

# **Location-Based Services and Geo-Information Engineering**

*Allan Brimicombe*  
*Chao Li*

**WILEY-BLACKWELL**

# **Location-Based Services and Geo-Information Engineering**

**Allan Brimicombe**

*University of East London, UK*

**Chao Li**

*University College London, UK*



**WILEY-BLACKWELL**

A John Wiley & Sons, Ltd., Publication



# **Location-Based Services and Geo-Information Engineering**

## **Mastering GIS: Technology, Applications and Management Series**

*GIS and Crime Mapping*

Spencer Chainey and Jerry Ratcliffe

*GIS Mastering the Legal Issues*

George Cho

*Geodemographics, GIS and Neighborhood Targeting*

Richard Harris, Peter Sleight and Richard Webber

*Integration of GIS and Remote Sensing*

Victor Mesev

*Landscape Visualization: GIS Techniques for Planning and Environmental Management*

Andrew Lovett, Katy Appleton and Simon Jude (forthcoming)

*GIS for Public Sector Spatial Planning*

Scott Orford, Andrea Frank and Sean White (forthcoming)

*GIS and Techniques for Habitat Management*

Nigel Waters and Shelly Alexander (forthcoming)

# Location-Based Services and Geo-Information Engineering

**Allan Brimicombe**

*University of East London, UK*

**Chao Li**

*University College London, UK*



**WILEY-BLACKWELL**

A John Wiley & Sons, Ltd., Publication

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

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

*Registered office*

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

*Other Editorial Offices*

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

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

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

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

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

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

***Library of Congress Cataloguing-in-Publication Data***

Brimicombe, Allan.

Location-based services and geo-information engineering / Allan Brimicombe, Chao Li.  
p. cm.

Includes bibliographical references and index.

ISBN 978-0-470-85736-6 — ISBN 978-0-470-85737-3

1. Location-based services. 2. Geographic information systems.

I. Li, Chao, 1962 June 11— II. Title.

TK5105.65.B75 2009

910.285—dc22

2008040264

ISBN: 978-0-470-85736-6 (HB)

ISBN: 978-0-470-85737-3 (PB)

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

Set in 11/13 Times by Integra Software Services Pvt. Ltd., Pondicherry, India.  
Printed in Singapore by Fabulous Printers Pte Ltd.

First Impression 2009

# Contents

*Mastering GIS Series: Foreword* xi

*Preface* xiii

<b>1</b>	<b>The Context of Location-Based Services</b>	<b>1</b>
1.1	Introduction	1
1.2	The Information Society	2
1.3	The Digital City	9
1.4	The New Mobility	15
1.5	Outline of Following Chapters	20
<b>2</b>	<b>Technological Convergence: Towards Location-Based Services</b>	<b>21</b>
2.1	Introduction	21
2.2	The Internet and World Wide Web	22
2.2.1	The Basics of the Internet	22
2.2.2	World Wide Web	24
2.3	New Information and Communication Technologies	31
2.3.1	Mobile Wireless Telecommunication Technologies	32
2.3.2	Other Wireless Networks	47
2.3.3	Wireless Mobile Devices	52
2.4	Geographical Information Systems	56
2.5	Convergence of Technologies	57
2.5.1	Convergence of Wireless Technologies	57
2.5.2	Wireless Mobile Internet	58
2.5.3	Internet GIS and Wireless GIS	62
2.5.4	Towards LBS	64



