10

 GeoEye 6.13

 geographic information system (GIS) 1.11, 11.60

 geoid 8.22

 geology, applications of remote sensing in 1.3,
 1.6, 5.34, 6.24, 6.50, 9.11

 geomagnetic investigations 1.1

 geometrical optics model 3.38

 geometrical correction 11.10

- geometrical scattering [3.49, 3.73, 3.81]
 geomorphology, applications of remote sensing in [1.6, 5.32, 6.24, 9.37]
 georeferencing [11.10]
 Geoscience Laser Altimeter System (GLAS) [8.10, 8.11]
 geostationary orbits and satellites [1.4, 6.6, 6.22, 6.42, 10.9, 10.15, 11.72]
 geostrophic balance [8.24]
 geosynchronous orbit [10.16]
 GeoTiFF format [11.67]
 germanium [6.3, 6.31]
 GIF [11.68]
 GIS [1.11, 11.60]
 glaciology, application of remote sensing to [1.6, 3.72, 5.31, 6.24, 7.11, 7.13, 7.14, 7.15, 7.18, 7.35, 8.10, 8.11, 8.13, 8.28, 8.32, 9.11, 9.18, 9.35, 9.38, 9.39, 11.57]
 GLAS [8.10, 8.11]
 GLCM [11.48]
 Global Navigation Satellite Systems (GNSS) [10.18]
 Global Ozone Monitoring by Occultation of Stars (GOMOS) [6.69]
 Global Positioning System (GPS) [10.3, 10.18, 10.32]
 GLONASS [10.19]
 GMT [10.21]
 GOES satellites [10.15]
 GOMOS [6.69]
 Google Earth [1.5, 1.9]
 GPS [10.3, 10.18, 10.32]
 GRACE mission [1.1, 8.23]
 gradient ratio (GR) [7.36]
 granite [3.77]
 graphics interchange format (GIF) [11.68]
 gravitational constant, geocentric [10.10]
 gravitational potential, Earth's [1.1, 8.22, 10.14]
 gravity measurements [1.1, 1.6]
 Great Barrier Reef [6.20]
 great circle [10.13]
 Greenland Ice Sheet [7.15, 7.18, 8.27, 8.32]
 greenness, in Kauth–Thomas transformation [11.28]
 Greenwich Mean Time (GMT) [10.21]
 grey level [11.6]
 grey level cooccurrence matrix (GLCM) [11.48]
 ground control point (GCP) [10.4, 11.10]
 ground penetrating radar [8.33]
 ground state [3.54]
 ground station [11.1]
 group velocity [3.13, 4.23, 8.2, 8.7, 8.29, 8.33]
 Gulf Stream [8.24]
 Hadamard transform [11.65]
 half-power beam width (HPBW) [7.3]
 half-wave dipole antenna [7.4]
 Hartley transform [2.17, 11.65]
 haze, atmospheric [4.13, 5.28]
 heat capacity [6.45]
 Heat Capacity Mapping Mission (HCMM) [6.51]
 heat-loss surveying [6.43]
 height resolution (sounding systems) [6.58, 6.60, 6.67, 6.70, 7.32]
 Heisenberg uncertainty principle [3.57]
 helium, liquid [6.31]
 hemispherical albedo *see* albedo, diffuse
 Henyey–Greenstein term in surface scattering [3.26, 3.28]
 Herschel, William [1.2]
 Hertz, Heinrich [1.2]
 HF band (radio frequencies) [2.2, 4.25]
 hidden layer [11.44]
 hierarchical data format (HDF) [11.67]
 high-boost filter [11.24, 11.72]
 highlighting [9.16]
 high-pass filter [11.25, 11.72]
 high resolution imager
 TIR [6.38]
 VNIR [6.14]
 High Resolution Infrared Radiometer (HRIR) [6.51]
 High Resolution Infrared Sounder (HIRS) [6.59]
 high resolution visible (HRV) [6.5]

- histogram, image [11.14]
 equalisation [11.16]
 matching [11.16]
 Hohmann transfer orbit [10.8]
 Hotelling transformation [11.32]
 horizontal polarisation [3.18, 3.35, 7.14]
 Hough transforms [11.58]
 hour angle, Sun [A3]
 HPBW [7.3]
 Huffman coding [11.64]
 humidity, atmospheric [4.1]
 hybrid classification [11.41]
 hydrogen
 atom [3.54]
 molecule [3.57]
 polarisability of [3.9]
 hydrology, applications of remote sensing in [5.34]
 Hyperion imager [6.21]
 hyperspectral
 classification [11.43]
 imager [6.21, 11.43, 11.47]
 ice
 albedo [3.72]
 emissivity (microwave) [7.35]
 emissivity (thermal infrared) [3.77]
 monitoring of *see* glaciology
 refractive index [3.72]
 thermal inertia and diffusivity [6.50]
 ICESat satellite [8.10]
 ideal gas [4.4]
 identity operator, in spatial filtering [11.19]
 IFOV [6.5, 6.17, 6.38, 7.8, 7.35]
 Ikonos [6.13]
 illuminance [5.3, 5.9]
 atmospheric [5.27, 5.36, A2]
 image [1.9, 11.6]
 classification [6.25, 11.33]
 formation [5.8, 6.4, 7.8, 9.20]
 processing [1.9, 1.11, 5.7, 5.16, 5.24, 11.5]
 imaging system [1.8, 4.26, 5.1, 6.1, 6.32, 7.22, 9.13, 11.1]
 impedance [3.3, 3.18]
 of free space [2.5, A1]
 incidence angle [3.20, 3.23, 3.35, 3.42, 3.81, 3.89, 7.22, 9.5, 9.6, 9.7, 9.9, 9.14, 9.30, 9.36, 9.43]
 inclination, satellite orbit [10.12, 10.16, 10.17, 10.20, 10.23, 10.25]
 incoherent
 average [9.26]
 radiation [2.9, 8.15]
 Indian Space Research Organisation (ISRO) [7.26]
 indium antimonide [6.2, 6.3, 6.31, 6.39]
 inertia
 moment of [3.56]
 thermal [6.47, 6.73]
 inertial sensor [10.4]
 information, Shannon [11.63, 11.72]
 infrared radiation [1.7, 2.1, 2.4, 2.29, 3.7, 3.71, 4.6, 4.13, 4.19, 6.1, 7.27, 8.1, 8.12, 9.2, 10.22, 11.7, 11.27],
 discovery of [1.1]
 thermal [1.7, 2.26, 2.29, 3.22, 3.56, 3.71, 3.76, 6.30, 6.55, 7.11, 10.16, 11.37]
 film [1.3, 5.1, 5.2, 5.34]
 input layer [11.44]
 instantaneous field of view (IFOV) [6.5, 6.17, 6.38, 7.8, 7.35]
 integral equation model [3.42]
 integral transform [2.17]
 intensity, luminous [5.3]
 interference, interferogram [6.62, 8.19, 9.32]
 interferometric SAR [9.30, 9.37, 9.43]
 interior orientation [5.16]
 International Satellite Cloud Climatology Programme (ISCCP) [4.15]
 International Space Station [5.13, 5.32]
 interpolation, image [11.13]
 inverse problem [6.58]
 ionosphere [3.12, 4.22, 8.30, 8.36, 10.1]
 irradiance [2.20, 2.29, 3.23, 3.25, 5.3, 9.3]
 irreversible (lossy) image compression [11.65]

- ISO number 5.4
 isodata algorithm 11.40, 11.55
 isotropic
 radiation 2.21, 3.27
 scattering 3.27, 3.35
 spatial filter 11.21
 ISRO 7.26
 Japan Advanced Meteorological Imager (JAMI) 6.42
 Jason-1 and-2 satellites 8.26, 10.32, 10.36
 Jeffries-Matusita (JM) distance 11.42
 Johnson noise 7.2
 Joint Photographic Expert Group (JFIF) 11.68
 JPEG image format 11.68
 K band 2.4
 Ka band 2.4
 kappa statistic 11.51
 Karhunen–Loëve transformation 11.32
 Kauth–Thomas transformation 11.28, 11.30
 kernel 11.22, 11.56, 11.72
 KFA-1000 5.31
 Ku band 2.4, 8.30
 kelvin 2.22
 Kepler elements 10.12
 Keyhole camera 5.32
 Kirchhoff, Gustav 3.38
 Kirchhoff
 law of radiation 2.28
 scattering model 3.38
 kites, as platforms for remote sensing 1.2
 K-means algorithm 11.40
 Komi Republic, Russia 11.59
 Kyoto, Japan 9.37
 L band 2.4, 7.12, 9.33
 Labruguière, France 1.3
 Lakselv, Norway 8.12
 Lambert, Johann Heinrich 3.27
 Lambertian scattering 3.26, 6.72, 9.8, 11.9
 land cover and land use mapping 6.24
 Landsat satellites 1.4, 4.14, 6.5, 10.22, 10.22, 10.23, 10.30, 11.3
 enhanced thematic mapper (ETM+) 6.5, 6.6, 6.32, 6.38, 6.44, 6.73, 11.16, 11.72
 multispectral scanner (MSS) 11.25, 11.28
 thematic mapper (TM) 6.5, 11.12, 11.25
 Lanyard satellite programme 5.32
 Laplace transform 2.17
 Laplacian operator (Laplace filter) 11.20, 11.24, 11.72
 lapse rate, atmospheric 4.5
 Large Format Camera 5.31
 laser profiling (laser altimetry) 1.8, 8.1, 8.18
 scanning 8.4, 8.35
 waveform 8.6, 8.11
 laser ranging to satellites 10.32
 layer (theme) concept in GIS 11.61
 layover 9.16, 9.36
 lead sulphide 6.2, 6.3, 6.31
 Lempel–Ziv–Welch algorithms 11.64, 11.68
 lemniscate 10.16
 length modulated radiometer 6.66
 lens
 compound 5.9
 simple 5.8
 LEO 10.9, 10.19
 LibRadTran 6.10
 lidar 1.8, 8.1, 9.1
 light, speed of 2.3, 2.18, 2.23, 3.3, 3.57, 4.23, 8.2, 8.7, 8.15, 8.19, 8.30, 9.6, 9.21, A1
 limb-sounding 6.60, 6.69, 6.72, 7.32
 limestone, VNIR reflectance spectrum of 3.75
 linear mixture modelling 11.46
 line
 concept in GIS 11.60
 pair 5.4
 scanner 6.5
 little-endian 11.66
 Lockheed ER2 10.3
 lock, radar altimetry 8.26
 looks, in radar image 9.26
 look-up table (LUT) 11.18

- lossless (reversible) image compression 11.63
 loss tangent 3.4
 lossy (irreversible) image compression 11.65
 low Earth orbit (LEO) 10.9, 10.19
 lowlighting 9.16
 low-pass filter 11.24
 low resolution VNIR imagers 6.17
 LOWTRAN 6.10, 6.35
 lumen 5.3
 luminance 5.3
 luminosity, solar 2.30
 luminous flux, intensity and exitance 5.3
 lunar occultation 6.62
 LUT 11.18
 lux 5.3
 LZ and LZW compression 11.64, 11.68
 magnesium oxide, used as optical reflectance calibrator 3.28
 magnetic permeability 2.2, 3.2
 main lobe 7.3
 major axis 10.10
 majority filter 11.55
 Manhattan 8.13
 manoeuvres, orbital 10.29
 map, thematic 11.36
 mask, image 11.37
 maximum likelihood classifier 11.39
 Maxwell, James Clerk 1.2, 2.2
 Maxwell's equations 2.2, 3.3, 3.18
 mean filter 11.19, 11.24, 11.72
 Measurement of Pollution in the Troposphere (MOPITT) 6.66
 median filter 11.25
 media, storage, for digital data 11.4
 medical imaging 1.1
 medium resolution imager
 TIR 6.39
 VNIR 6.17
 mercury cadmium telluride (MCT) 6.31, 6.38, 6.39, 6.42
 mercury-doped germanium 6.31
 mesophyll 3.76, 11.27
 mesosphere 4.4
 metadata, image 11.65
 metals
 dielectric constant 3.10
 emissivity (thermal infrared) 3.77
 thermal inertia and diffusivity 6.50
 meteorology, application of remote sensing to 1.6, 4.18, 6.22, 6.30, 6.43, 6.55, 8.1, 10.12, 10.15,
 Meteor-M satellite 6.17
 METEOSAT 1.2, 6.22, 10.15
 Second Generation (MSG) 6.22
 methane, atmospheric 4.6, 6.65, 6.67
 Metop satellite 7.27, 7.31, 9.12
 Metric Camera 5.31
 Michelson interferometer 6.62
 microwave
 radiation 1.7, 2.1, 2.4, 2.10, 2.17, 2.29, 2.36, 3.18, 3.22, 3.32, 3.57, 3.71, 3.77, 4.7, 4.17, 4.21, 4.23, 4.25, 4.29, 7.1, 8.14, 9.1
 region, validity of Rayleigh–Jeans approximation in 2.25
 scatterometry 9.5
 Microwave Humidity Sounder (MHS) 7.31
 Microwave Imaging Radiometer with Aperture Synthesis (MIRAS) 7.9, 7.12, 7.26
 Microwave Limb Sounder (MLS) 7.32
 Midori satellite 9.11
 Mie, Gustav 3.49
 Mie
 coefficients 3.50
 scattering 3.49, 3.81, 4.21,
 migrating means algorithm 11.40
 military applications of remote sensing 1.1, 1.3, 1.6, 5.2, 5.32, 6.13, 9.13, 9.17, 10.20, 10.31,
 minerals
 albedo 3.73
 emissivity (thermal infrared) 3.77
 thermal properties 6.50
 VNIR reflectance spectra of 3.75

- minimum noise fraction (MNF) 11.43
 Minnaert model of surface scattering 3.26, 3.28, 3.89
 minor axis 10.10
 Mir (space station) 5.31, 8.10, 9.2
 mixed pixel 11.46
 mixture modelling 11.46
 MODIS 4.13, 6.17, 6.20, 6.29, 6.35, 6.39, 6.46, 6.52, 11.46
 MODTRAN 6.10, 6.35
 modulation transfer function (MTF) 5.5, 5.12
 molar mass, atmospheric 4.4
 Molniya orbit 10.17
 moment of inertia 3.56
 monopole antenna 7.4
 monostatic radar equation 9.4
 Moore, William 10.6
 MOPITT instrument 6.66
 Morocco 6.18
 mosaic, image 11.10
 Moscow 5.33, 11.3
 motion blurring 6.4
 MrSID format 11.67
 MSG 6.22
 MSS 11.25, 11.28
 MSU-50 6.17
 MTS-2 satellite 6.42
 Multifrequency Scanning Microwave Radiometer (MSMR) 7.26
 multi-look processing 9.26
 multiple scattering 3.68, 3.73, 3.76, 3.80, 3.82, 3.89, 9.7, 11.27
 multispectral classification 11.38, 11.55, 11.63
 Multispectral Scanner 11.25, 11.28
 multitemporal remote sensing 9.12
 Nadar (Tournachon) 1.2
 nadir 6.15, 6.22, 6.35, 6.55, 6.61, 6.64, 7.8, 7.12, 7.27, 7.32, 8.27, 9.1, 11.2, 11.9,
 National Aeronautics and Space Administration (NASA) 1.4, 6.21, 7.11, 7.36, 10.3
 National Oceanographic and Atmospheric Administration (NOAA) 6.39, 6.52, 6.68, 7.27, 7.31
 Navstar satellites 10.19
 nearest-neighbour resampling 11.13
 neodymium doped yttrium aluminium garnet (Nd: YAG) 8.1
 near field 2.35
 near-infrared 2.4
 negative, photographic 5.1
 network, concept in GIS 11.60
 neural network *see* artificial neural network
 NIMBUS-5 satellite 7.9
 nitrogen
 atmospheric 4.1
 liquid 6.31
 oxides of 6.69
 NOAA satellites *see* POES
 nodal period, satellite 10.10, 10.14
 node, concept in GIS 11.60
 nodes, ascending and descending 10.12, 10.21
 noise
 reduction in an image 11.18
 system 7.8
 thermal (Johnson, Nyquist) 7.2
 nonesuch, in Kauth–Thomas transformation 11.28
 non-linear spatial filters 11.25
 non-polar materials, dielectric properties of 3.9
 non-selective scattering 3.51, 3.73, 3.81
 Noril'sk, Russia 11.28
 normalised
 burn ratio (NBR) 11.29
 difference snow index (NDSI) 11.29
 difference vegetation index (NDVI) 11.27, 11.37
 difference water index (NDWI) 11.29
 normalised spatial filter 11.19
 North America 7.16
 Norway 11.12
 n-type 6.2
 Nyquist
 noise 7.2

- Nyquist (cont.)
 frequency 2.17
 sampling theorem 2.17
- objective lens 5.10, 6.6, 6.17, 6.38, 6.39, 6.42, 6.70
 oblique aerial photography 5.12
 oblique path through atmosphere, optical length of 4.10
 occultation 6.62, 6.69
 ocean colour 1.6, 6.19, 6.28, 6.30
 ocean currents 1.6, 6.29, 8.24
 oceanography, applications of remote sensing in 1.6, 6.18, 6.29, 7.10, 7.18, 9.9, 9.18, 9.38
 Oceansat 7.26
 OCM-2 6.19
 Odin satellite 7.32
 Okean satellites 9.18
 okta 4.14
 Operational Line Scanner (OLS) 6.1
 optical
 depth 3.65
 thickness 3.62, 3.63, 3.73, 3.80, 3.82, 3.84, 4.7, 4.13, 6.30, 6.56, 7.15, 7.20, 7.27, 7.30
 order of spectrum 6.8
 orientation, exterior 5.15
 orthophotograph orthoimage 5.24, 11.14
 output layer 11.44
 oversampling 5.7
 oxygen, atmospheric 3.58, 4.1, 4.6, 6.55
 polarisability of 3.9
 profiling 6.69
 60 GHz absorption line 4.7, 7.20, 7.27
 119 GHZ absorption line 7.20, 7.27
 ozone, atmospheric 1.6, 4.5, 4.6, 6.55, 6.60, 6.65, 6.69, 7.32
- P band 2.4
 PALSAR instrument 9.37
 pancake ice 11.57
 panchromatic photography 3.71, 5.2
 Papua New Guinea 6.20
 PAR 11.27
 paraboloidal dish antenna 7.1, 7.4
 parallel polarisation 3.17
 parallelepiped classifier 11.38
 partial pressure 4.1, 4.5
 passive microwave radiometer 2.28, 2.36, 3.22, 7.1
 passive systems 1.7, 5.1, 6.1, 7.1
 PCT 11.32
 perceptron 11.44
 perigee 10.9
 period, nodal (satellite) 10.10, 10.14
 permeability, magnetic 2.2, 3.2
 permittivity, electric 2.2, 3.2
 perpendicular polarisation 3.17
 per-pixel classifier 11.46
 phase
 function, scattering 3.48, 3.69
 velocity 3.3, 3.13, 4.23, 8.30
 phased array 7.9, 7.35
 photocathode 6.2
 photoconductive 6.3, 6.38
 photodiode 6.2, 6.31, 8.1
 photoelectric effect 6.2
 photogrammetry 5.15
 photography 5.1
 digital 1.9, 5.1, 5.6
 infrared 5.1, 5.2, 5.34
 oblique 5.12
 photometric units 5.2
 photomultiplier 6.1, 6.68
 photon 2.5, 5.1, 6.2, 6.31, 7.1
 photosynthetically active radiation (PAR) 11.27
 photovoltaic 6.3, 6.42
 phytoplankton 6.19, 6.29
 pigments, plant 3.73, 3.75, 3.76, 6.18, 6.29, 11.27
 pincushion distortion 5.13
 pinhole optics 5.13
 pitch, aircraft 10.4

- pixel [5.7, 5.16, 6.5, 11.5]
 value [11.6]
 planetary reflectance [11.9]
 Plan Position Indicator [1.2, 9.13]
 Planck, Max [2.23]
 Planck
 constant [2.5, 2.24, 3.53, A1]
 integral [2.25]
 law [3.53]
 radiation law [2.23, 2.29, 2.38, 6.32, 7.1]
 plane polarisation [2.7]
 planetary
 imaging [1.1]
 reflectance [11.9]
 plankton [1.6, 6.19, 6.29]
 plasma, plasma frequency [3.12, 3.15, 4.23]
 PNG format [11.68]
 POES satellites [6.39, 6.52, 6.68, 7.27, 7.31]
 point
 concept in GIS [11.60]
 principal [5.13, 5.17, 5.20, 5.21, 5.36]
 polar materials, dielectric properties of [3.9]
 polar orbit [10.12]
 polarisability [3.7, 3.45]
 polarisation [1.8, 2.6, 2.38, 3.3, 3.17, 6.30, 7.11,
 7.12, 9.5, 9.36, 9.39]
 ratio [7.36]
 polygon, concept in GIS [11.60]
 Porsangmoen, Norway [8.6, 10.4]
 portable network graphics (PNG) format [11.68]
 power pattern [7.3]
 PPI [1.2, 9.13]
 precession, orbital [10.14]
 precipitation *see* rain, snow
 precipitation radar [4.22]
 preprocessing, image [11.6]
 pressure
 distribution, atmospheric [4.4]
 broadening [3.58, 3.90]
 modulated radiometer [6.64, 6.66]
 partial [4.1, 4.5]
 principal components [11.30]
 transformation [11.32, 11.43]
 principal point [5.13, 5.17, 5.20, 5.21, 5.36]
 prism [6.8, 6.72]
 processing layer [11.44]
 producer's accuracy [11.50]
 prograde orbit [10.8, 10.12]
 pseudocolour image [11.37]
 p-type [6.2]
 pulse compression [8.31, 9.13]
 pulse-limited operation [8.18, 8.35, 9.7]
 pulse repetition frequency (PRF) [8.3, 8.30]
 pulse rise time [8.2]
 pushbroom imaging [6.5, 6.13, 6.21]
 pyroelectric detector [6.31]
 quantum detector [6.31]
 quantum number [3.54, 3.55, 3.56]
 RA-2 radar altimeter [8.30]
 radar [1.3]
 altimeter [1.8, 8.14, 9.7]
 equation [9.3, 9.5, 9.43]
 stereogrammetry [9.36]
 Radarsat [9.39]
 radiance [2.20]
 radiant exitance *see* exitance
 radiation resistance [7.2]
 radiative transfer equation [3.59, 3.90, 6.55, 7.19]
 radiative transfer models, atmospheric
 [6.10, 6.35]
 radio astronomy [7.1]
 radio echo-sounding [8.32]
 radiometric correction [11.7]
 radiosonde [6.35]
 radio waves, discovery of [1.2]
 rain [3.81, 4.18]
 effect on microwave radiation [3.81, 4.21, 7.12,
 7.21, 9.11]
 effect on VNIR radiation [4.19]
 physical properties of [4.18, 4.29]
 remote sensing of [1.6]

- rain (cont.)
 radar [3.81, 4.22]
 rainbow [3.51, 4.19]
 random polarisation [2.8, 3.20]
 range
 ambiguity [8.4]
 as texture measure [11.47]
 direction [9.15, 9.19]
 slant [9.14]
 walk [9.28]
 ranging systems [1.5, 8.1]
 raster representation [11.60]
 ratio vegetation index (RVI) [11.27]
 Rayleigh, John William Strutt, third Lord [3.29]
 Rayleigh
 distribution [9.25, 9.43]
 roughness criterion [3.29, 3.35, 9.7]
 scattering [3.45, 3.50, 4.6, 4.8, 4.14]
 Rayleigh–Jeans approximation [2.25, 2.28, 3.62, 6.36, 7.20]
 real aperture radar [9.13]
 reconnaissance film [5.5]
 rectilinear propagation [5.13]
 red edge [3.76]
 reduced mass [3.54]
 reference accuracy [11.50]
 reflectance, planetary [11.9]
 reflection coefficient *see* Fresnel coefficients
 reflectivity [1.8, 2.28, 3.24, 3.76, 3.89, 5.27, 8.3]
 refractive index [3.3, 3.4, 3.43, 3.88]
 of air [3.14]
 of water [3.72]
 contrast [3.22]
 region-growing [11.55]
 relaxation time [3.9]
 relay satellite [11.3]
 reliability accuracy [11.50]
 relief displacement [5.18, 5.36]
 remote sensing, definition and history of [1.1]
 repeat-orbit interferometry [9.34]
 resampling, image [11.13]
 residence time, atmospheric water [4.18]
 resistance, radiation [7.2]
 resolution
 azimuth [9.15, 9.17, 9.19]
 radiometric [9.13]
 range [8.2, 8.20, 8.35, 9.15]
 slant range [9.15]
 spatial [5.4, 5.7, 5.11, 6.6, 6.32, 7.1, 7.15, 8.3, 8.18, 9.15, 9.21]
 spectral [6.7, 6.32, 6.62]
 vertical (sounding systems) [6.58, 6.60, 6.67, 6.70, 7.32]
 resonant
 absorption [3.59]
 scattering [3.59]
 retrograde orbit [10.12]
 reverse bias [6.3]
 reversible (lossless) image compression [11.63]
 rezel [6.5, 6.7, 6.13, 6.17, 11.5]
 rise time, pulse [8.2]
 Ritter, Johann [1.1]
 RLSBO [9.18]
 roads, detection of [1.6, 3.82, 6.24, 9.37]
 Roberts filter [11.21]
 rocket
 dynamics [10.5]
 multiple-stage [10.9]
 roll, aircraft [10.4]
 rotation, molecular [3.56]
 roughness, surface [3.35, 3.77, 3.79, 9.12, 9.36]
 of ocean [3.82, 7.11, 8.25, 9.9, 9.38]
 row-major order [11.65]
 rule, for image classification [11.38]
 run-length encoding [11.64]
 RVI [11.27]
 S band [2.4, 8.30]
 Sainte-Croix, Lac de, France [11.16]

- salinity, effect on microwave emissivity of sea water [3.79, 7.11x2]
 salt, multiple scattering by [3.68]
 SAM [11.43]
 sampling theorem [2.17]
 sand
 emissivity (thermal infrared) [3.77]
 microwave backscattering coefficient [3.82]
 SAR [9.15, 9.16, 9.19, 10.1, 11.7],
 interferometry [9.30]
 satellite [10.5]
 relay [11.3]
 saturated vapour pressure [4.1]
 SBUV [6.68]
 scalar approximation [3.38, 3.89]
 scale, photographic [5.10]
 scale height, atmospheric [4.5, 4.26, 4.29, 6.57]
 Scanning Multichannel Microwave Radiometer (SMMR) [7.15]
 scattering
 and backscattering coefficient [3.23, 3.32, 3.65, 3.68, 3.69, 3.73, 3.81, 3.90, 4.16, 4.21, 6.60, 9.1, 9.4, 9.5, 9.8, 9.9, 9.16, 9.23, 9.36, 9.43]
 cross-section [3.44, 3.65]
 elastic [3.65]
 geometric [3.9]
 matrix [9.5]
 Mie [3.49, 3.81, 4.21]
 molecular [3.58]
 multiple [3.68, 3.73]
 non-selective [3.51]
 Rayleigh [3.45, 3.50, 4.6, 4.8, 4.14]
 selective [3.48]
 volume [3.68, 3.73]
 zone [8.15, 9.14]
 scatterometry, microwave [9.5]
 sea ice *see* ice
 sea level [8.26]
 sea state [3.79, 8.25]
 sea-state bias [8.26]
 sea surface [8.19]
 microwave backscattering coefficient [3.82, 9.9, 9.43]
 microwave emission from [3.79, 7.10]
 temperature [1.6, 6.42, 6.45, 7.10, 11.37]
 topography [8.10, 8.22, 8.25]
 sea water
 absorption length (microwave) [3.88, 7.11]
 dielectric constant of [3.10, 3.88]
 emissivity (microwave) [3.78, 7.35]
 emissivity (thermal infrared) [3.77, 6.45]
 salinity of [3.79]
 SeaWinds scatterometer [9.11]
 sediment, suspended [6.19, 6.28, 6.30]
 Seebeck effect [6.31]
 seed pixel [11.55]
 segmentation, image [11.54, 11.59]
 seismic investigations [1.1]
 semiconductor diode [6.2]
 semimajor and semiminor axes, satellite orbit [10.10]
 sensitivity [6.32, 7.7]
 separability of image classes [11.41]
 SEVIRI imager [6.22]
 shadowing, radar [9.16]
 Shannon information [11.63, 11.72]
 shapefile [11.68]
 sharpening filter [11.19, 11.21, 11.24, 11.72]
 Shenandoah river [11.33]
 SHF band [2.4]
 ships, detection of [9.38]
 short-wave infrared [6.15]
 Shuttle Radar Topography Mission (SRTM) [9.37]
 sidelobe [7.3]
 side-looking (airborne) radar (SLR, SLAR) [9.13]
 sidereal day [10.15]
 signature, spectral [3.74]
 significant wave height (SWH) [8.25]
 signal to noise ratio [8.2]
 silicon [6.4, 6.13]
 sinc function [2.15, 2.33]
 size parameter, dimensionless [3.45, 3.73, 4.21]

- sky
 colour of 4.7
 illuminance of 5.27
 skylight 4.7, 6.9
 slant range 9.14
 distortion 9.15
 SLAR 9.13
 slicks, detection of 9.38
 slope-induced error 8.27
 SLR 9.13
 small perturbation model 3.32
 SMMR 7.15
 smoke, atmospheric 4.13
 smoothing filter 11.19, 11.24, 11.72
 SMOS satellite 7.9, 7.12
 SMR 7.32
 Snell's law of refraction 3.17, 3.83, 6.8
 snow
 albedo 3.73
 discrimination from cloud 11.29
 emissivity (microwave) 3.80
 emissivity (thermal infrared) 3.77
 microwave backscattering coefficient 3.82x2,
 3.83
 monitoring of 1.6, 6.24, 7.15, 7.30, 8.10, 8.12,
 9.12, 9.38
 optical thickness of 9.36
 VNIR reflectance spectrum of 3.75
 volume (multiple) scattering by 3.68, 3.72,
 3.83, 3.89
 water equivalent (SWE) 7.17
 SNR 8.2
 Sobel filter 11.21, 11.24
 soil
 albedo 3.73
 emissivity (microwave) 3.79
 emissivity (thermal infrared) 3.77
 microwave scattering from 9.12
 moisture 1.6, 3.79, 5.34, 6.50,
 7.15, 9.11, 9.37
 sounding radar 8.33
 VNIR reflectance spectrum of 3.75, 6.22
 solar
 activity 4.25
 elevation 5.27
 occultation 6.62
 radiation 1.7, 2.1, 2.29, 4.5, 5.27, 5.36, 6.9,
 6.33, 6.72, 11.9
 Solar Backscatter Ultraviolet Radiometer (SBUV)
 6.68
 solid angle
 beam 7.2
 definition of 2.19
 sonar investigations 1.1
 soot, atmospheric 4.13
 sounding, atmospheric 1.8, 3.64, 6.55, 6.59, 6.65,
 6.69, 7.27, 7.30, 7.32
 South America 6.67
 Spacelab 5.31
 Space Shuttle 5.31, 9.37, 111.1
 Space Station *see* International Space Station, Mir
 spatial
 averaging 11.18
 dependency matrix 11.48
 filtering 11.18, 11.72
 frequency 3.34, 5.6, 11.23
 resolution 5.4, 5.7, 5.11, 6.6, 6.32, 7.1, 7.15,
 8.3, 8.18, 9.15, 9.21
 Special Sensor Microwave Imager (SSM/I) 7.13,
 7.15x2, 7.17, 7.18, 7.22
 Sounder (SSMIS) 7.22, 7.27
 Special Sensor Microwave Temperature Sounder
 (SSM/T) 7.22
 specific heat capacity 6.45
 speckle 9.23, 9.43, 11.25
 spectral
 angle mapping (SAM) 11.43
 radiance 2.22, 3.62, 6.43, 11.73
 reflectance 3.74
 resolution 6.7, 6.32, 6.62
 signature 3.74
 Spectralon 3.28
 spectrometer
 grating 6.8, 6.21, 6.62, 6.64, 6.68, 6.69

- prism 6.8
- spectroradiometer 3.74
- spectrum
 - electromagnetic 2.3
 - calculated using Fourier transform 2.6
 - reflectance 3.74, 11.43
- specular reflection 3.26x2, 3.21, 9.7
- speed, film 5.4
- spheroid, Earth 8.22, 10.14
- spin-scan imaging 6.6, 6.22
- split and merge 11.56
- split window technique 6.35
- SPOT satellites 6.5, 11.59
- SRTM 9.37
- SSM/I 7.13, 7.15x2, 7.17, 7.18, 7.22
- SSMIS 7.22, 7.27
- SSM/T 7.22
- SST 6.42, 6.45, 7.10, 11.37
- starlight 6.69
- stationary phase model 3.38, 3.89
- station-keeping, satellite 10.29
- station mask 11.2
- Stefan–Boltzmann constant 2.26
- Stefan’s law 2.26
- stellar occultation 6.62, 6.69
- step-stare imaging 6.4
- steradian 2.19
- stereogrammetry 5.15, 5.21
 - radar 9.36
- stereophotography 1.3, 5.21, 5.36
- Stokes vector 2.9, 2.38, 3.3, 3.88, 9.5
- storage media for digital data 11.4
- stratosphere 4.3
- striping, image 11.25
- structure function 4.26
- subcycle, orbital 10.24
- submillimetre radiation 7.32
- Submillimetre Radiometer (SMR) 7.32
- sub-pixel classification 11.46
- sub-satellite track 10.12
 - direction 10.25
- sulphide minerals 3.65
- sulphur dioxide, atmospheric 6.65
- Sun, radiation from *see* solar radiation
- Sun, position of A3
- sun-synchronous orbit 10.20, 10.35
- supervised classification 11.38
- support vector machine (SVM) 11.45
- surface, as concept in GIS 11.60
- Svalbard 5.31, 8.13, 9.17
- SWE 7.17
- Sweden 11.12
- SWH 8.25
- SWIR 6.15
- synthetic aperture radar 9.15, 9.16, 9.19, 10.1, 11.7
 - interferometry 9.30
- system noise 7.7
- tagged image file format (TIFF) 11.67
- tandem SAR interferometry 9.35
- TanDEM-X mission 9.38
- tangent plane approximation 3.38
- tasselled-cap transformation 11.29
- TDRS 11.3
- TEC 4.25, 8.37
- tectonic motion 1.6
- temperature
 - anomaly 6.52
 - antenna 7.2
 - cloud-top 6.53, 7.30
 - distribution within atmosphere 4.4
 - land surface 1.6, 6.43, 7.15
 - sea surface 1.6, 6.43, 7.10, 11.37
 - sounding (atmospheric) 1.8, 3.64, 6.55, 6.65, 6.69, 7.27, 7.32
- template matching 11.56
- TerraSAR-X/TanDEM-X mission 9.35
- Terra satellite 6.14, 6.17, 6.66, 10.31
- tesseral addressing 11.64
- text (ASCII) image format 11.66
- texture classification 11.47
- thematic map 11.36

- Thematic Mapper (TM) 6.5,
 11.12, 11.25.
 theme (layer) concept in GIS 11.61
 thermal
 conductivity 6.45
 detector 6.31
 diffusivity 6.48
 emission 2.21
 inertia 6.47, 6.73
 infrared radiation (TIR) 1.7, 2.4, 2.26, 2.29,
 3.22, 3.56, 3.76, 6.15, 6.17, 6.22, 6.30, 6.55,
 6.62, 6.64, 6.67, 6.72, 7.11, 10.16, 11.37,
 11.73
 noise 7.2
 thermistor bolometer 6.31
 thermocouple 6.31
 thermopile 6.31
 thermosphere 4.3
 tides, ocean 1.6, 8.22, 10.26, 10.36
 TIFF (tagged image file format) 11.67
 time
 domain scatterometry 9.7
 equation of A3
 Greenwich 10.21
 universal 10.21
 TIROS-1 satellite 1.4
 TM 6.5, 11.12, 11.25,
 Tokyo 6.16
 TOPEX radar altimeter 8.26
 topography
 land surface 1.6, 5.24, 6.24, 8.10, 8.27, 9.37,
 11.60
 sea surface 1.6, 8.10, 8.22
 Torres Strait 6.20
 total electron content (TEC) 4.25, 8.37
 total precipitable water 4.8, 8.9
 Tournachon, Gaspard-Félix 1.2
 tracking and data relay satellite (TDRS) 11.3
 tracking, radar altimetry 8.26, 8.31
 training data 11.38, 11.45, 11.49
 transfer function
 in artificial neural network 11.44
 in image contrast modification 11.14
 transfer orbit 10.8
 transformation, band 11.26
 transformed divergence 11.42
 transition, electronic 3.54
 transmission coefficient *see* Fresnel coefficients
 transparency 3.63
 trench, ocean floor 8.23
 trihedral reflector 3.82, 10.32, 11.10
 troposphere 4.3, 4.29
 Tsiolkovsky, Konstantin 10.6
 turbulence
 atmospheric 4.26
 ionospheric 4.27
 two-look technique 6.35, 6.72
 two-stream model 3.65, 3.89
 type 1/ type 2 waters 6.29
 UHF 2.4
 ultraviolet radiation 1.2, 1.7, 2.4, 3.11, 3.54, 3.65,
 4.5, 4.6, 4.23, 5.2, 6.55, 6.60, 6.68
 discovery of 1.1
 uncertainty principle 2.16, 3.57
 United States Air Force 5.5
 universal time (UT) 10.21
 unmanned aerial vehicle (UAV) 10.1
 unmixing, spectral 11.46
 unpolarised radiation 2.8, 3.20
 unsupervised classification 11.40
 upwelling 2.21, 7.20
 urban areas
 albedo of 3.73
 emissivity (thermal infrared) 3.77
 mapping and monitoring of 1.6x2, 5.34, 6.24,
 8.11
 microwave backscattering coefficient 3.82
 user's accuracy 11.50
 Van Allen belts 10.1, 10.20
 variance, as texture measure 11.47
 variance–covariance matrix 11.31
 vector representation 11.60

- vegetation
 albedo of 3.73
 emissivity (microwave) 3.78
 emissivity (thermal infrared) 3.77
 fluorescence of 3.65
 index 3.59, 11.27
 microwave backscattering coefficient 3.82, 9.11
 mapping and monitoring of 1.3, 1.6, 5.6, 5.34, 6.25, 6.29, 7.15, 8.6, 8.10, 8.11, 9.11, 9.37
 VNIR reflectance spectrum of 3.75, 5.2
 pigments 3.73, 3.75, 3.76, 6.18, 6.29, 11.27
- velocity
 group 3.13, 4.23, 8.2, 8.7, 8.29, phase 3.3, 3.13, 4.23, 8.30
 vertical polarisation 3.18, 3.20, 3.78, 7.12, 7.14, 7.15, 9.5
 vertical resolution (sounding systems) 6.58, 6.60, 6.67, 6.70, 7.32
 very high resolution imagers 6.13
 VHF band 2.2, 8.33
 vibration, molecular 3.56
 Virginia, USA 11.32
 visibility, interference fringe 6.63
 visible radiation 1.2, 1.7, 1.9, 2.4, 2.26, 3.14, 3.22, 3.51, 3.55, 3.65, 3.68, 3.71, 4.6, 4.7, 4.13, 4.14, 4.26, 9.36, 11.29
- Visible and Infrared Spin-Scan Radiometer (VISSR) 6.22
 visible and near-infrared radiation (VNIR) 1.7, 3.71, 4.6, 4.8, 4.16, 4.19, 5.1, 6.1, 8.1, 11.32
 vision, binocular 5.24
 volcanoes, monitoring of 6.24, 6.33, 8.12, 9.35
 volume scattering 3.68, 3.73, 3.76, 3.80, 3.82, 3.89, 9.7, 11.27
- wakes, detection of 9.38
 water
 absorption coefficient 3.6, 3.73, 6.45, 7.11
 albedo of 3.72
 bodies, remote sensing of 1.6, 6.24, 6.28
 dielectric constant of 3.10, 3.79, 4.17, 7.11
 emissivity (microwave) 7.12
 emissivity (thermal infrared) 3.77, 6.45
 reflection of VNIR radiation from 3.20, 3.72, 3.74, 6.10, 8.12
 refractive index of 3.22, 3.72
 sphere, scattering and absorption by 3.51, 3.73, 3.80, 4.21
 thermal inertia and diffusivity 6.50
 type 1/ type 2 6.29
 VNIR reflectance spectrum of 3.75
 water vapour, atmospheric 1.6, 3.58, 4.1, 4.8, 4.17, 6.34, 6.55, 8.8, 8.21
 absorption of electromagnetic radiation by 4.6, 4.8, 6.32, 7.20, 7.31
 column integral of 4.8
 delay of laser pulses by 8.8, 8.35
 delay of microwave pulses by 8.30, 8.36
 partial pressure 4.1
 polarisability of 3.9
 profiling 6.65, 6.69, 7.30, 7.31, 7.32x2
 residence time in atmosphere 4.17
 saturated vapour pressure of 4.1
 total precipitable 4.8, 4.18, 8.9, 8.30
 waves, ocean 1.6, 7.12, 8.20, 8.25, 9.9, 9.28, 9.38
 waveform
 laser profiler 8.6, 8.11
 radar altimeter 8.15, 8.26, 8.30
 wavelength 2.3
 free space 3.7, 3.88
 wavelet transform 2.17, 11.65, 11.68
 wavenumber 2.3, 3.12, 3.55
 weather radar *see* rain radar
 website 1.10
 weighting function 6.56, 6.61, 7.27, 7.30, 7.35
 WGS84 ellipsoid 8.23, 8.24
 whiskbroom imaging 6.5, 6.17, 6.38, 6.39, 8.4
 width, coherence 8.19
 Wien's law 2.26
 wind lidar 9.2

- wind speed and velocity [1.6] 2, 3.79, 3.82, 6.30, 7.12, 8.25, 9.9, 9.43
- windows, atmospheric 1.7, 4.8, 6.55
- wood, thermal properties 6.50
- work function 6.2
- World Reference System (WRS) 10.22
- X band 2.4, 3.81
- xanthophyll 3.76
- XPGR 7.15
- X-rays 4.23
- Yagi antenna 7.4
- yaw, aircraft 10.4
- yellowness, in Kauth–Thomas transformation 11.28
- Zenit satellite programme 5.32
- zenith angle 4.9



Figure 3.24. The angular distribution of light in a rainbow (here a double rainbow can be seen) indicates the complexity of scattering when the size parameter is large.

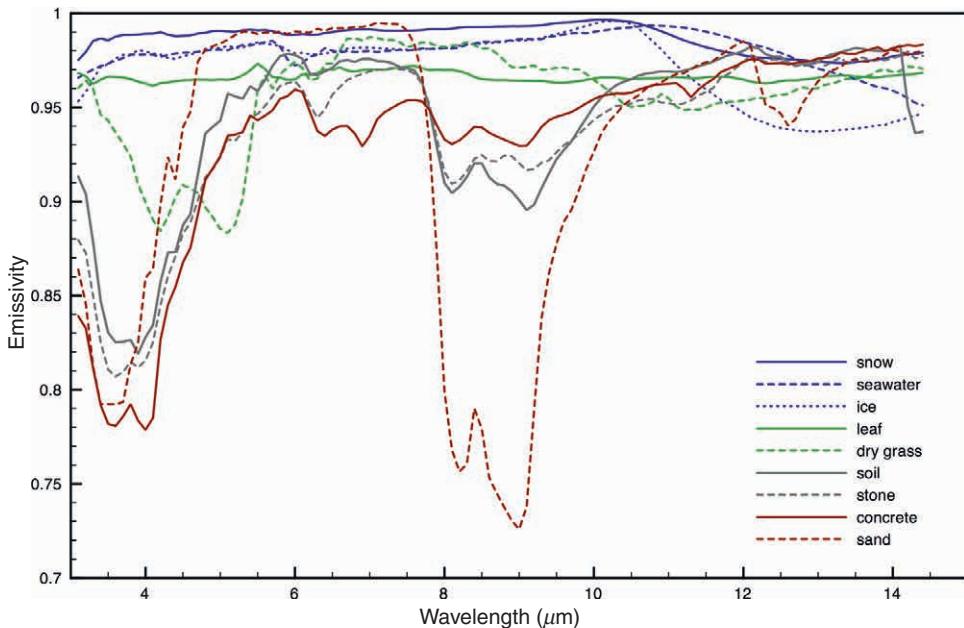


Figure 3.38. Typical emissivities of various materials at normal incidence in the range 3–15 μm . The data have been averaged from a large number of MODIS samples (Anon). The category ‘concrete’ in fact also includes a wide variety of building materials such as brick, tile and asphalt. The figure excludes metal surfaces, which have much lower emissivities.

AIRS TOTAL PRECIPITABLE WATER VAPOR (mm), May 2009

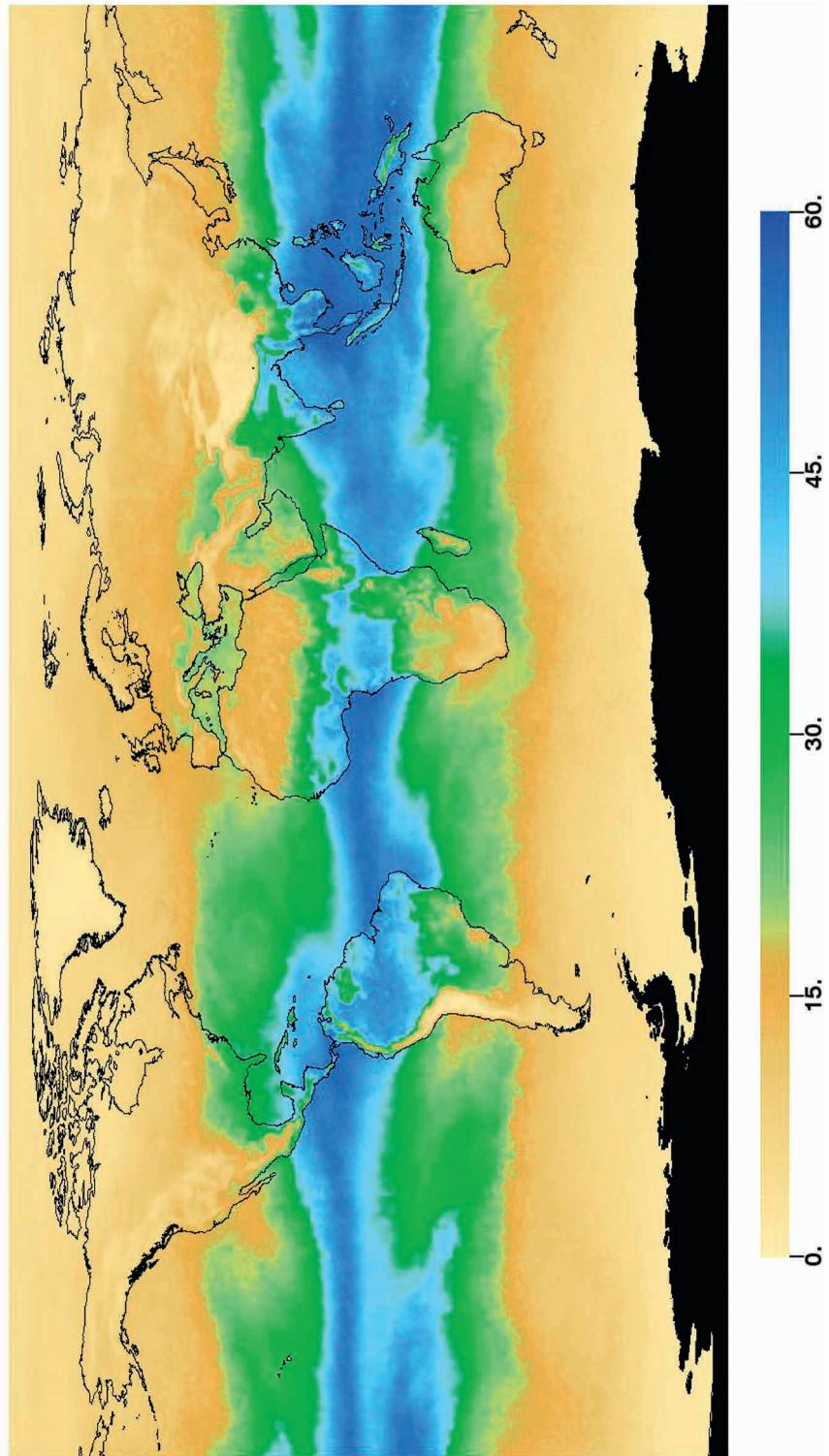


Figure 4.5. Total Precipitable Water in May 2009, derived from AIRS satellite data. (Figure reproduced by courtesy of NASA.
Source: NASA JPL <http://photojournal.jpl.nasa.gov/catalog/PIA12097>)

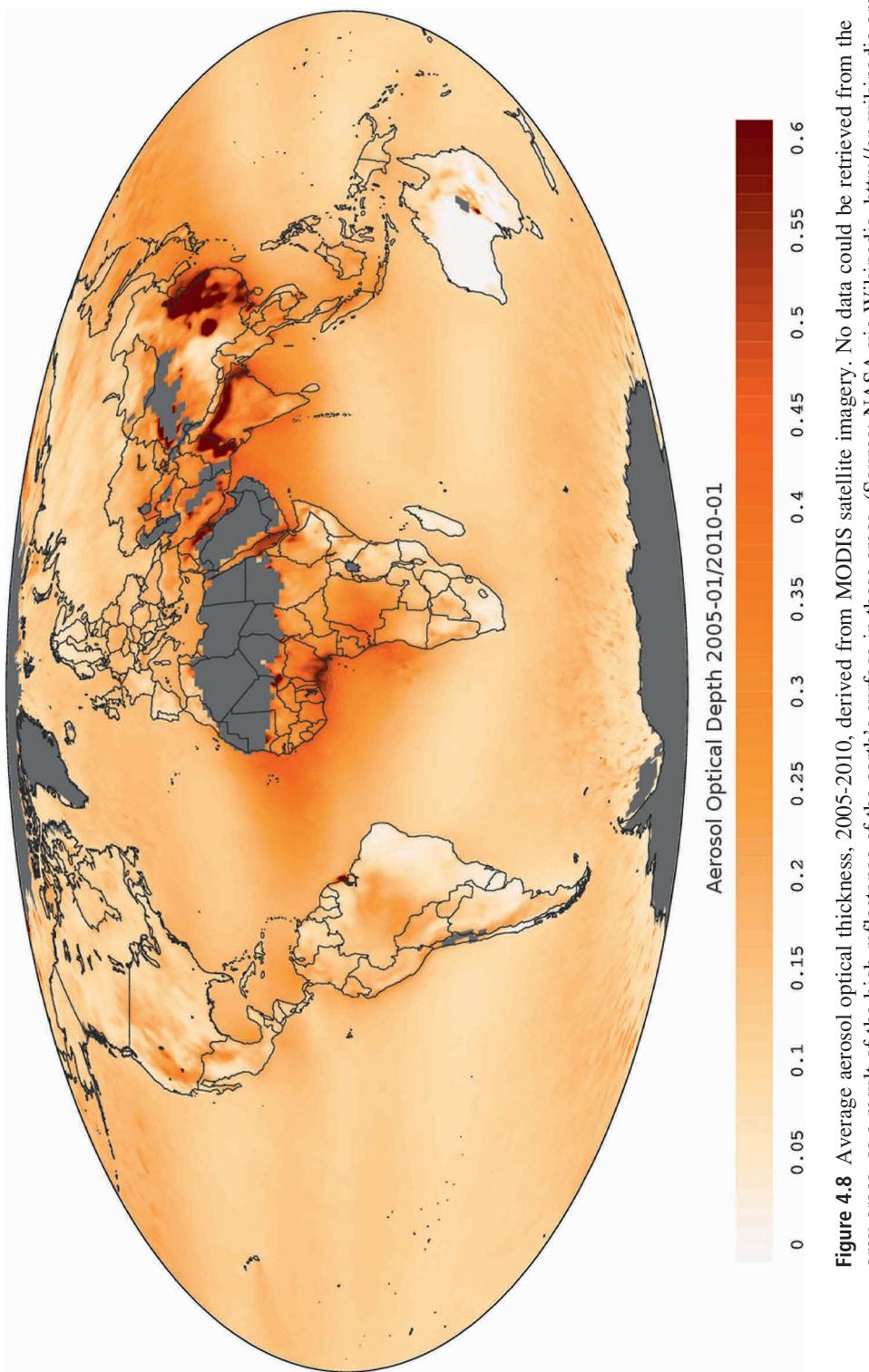


Figure 4.8 Average aerosol optical thickness, 2005-2010, derived from MODIS satellite imagery. No data could be retrieved from the grey areas, as a result of the high reflectance of the earth's surface in these areas. (Source: NASA via Wikipedia. http://en.wikipedia.org/wiki/File:Modis_aerosol_optical_depth.png).

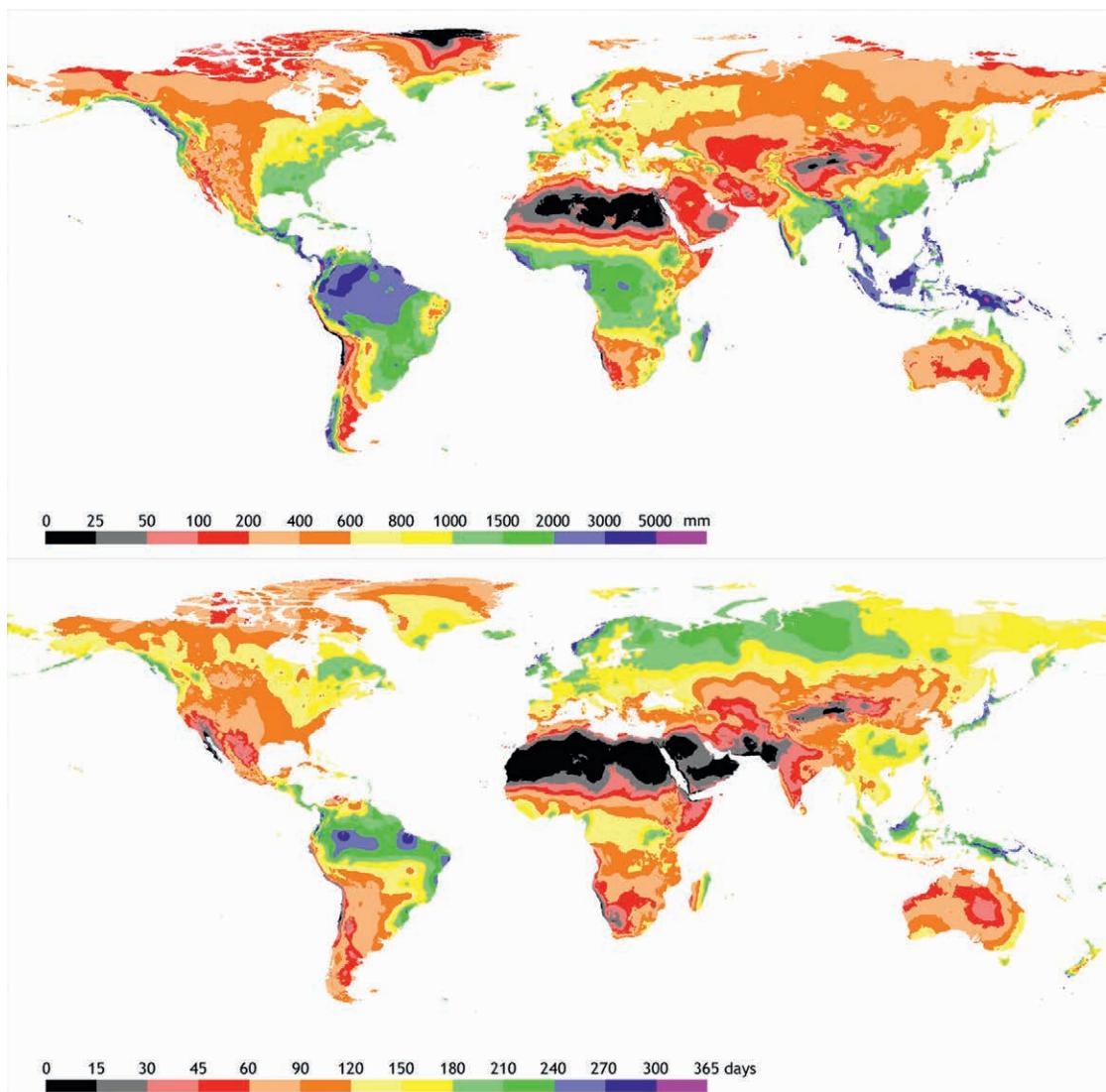


Figure 4.11. Global distribution of average annual rainfall (top) and average number of rainy days (bottom). The figures have been calculated from data compiled from terrestrial monitoring stations by the Climatic Research Centre, University of East Anglia, UK (CRU TS 2: (Mitchell and Jones 2005)).



ISS013E77377

Figure 5.9. An oblique photograph, showing wide coverage and variable scale. The photograph (ISS013-E-77377) was acquired from the International Space Station on 5 September 2006 and shows the Jungfrau, Mönch and Eiger mountains in the Bernese Alps, Switzerland. Photograph credit: NASA. (http://www.nasa.gov/mission_pages/station/expeditions/expedition17/earthday_imgs.html).

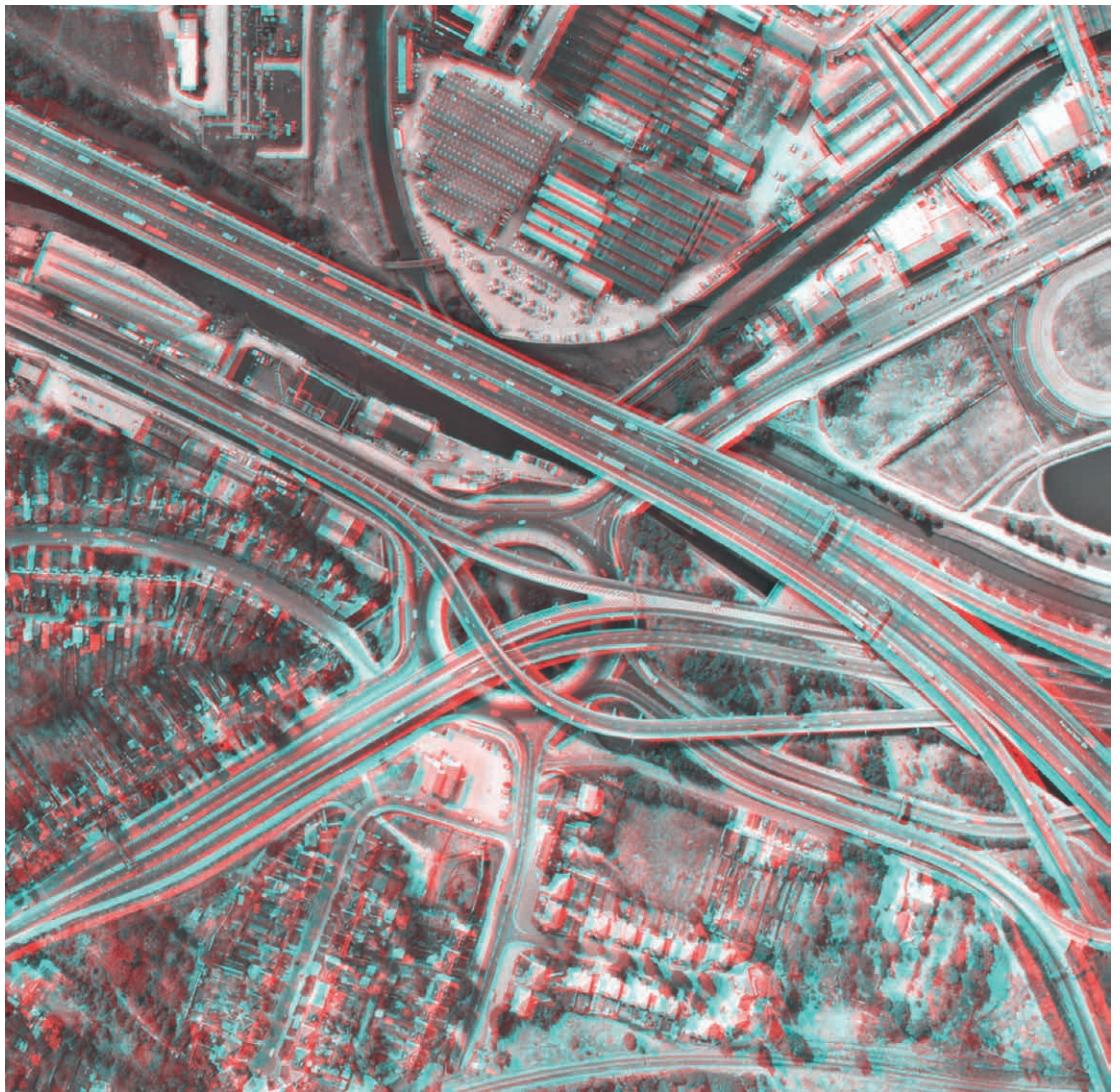


Figure 5.17. Anaglyph created from the stereophotographs of figure 5.15. If the anaglyph is viewed with a red filter over the left eye and a cyan filter over the right eye, the three-dimensional effect can be seen.



Figure 5.20. Mapping camera photograph showing the upper part of a glacier in Svalbard. Photograph reproduced by courtesy of Natural Environment Research Council, UK.



Figure 6.11. Extract of true colour GeoEye image showing the centre of Cambridge, UK, recorded on 4 September 2005. (Image reproduced by courtesy of GeoEye.)



Figure 6.12. Extract of FCIR ASTER image showing Tokyo, Japan, on 22 March 2000. Image by courtesy of NASA/GSFC/METI/ERSDAC/JAROS and the US/Japan ASTER Science Team. (<http://asterweb.jpl.nasa.gov/gallery-detail.asp?name=Tokyo>).



Figure 6.13. True-colour MODIS image of Morocco, acquired on 3 August 2000. (Image courtesy of Jacques Descloitres, MODIS Land Group, NASA GSFC. Image downloaded from NASA Visible Earth at <http://visibleearth.nasa.gov>.)

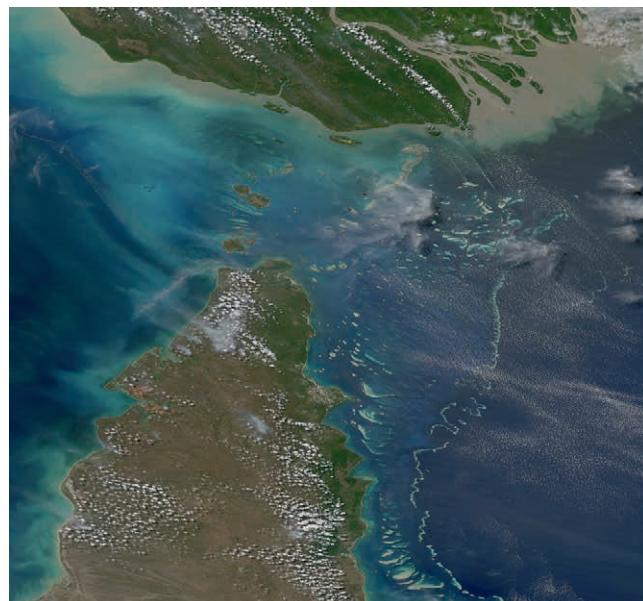


Figure 6.14. Extract of MODIS ocean colour imagery. The image was acquired on 9 August 2011 and shows Papua New Guinea, the Torres Strait, and the Cape York Peninsula, Australia. The Great Barrier Reef is clearly visible to the east of the peninsula. (Original image downloaded from <http://oceancolor.gsfc.nasa.gov> and reproduced by courtesy of NASA.)

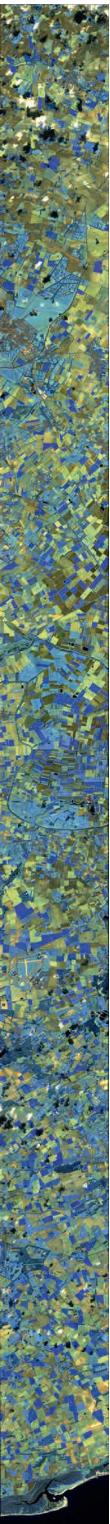


Figure 6.15. Hyperspectral imagery from East Anglia, UK, recorded on 21 October 2008. The strip is 7.7 km wide and 108 km long and extends from the Norfolk coast (left) to near Saffron Walden (right).

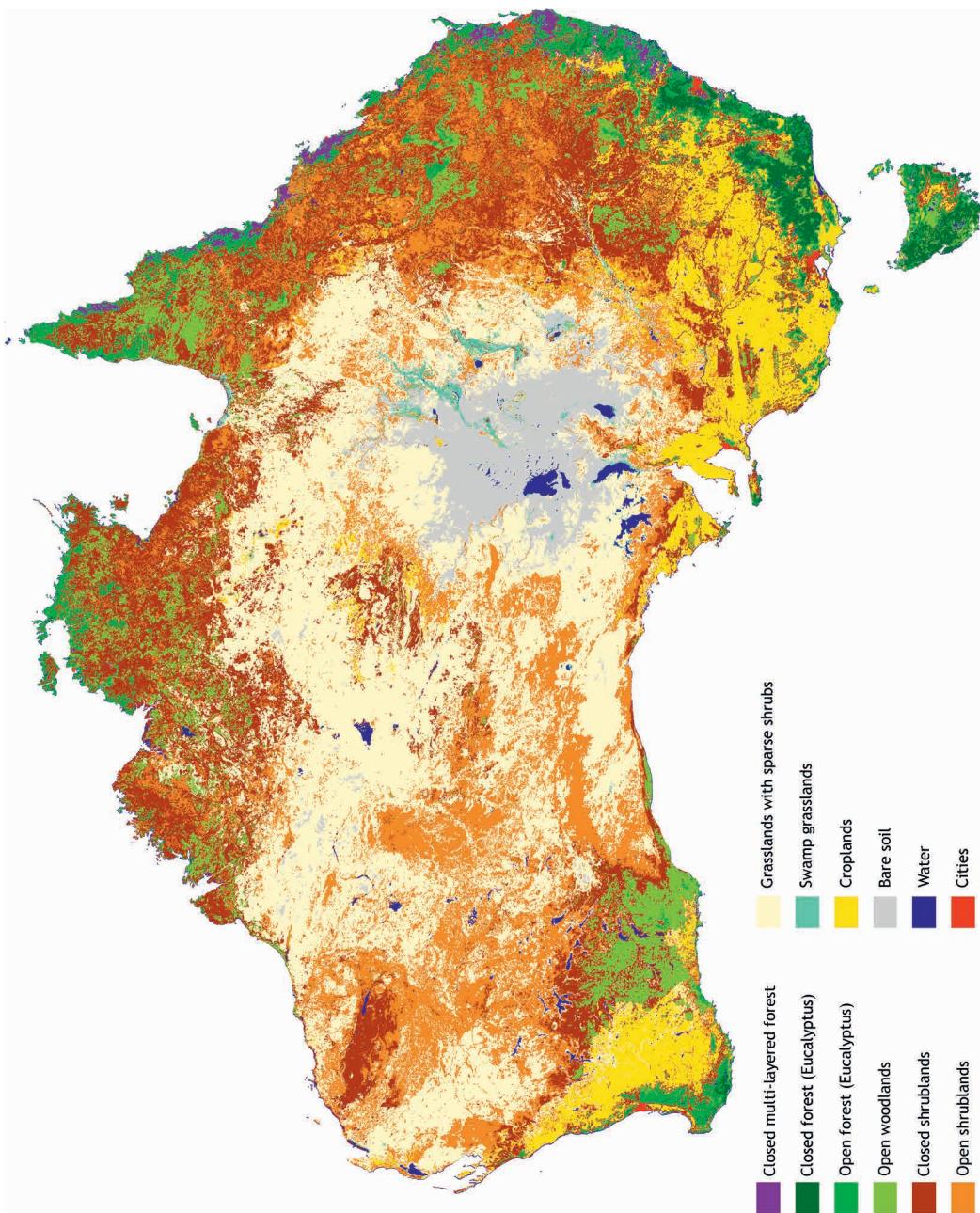


Figure 6.18. Part of the Global Land Cover 2000 dataset, constructed from data collected by the Vegetation instrument on the Spot-4 satellite. The dataset has a spatial resolution of 1 km. (Data downloaded from <http://bioval.jrc.ec.europa.eu/products/glc2000/products.php>. Citation: (Mayaux and Bossard 2000).)

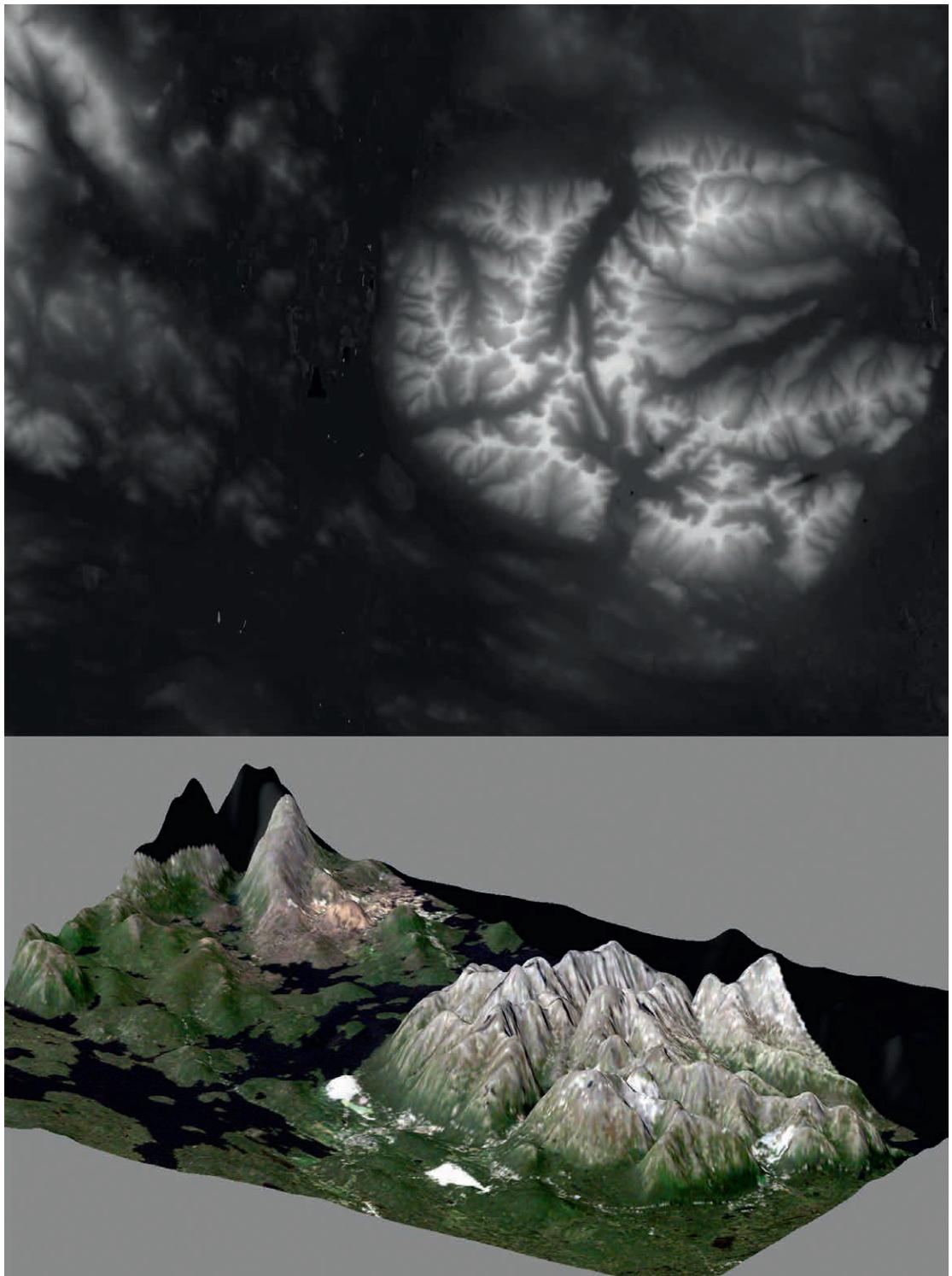


Figure 6.19. Above: Greyscale representation of a Digital Elevation Model DEM derived from ASTER imagery. Below: Terrain visualised by draping a true-colour Landsat image over the DEM. (Rees 2012).

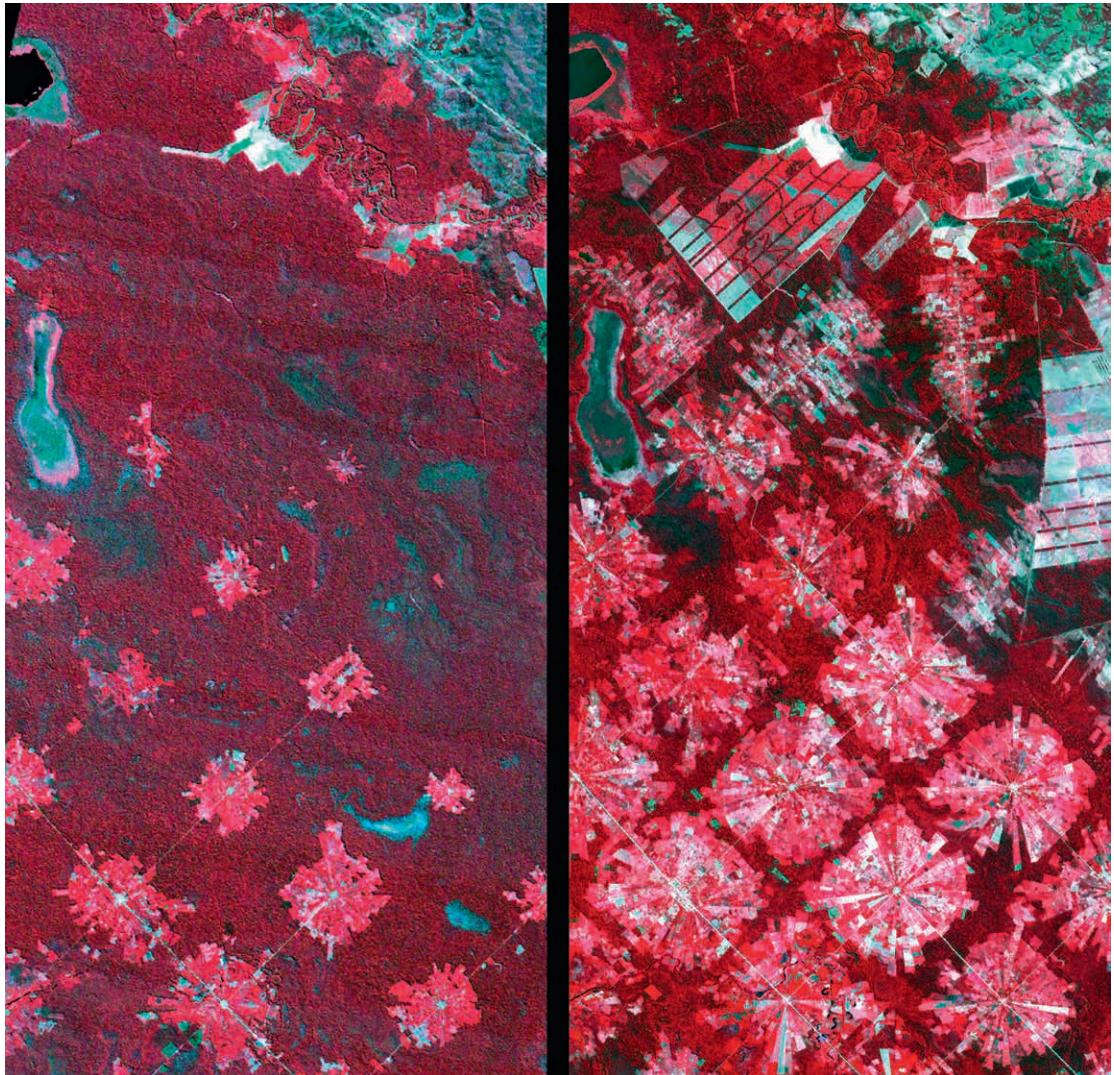


Figure 6.20. Landsat (left) and ASTER (right) false-colour infrared images of the same area of Bolivian forest, in 1986 and 2001. The square areas are villages and soybean plantations. Images downloaded from <http://asterweb.jpl.nasa.gov/gallery-detail.asp?name=bolivia> and reproduced by courtesy of NASA/GSFC/METI/ERSDAC/JAROS and US/Japan ASTER Science Team.

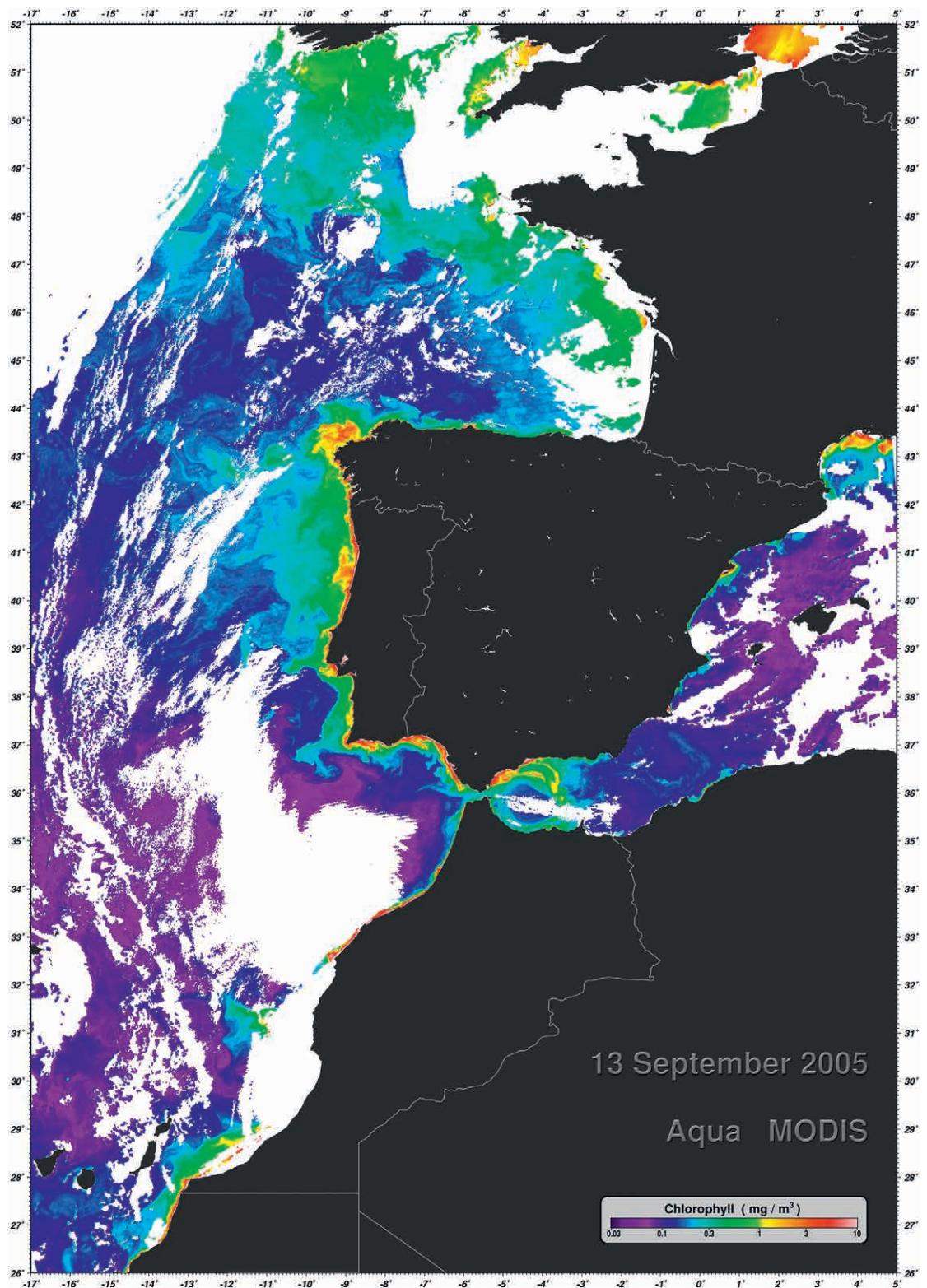


Figure 6.21. Chlorophyll concentrations estimated from MODIS imagery of the eastern North Atlantic. (Image reproduced by courtesy of NASA. Source: http://oceancolor.gsfc.nasa.gov/cgi/image_archive.cgi?c=CHLOROPHYLL.)

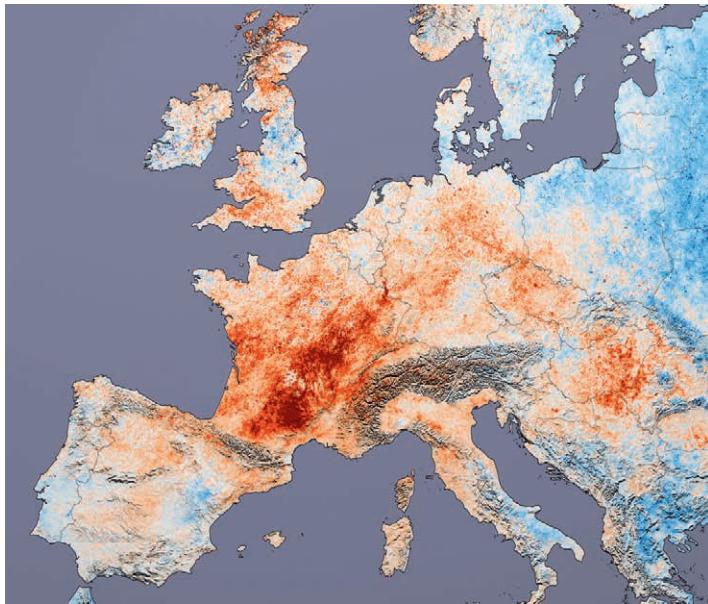


Figure 6.29. Visualisation of the temperature anomaly during the period 20 July 2003 to 20 August 2003, relative to the average for the corresponding periods in 2001, 2002 and 2004. The anomalies were calculated using MODIS TIR imagery. (Image by Reto Stöckli, Robert Simmon and David Herring, NASA Earth Observatory, based on data from the MODIS land team.)

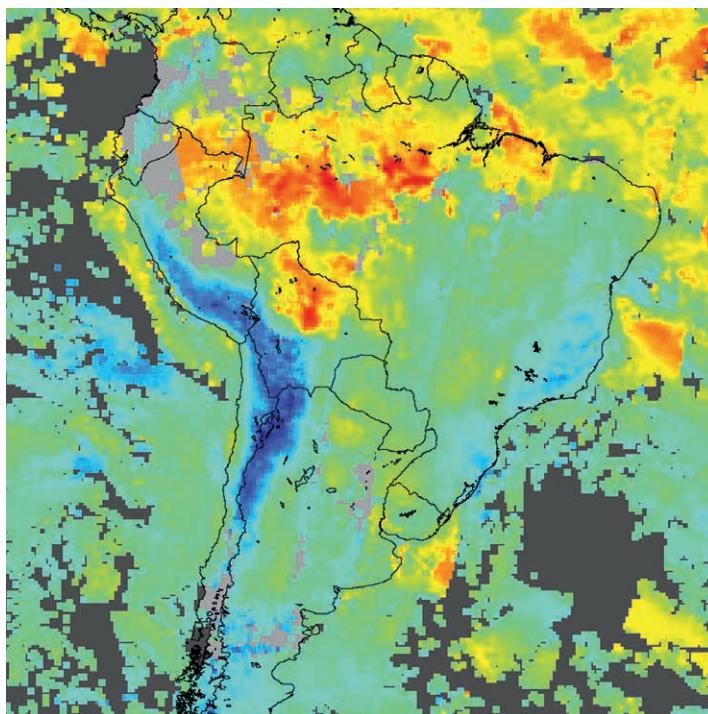


Figure 6.44. Column-integrated total concentration of carbon monoxide over South America, derived from MOPITT data. The image is a one-week gridded composite for the period 19-26 July 2004, and shows higher levels of carbon monoxide (red and yellow colours) in the north of Brazil as a result of fires. (Image downloaded from NASA Visible Earth at http://visibleearth.nasa.gov/view_rec.php?id=19432 and reproduced by courtesy of NASA.)

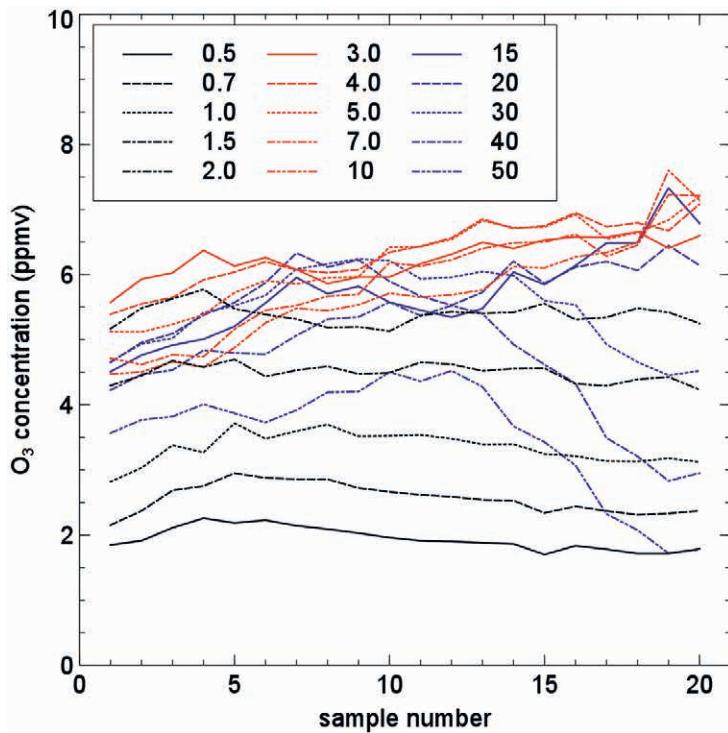


Fig. 6.45. Retrieved ozone concentrations from SBUV/2 data along a transect from 61.27 °N, 83.62 °W (sample 1) to 26.53 °N, 98.04 °W (sample 20), representing about 10 minutes of data. The key shows the atmospheric pressure levels in millibars.

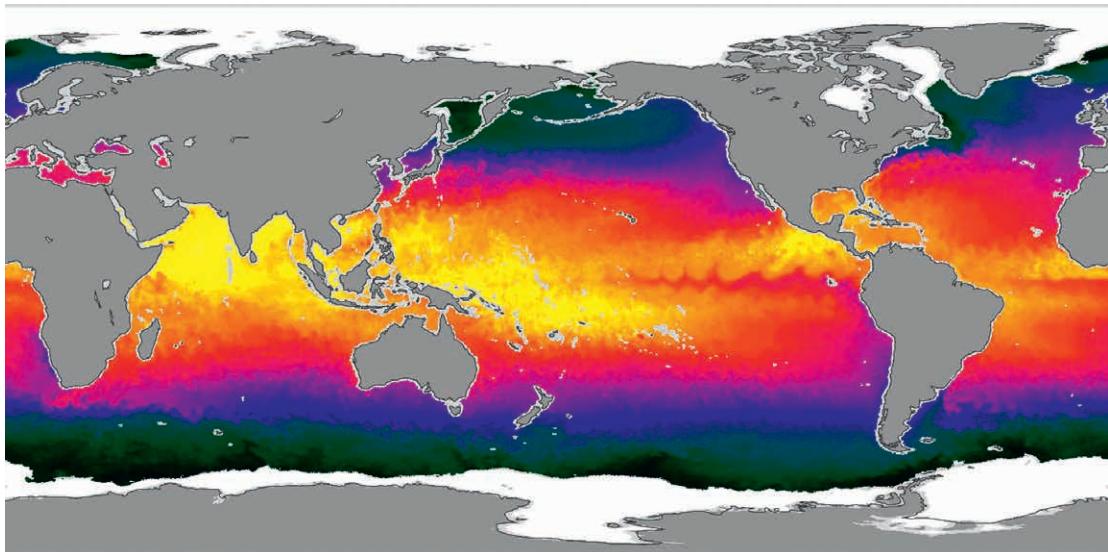


Figure 7.6. Global SST on 1 June 2003, determined using data from the AMSR-E microwave radiometer. The colour scale ranges from -2 °C (dark green) to +35 °C (bright yellow). Areas of sea ice are shown in white. Areas close to land are masked, reflecting the coarse spatial resolution of the data. Image downloaded from <http://aqua.nasa.gov/highlight.php?id=15> and reproduced by courtesy of NASA.

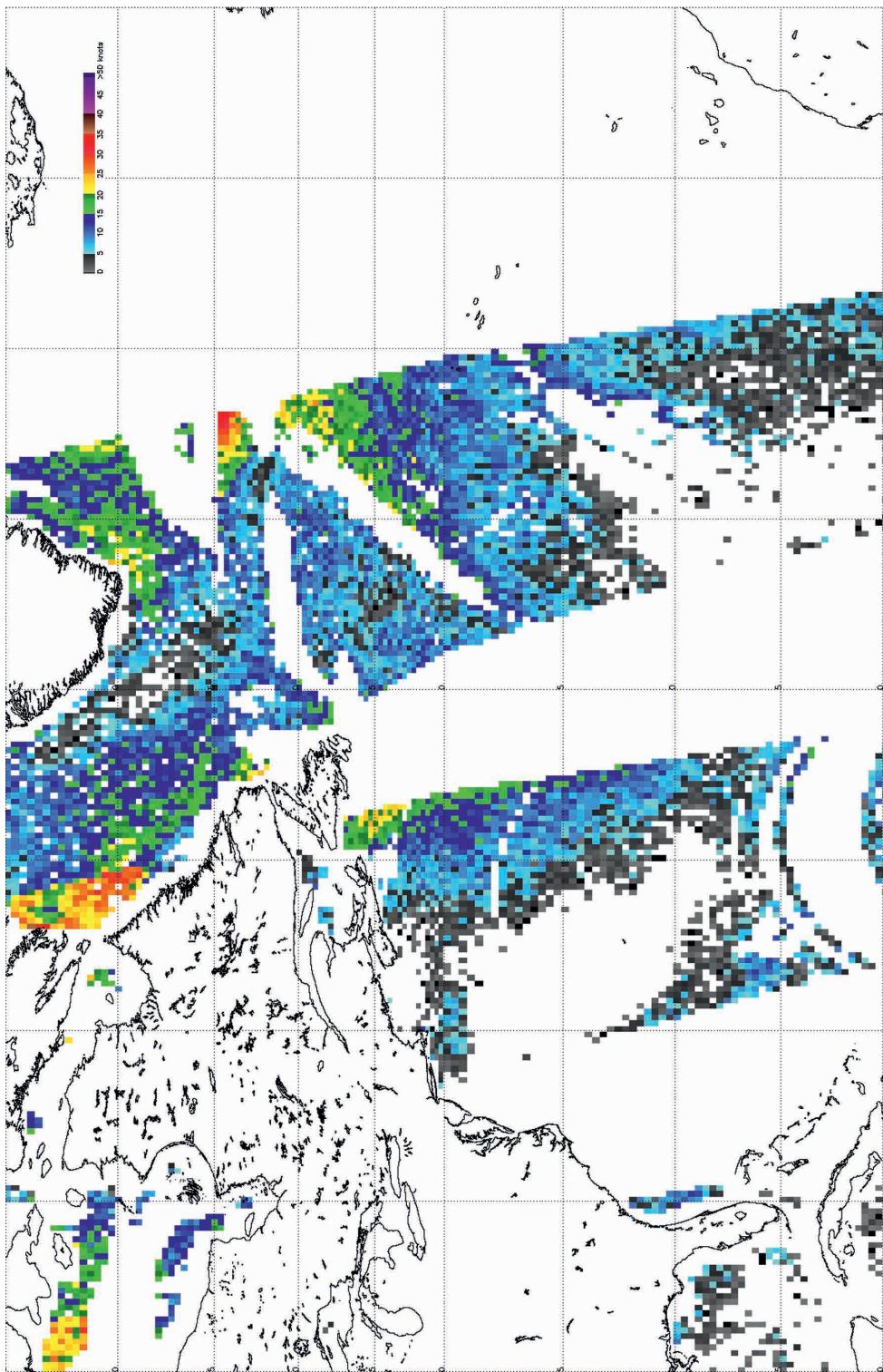


Figure 7.8. Surface wind speeds over the ocean determined from SSM/I passive microwave data. The data were recorded on 12 October 2011. Figure compiled from data supplied by NOAA NESDIS Center for Satellite Applications Research at <http://manati.orbit.nesdis.noaa.gov/datasets/SSMIData.php>.



Figure 7.12. Global distribution of soil moisture on 1 April 2005, deduced from AMSR-E passive microwave data. The retrieval algorithm failed in the areas shown as black. (Njoku, E. N. 2004, updated daily. *AMSR-E/Aqua Daily L3 Surface Soil Moisture, Interpretive Parameters, & QC EASE-Grids V002*, [1 April 2005]. Boulder, Colorado USA: National Snow and Ice Data Center. Digital media.)

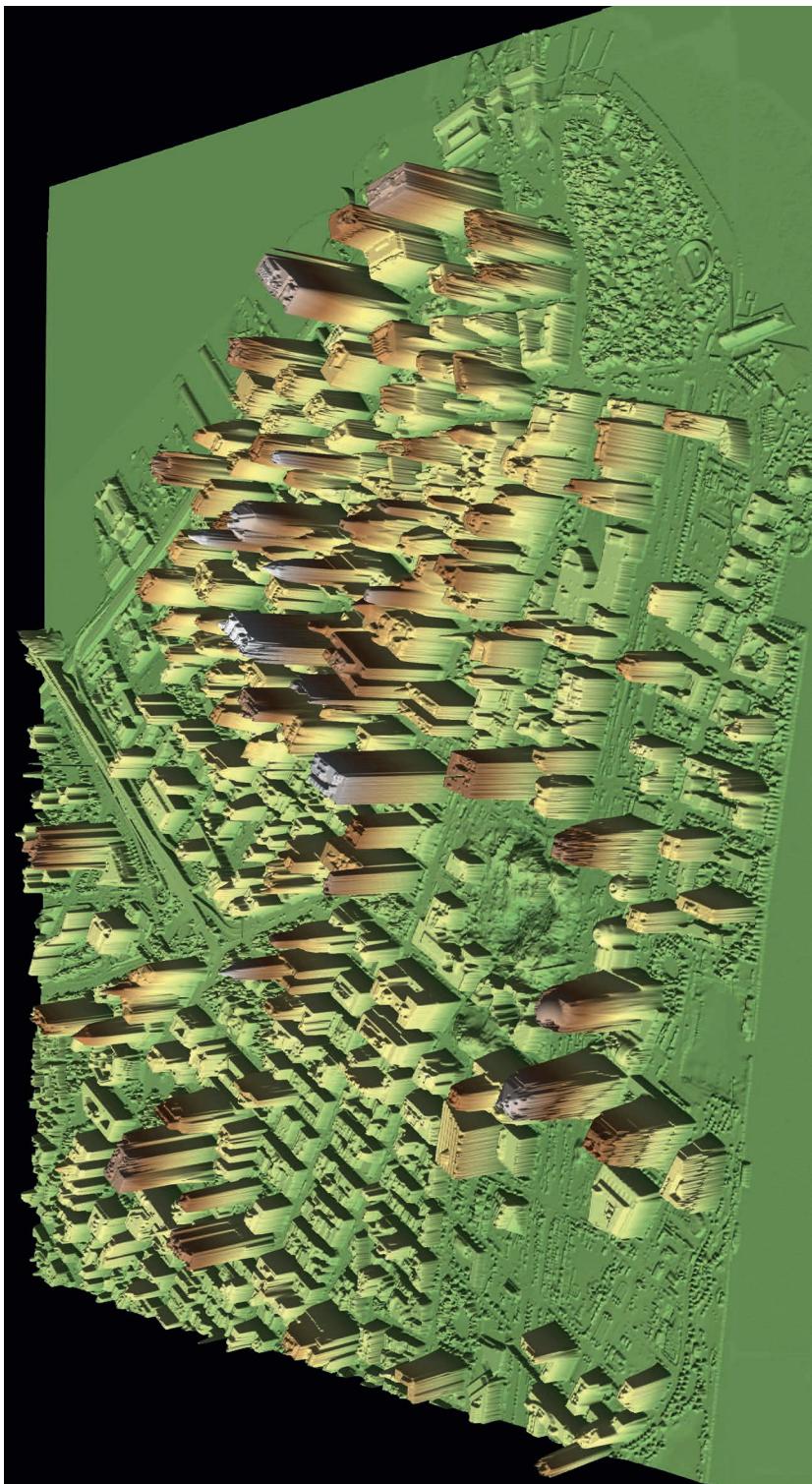


Figure 8.11. Visualisation of airborne LiDAR data collected over Lower Manhattan on 27 September 2001. (The data were acquired by NOAA and processed by the US Army Joint Precision Strike Demonstration. Image downloaded from <http://www.noaanews.noaa.gov/stories/s798.htm> and reproduced by courtesy of NOAA/US Army JPSD.)

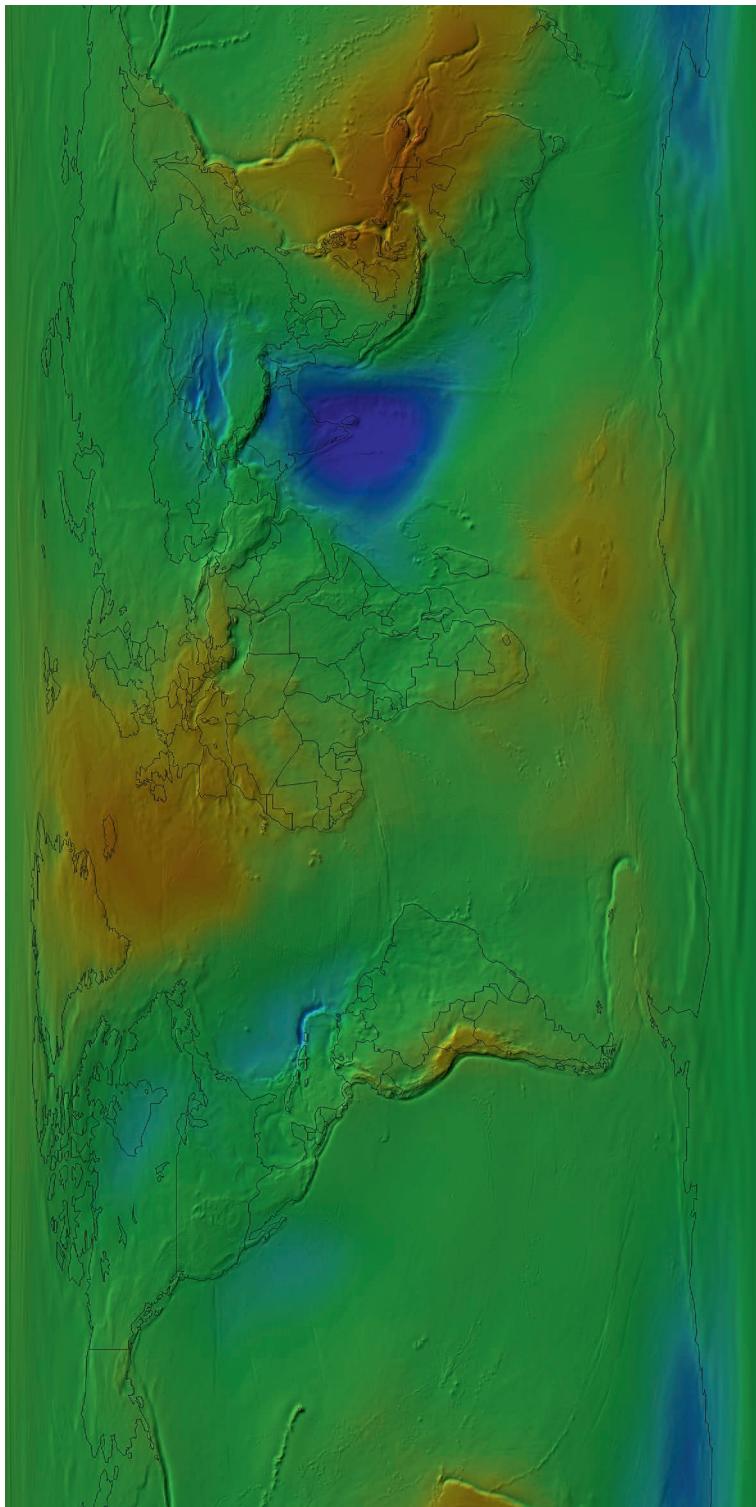


Figure 8.22. Visualisation of the global geoid EGM2008. The colour scale shows the height of the geoid above the WGS84 ellipsoid, ranging from –106 m (blue) to +86 m (orange). (Source: <http://earth-info.nga.mil/GandG/wgs84/gravitymod/egm2008/oceano.html>)

(a)

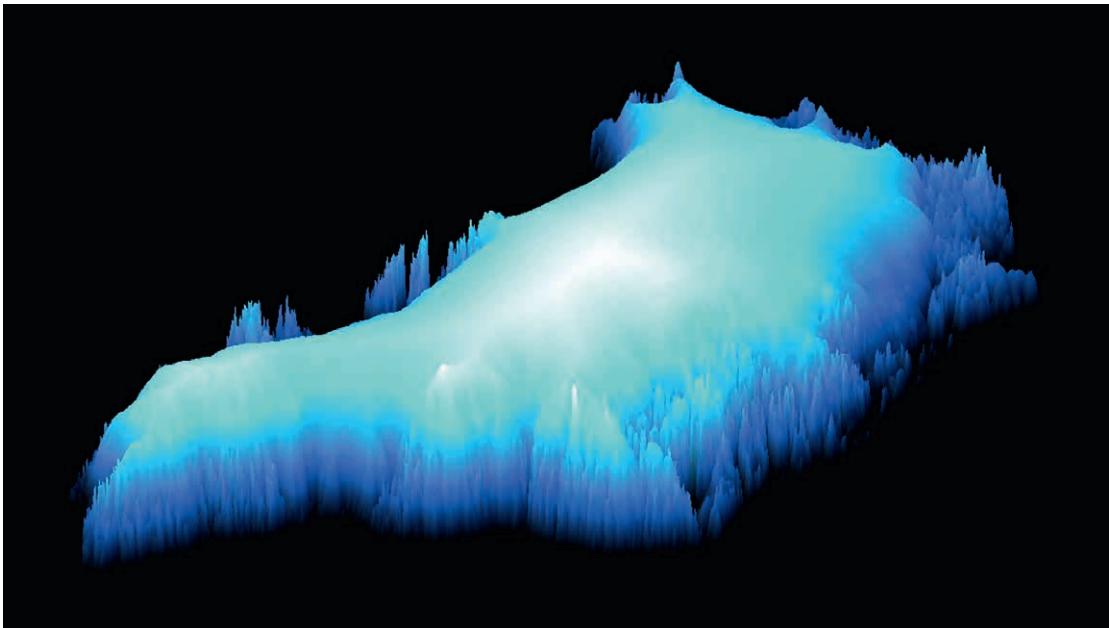


Figure 8.27. (a) Visualisation of the surface topography of Greenland using radar altimeter data.

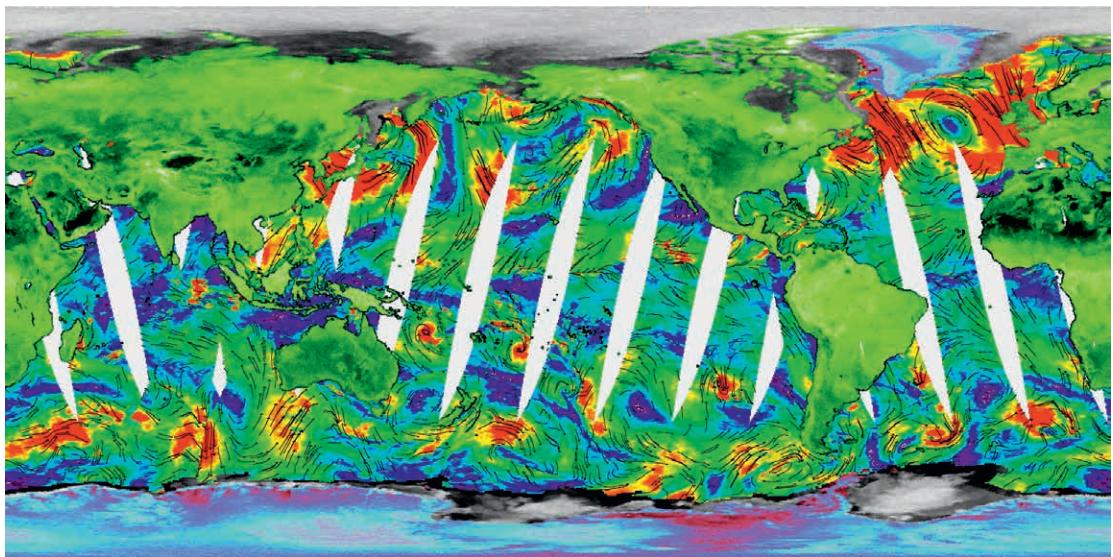
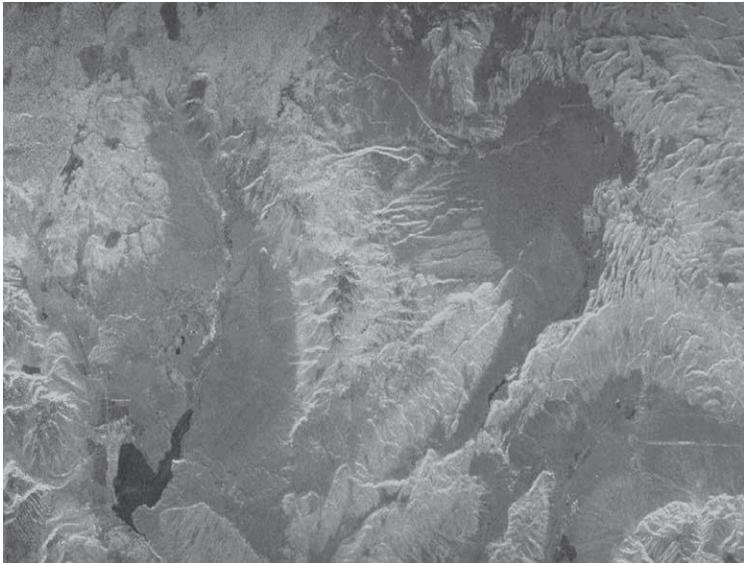


Figure 9.8. Composite image derived from SeaWinds scatterometer data collected by the ADEOS-2 (Midori-2) satellite on 28 January 2003. Variations in backscattering coefficient over land surfaces are represented in shades of green. Over oceans, the colours represent different wind speeds from blue (low) to red (high), and arrows show the inferred circulation of the wind. (Image downloaded from <http://earthobservatory.nasa.gov/IOTD/view.php?id=3248> and reproduced by courtesy of NASA/JPL SeaWinds science team.)

(a)



(b)

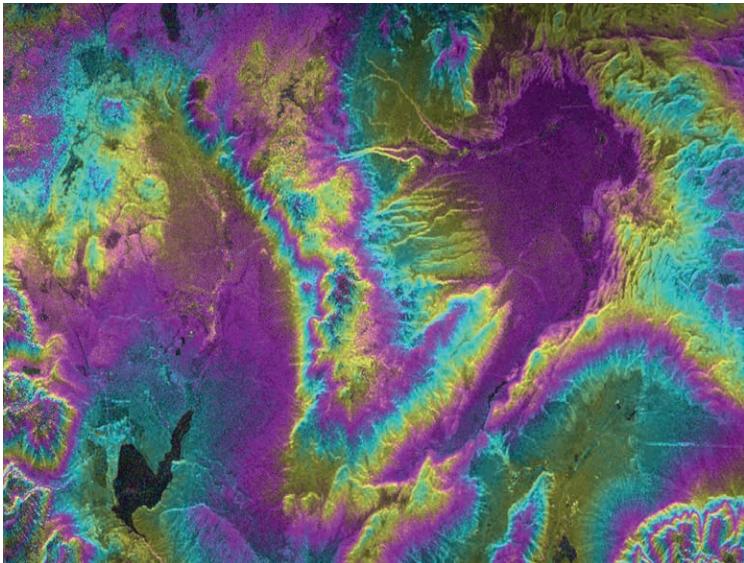


Figure 9.23. (a) One of a pair of SAR images used to construct an interferogram. (b) the corresponding interferogram. The images were acquired by the L-band SIR-C system carried on the Space Shuttle Endeavour, in April and October 1994, with a baseline of around 100 m. They show an area 59×34 km of Long Valley, California. The intensity of the interferogram is derived from the radar images, while the hue represents the phase. (Images downloaded from <http://www.archive.org/details/VE-IMG-626> and reproduced by courtesy of NASA Jet Propulsion Laboratory.)

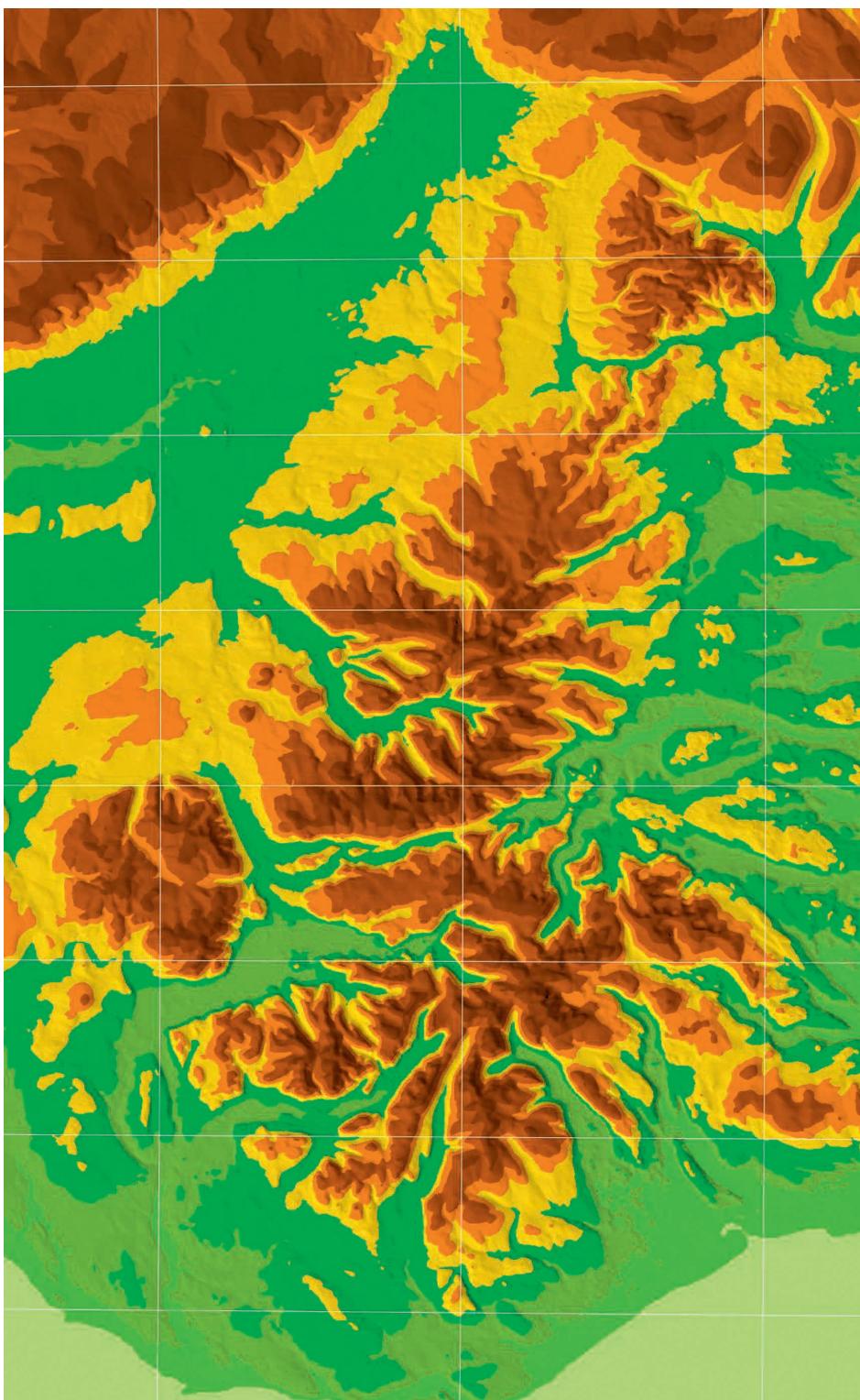


Figure 9.26. Visualisation of SRTM data of north-western England. The graticule has intervals of 10 minutes of latitude and longitude.

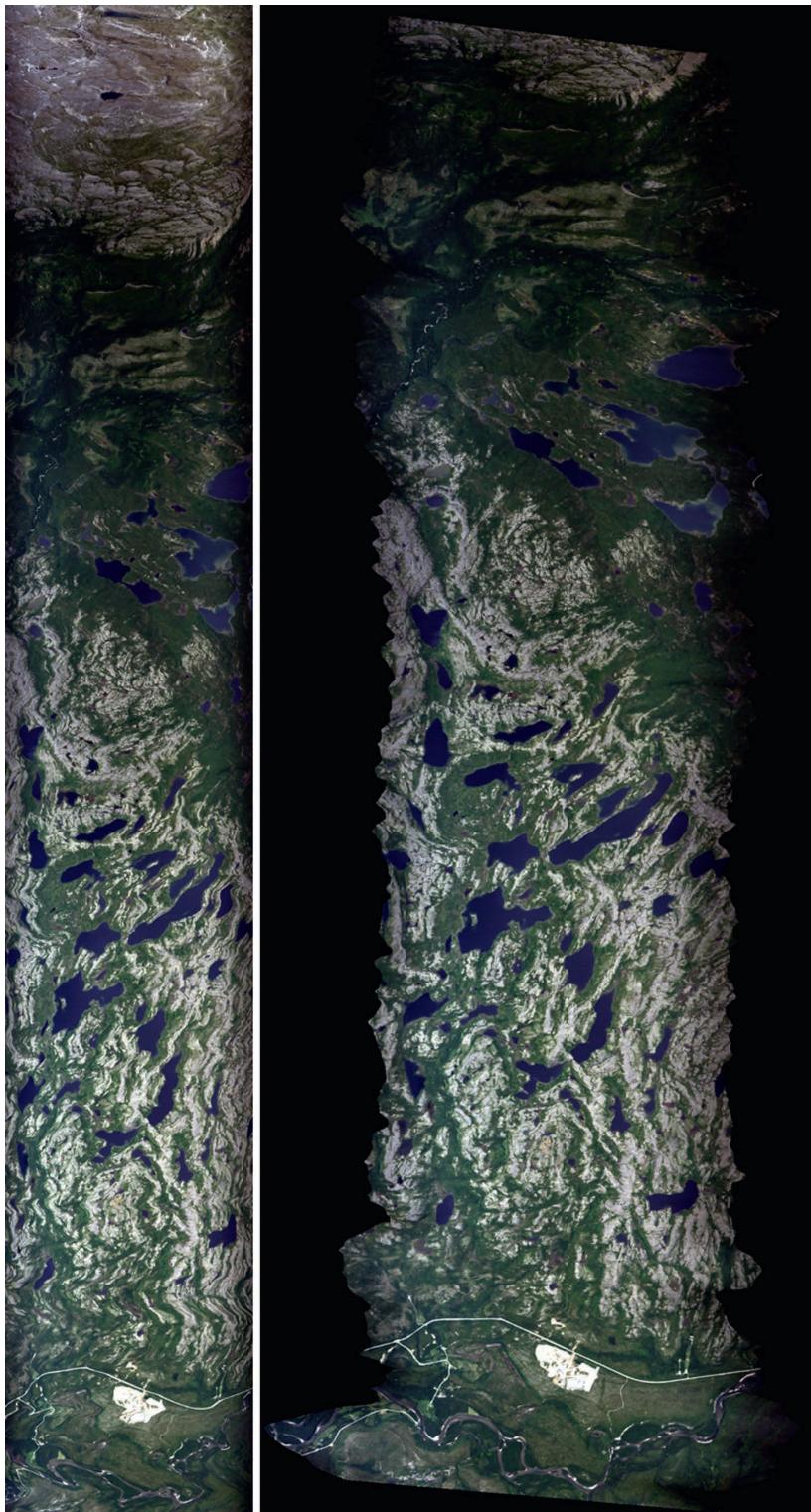


Figure 10.3. True-colour representation of airborne scanner imagery before (above) and after (below) correction for the effects of roll, pitch and yaw. The aircraft was flying from left to right, as can be seen from the fact that the initial roll was subsequently corrected by the pilot. (Airborne Thematic Mapper image of Porsangmoen, Norway, collected by UK Natural Environment Research Council's Airborne Research and Survey Facility in 2004.)



Figure 10.5. Launch of a Dnepr rocket from Baikonur Cosmodrome, Kazakhstan. The Dnepr is a converted intercontinental ballistic missile and it is launched from an underground silo. It has been used to launch more than a dozen remote sensing satellites. (Photograph from <http://www.kosmotras.ru> by courtesy of ISC Kosmotras.)

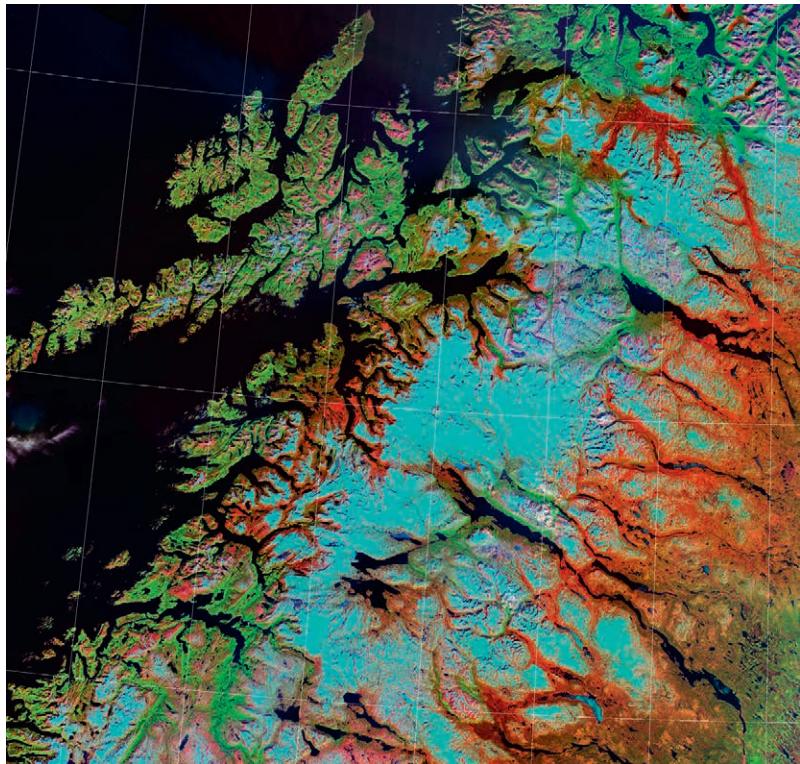
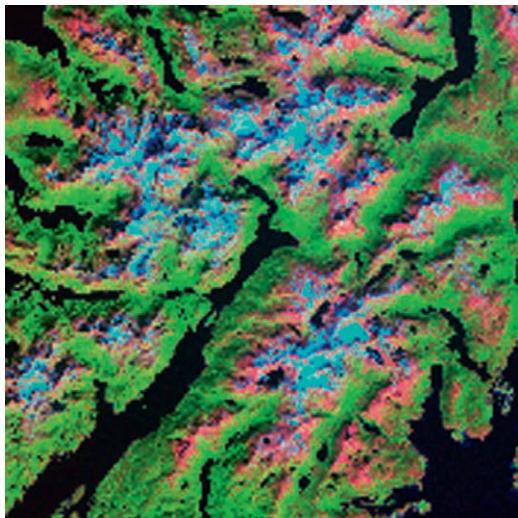
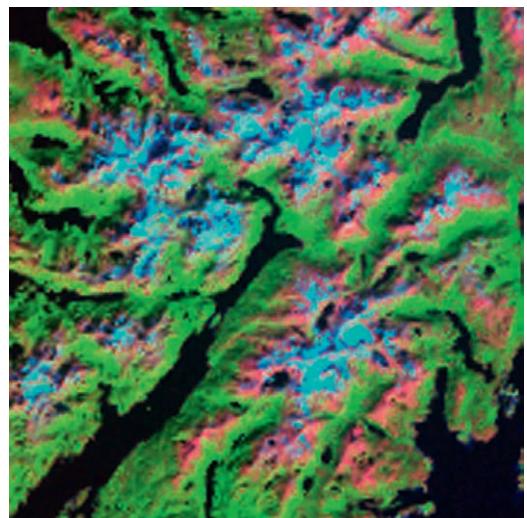


Figure 11.6. A georeferenced satellite image. The image is an extract of a mosaic of Landsat TM images showing parts of northern Norway and Sweden. It covers an area of approximately 57×60 km. The latitude-longitude grid has been superimposed after georeferencing the image.

(a)



(b)



(c)

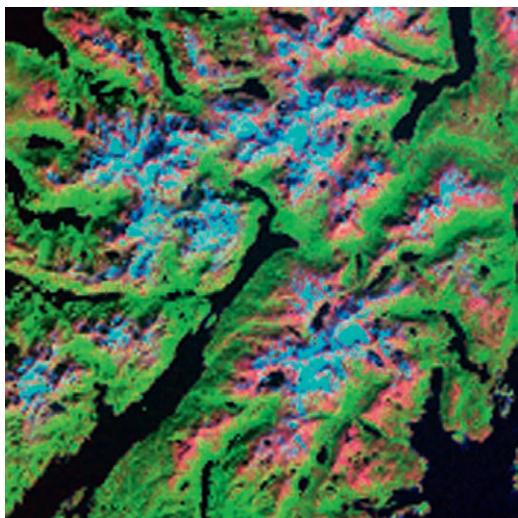


Figure 11.7. Extract of the image in Figure 11.6 reprojected to a geographical projection using (a) nearest-neighbour, (b) bilinear, and (c) bicubic resampling.

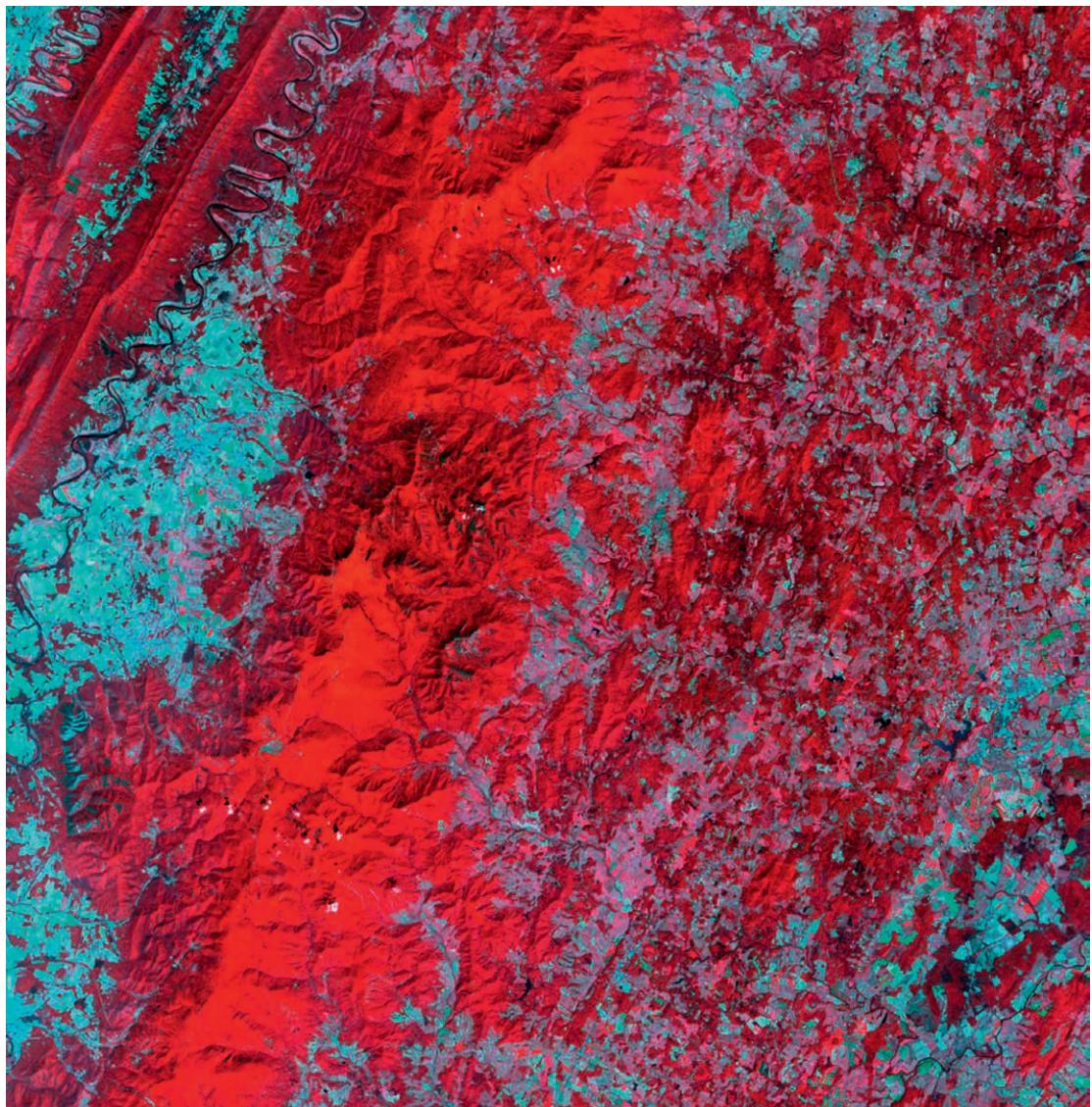


Figure 11.26. A 321 composite ASTER image showing the northern part of the state of Virginia, USA. The sinuous feature at top left is the Shenandoah River.

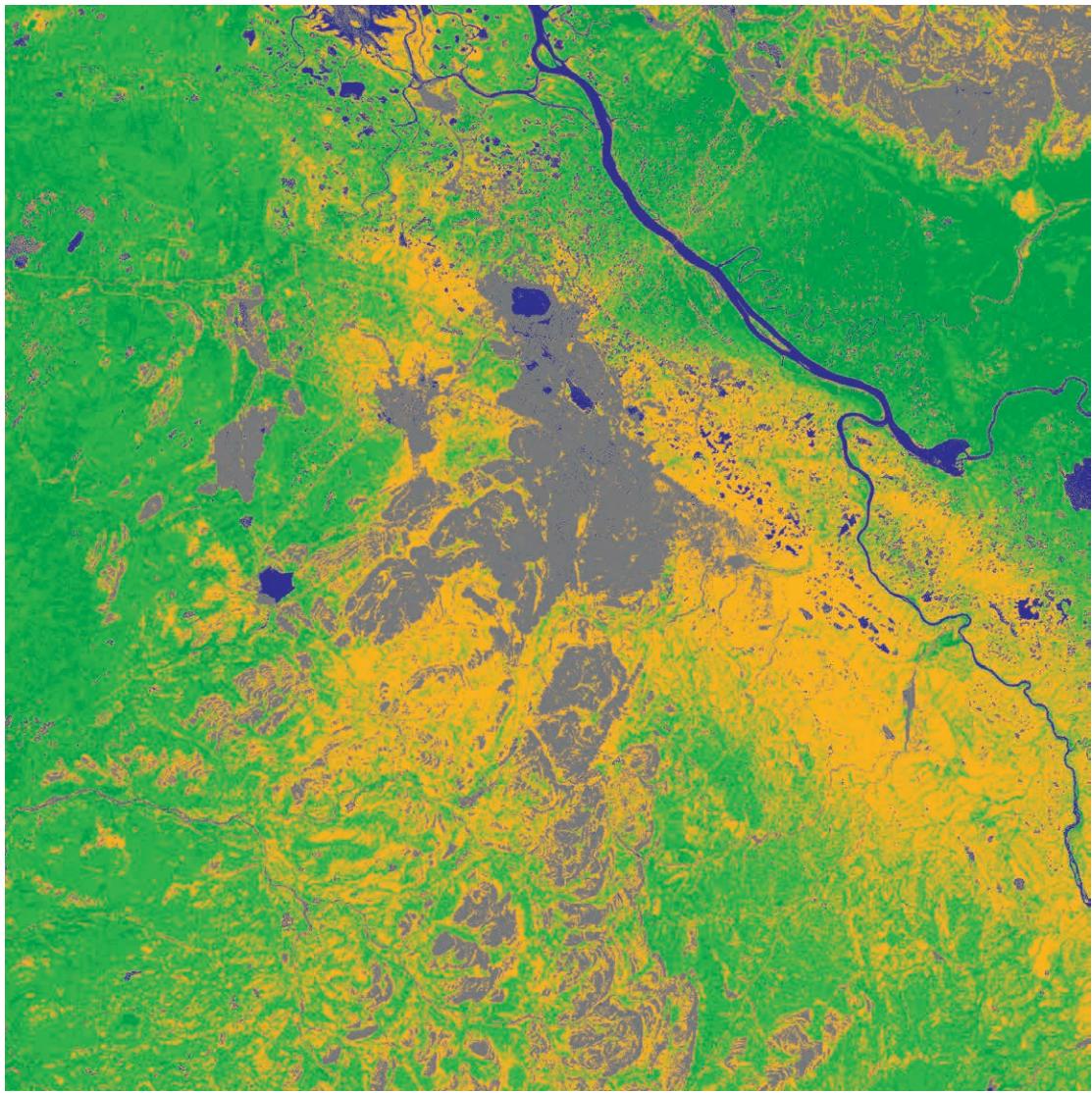


Figure 11.29. Pseudocolour representation of the NDVI image of Figure 11.23c. Blue denotes values NDVI below 0, grey between 0 and 0.25, orange between 0.25 and 0.5, light green between 0.5 and 0.7 and dark green above 0.7. The choice of colours suggests the distribution of water (blue) and pollution-damaged vegetation (orange).

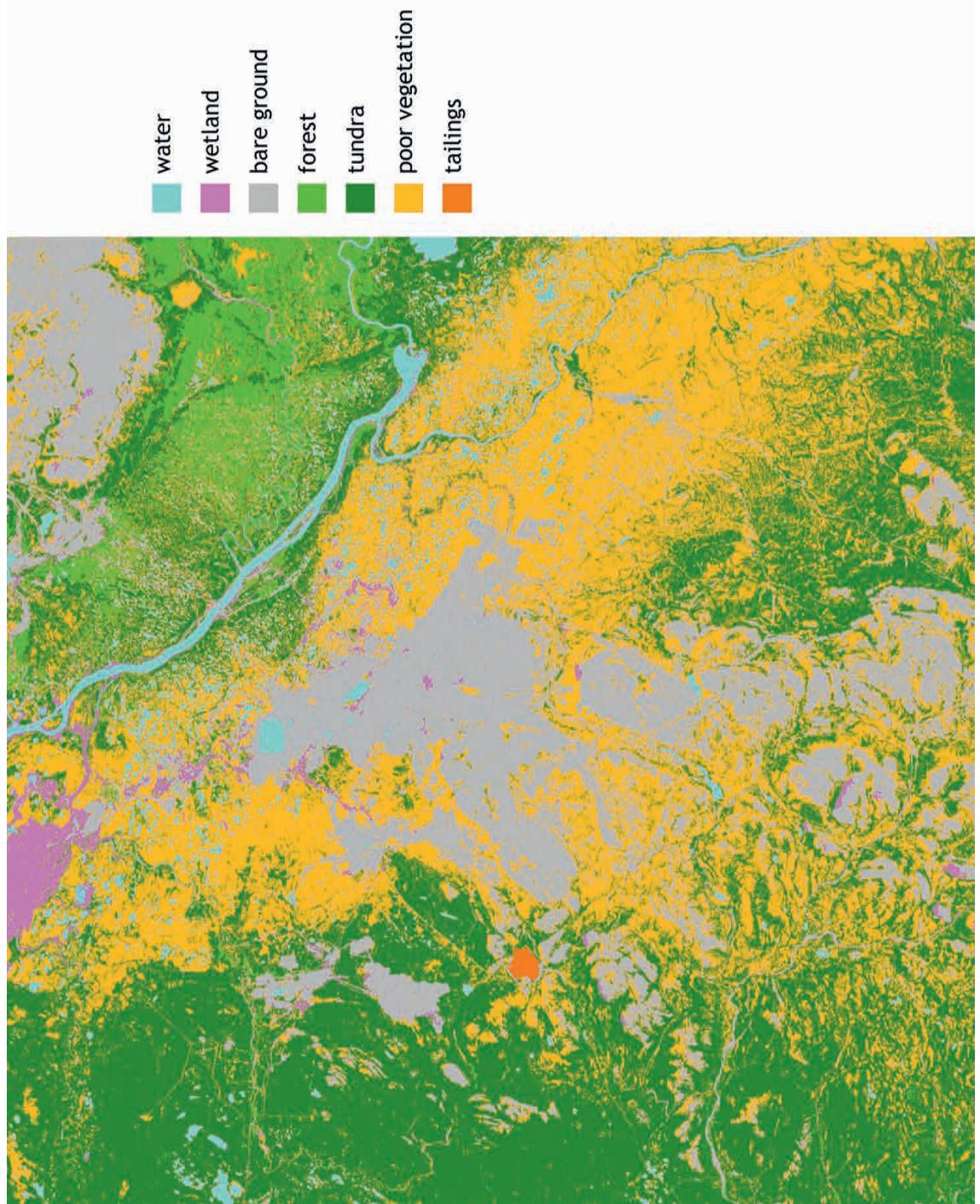


Figure 11.31. Supervised classification of the image of Figure 11.23c.

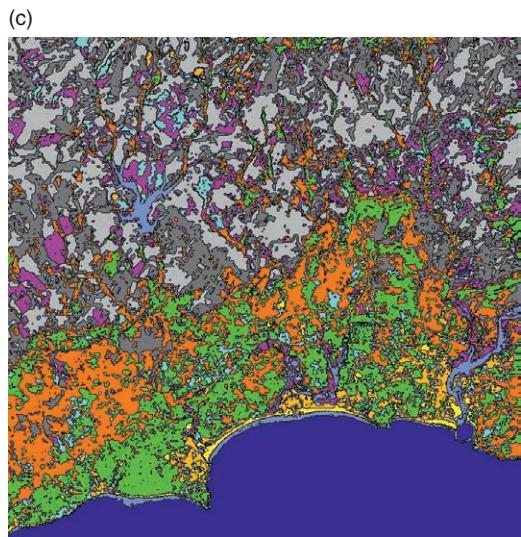
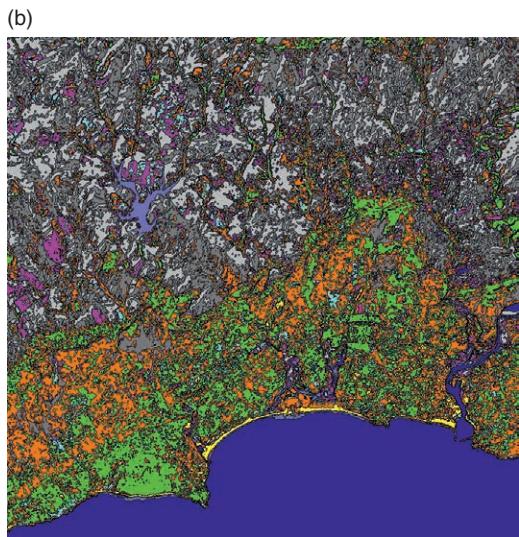
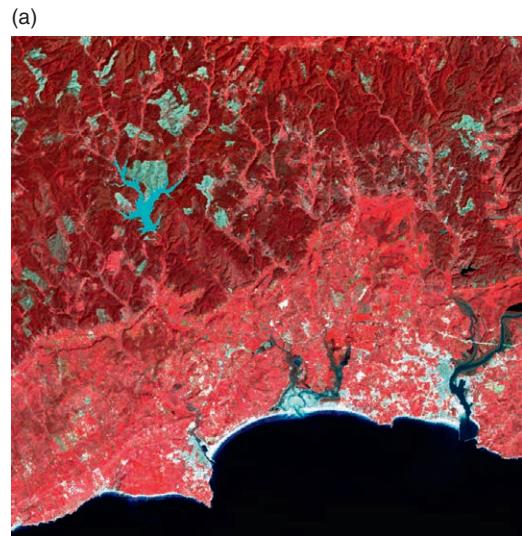
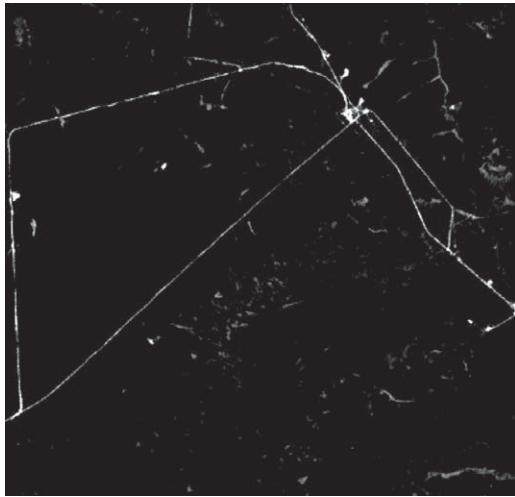
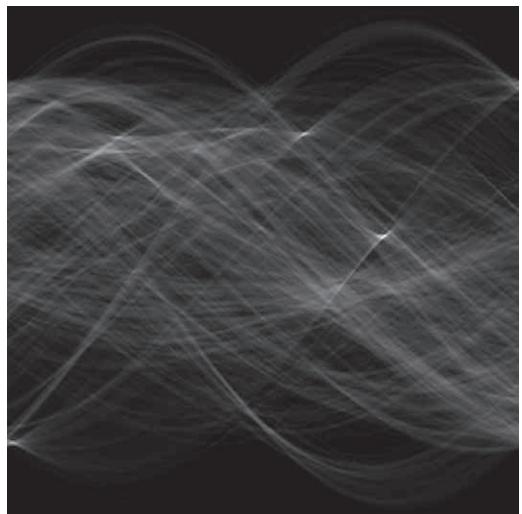


Figure 11.39 (a) 8.6 km square extract of a FCIR Landsat image centred on the Barragém de Bravura, Portugal. (b) Segmentation into ten classes using minimum-distance multispectral classification. (c) Segmentation into ten classes using ECHO.

(a)



(b)



(b)

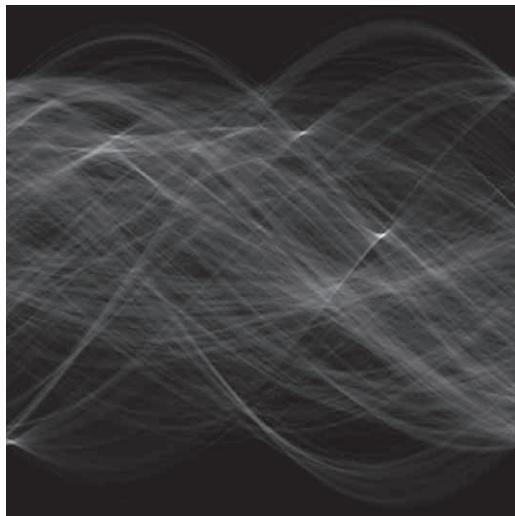


Figure 11.44. (a) Original image (extract of a SPOT image showing oil pipelines in the Komi Republic, Russia). (b) Accumulator array. (c) Original image with the three most prominent detected straight lines superimposed on it. (Hough transform implemented using a Java plugin (Burger and Burge 2005) for ImageJ.)