## Contents

<b>3</b>	<b>GIS and Geo-Information Engineering</b>	<b>67</b>
3.1	Introduction	67
3.2	Where is . . . ? How Do I Get there?	68
3.3	Defining GIS	73
3.3.1	An Historical Perspective	74
3.3.2	On-Going Development Trends	76
3.4	Exploring GIS Software	81
3.4.1	Measurement and Scale	82
3.4.2	Map Projection	85
3.4.3	Symbology	87
3.4.4	Data Primitives and Data Layers	89
3.4.5	Feature Attributes	91
3.4.6	Creating Thematic Maps	94
3.4.7	Scaling the Applications	96
3.5	Issues of GIScience	98
3.6	GI Engineering: the Rise of Ubiquitous GIS?	103
<b>4</b>	<b>Location-Based Services</b>	<b>109</b>
4.1	Introduction	109
4.2	Are Location-Based Services New?	110
4.3	From Locating Services to Location-Based Services	114
4.3.1	Using the Web for Services	114
4.3.2	Using Navigation Systems for Services	118
4.3.3	Using Mobile Phones for Services	119
4.3.4	Location-Based Services	121
4.4	E911 and E112 Mandates	123
4.4.1	E911	124
4.4.2	E112	125
4.5	Keitai	126
4.6	LBS Architecture	128
4.7	Application Areas	132
4.8	Implications of LBS for GIScience	133
<b>5</b>	<b>Data for Location-Based Services</b>	<b>137</b>
5.1	Introduction	137
5.2	The Size and Granularity of the Problem	139
5.3	Data Collection Technologies	144
5.3.1	GPS and Inertial Navigation Systems	145
5.3.2	Remote Sensing	146

5.3.3	Ground Survey	148
5.3.4	Non-Traditional Approaches to Data Collection	150
5.3.5	Update Frequencies	152
5.4	Data Quality Issues	155
5.4.1	Understanding Error and Uncertainty	156
5.4.2	Assessing Data Accuracy	159
5.4.3	Metadata	161
5.5	Organizing and Accessing Data	166
<b>6</b>	<b>Locating the User</b>	<b>169</b>
6.1	Introduction	169
6.2	Positioning Technologies	171
6.3	Global Positioning System	173
6.3.1	Basic Principles of GPS	174
6.3.2	A More Detailed Look at GPS	178
6.3.3	Principle of Differential GPS	182
6.3.4	Indoor GPS	185
6.3.5	Other Satellite Systems	186
6.3.6	An Example of GPS Use	186
6.4	Network-Based Positioning Technologies	189
6.4.1	Network Cell Identification	189
6.4.2	Angle of Arrival	192
6.4.3	Time Delay Methods	193
6.4.4	Advanced Forward Link Trilateration and Enhanced Forward Link Trilateration	197
6.5	Short Range Positioning Technologies	197
6.5.1	WiFi-Based Positioning	198
6.5.2	Bluetooth Technology Used for Positioning	200
6.5.3	Radio Frequency Identification	201
6.5.4	Other Non-Radio Signal Technologies Used for Short Range Positioning	203
6.6	Hybrid Positioning Approaches	205
<b>7</b>	<b>Context in Location-Based Services</b>	<b>209</b>
7.1	Introduction	209
7.2	Context and Context-Awareness	211
7.3	Context in LBS	214

## Contents

7.4	Environment as Context	217
7.5	Technology as Context	221
7.6	User as Context	225
7.7	Dynamics of Context	228
<b>8</b>	<b>The Spatial Query</b>	<b>235</b>
8.1	Introduction	235
8.2	Geometric Data	237
8.2.1	Vector	237
8.2.2	Raster	240
8.2.3	Object-Oriented	242
8.3	Topological Data	245
8.4	Attribute Data	250
8.5	Indexing Spatial Databases	250
8.6	Issues of Data Temporality	253
8.6.1	Space Versus Time Dominant Views	254
8.6.2	Space–Time Paths and Topology	257
8.6.3	Database Implications for LBS	260
8.7	Spatial Queries	262
8.7.1	Relational Algebra	263
8.7.2	SQL and Extended SQL	265
8.7.3	Querying Graphs	267
8.7.4	Query Optimization	274
<b>9</b>	<b>Communication in Location-Based Services</b>	<b>279</b>
9.1	Introduction	279
9.2	Modes of Communication in LBS	280
9.3	Maps in LBS	288
9.3.1	Making Maps for Information Communication	289
9.3.2	Digital Maps Used On-Screen	290
9.3.3	Map Generalization	293
9.4	Issues Around Modes of Communication in LBS	297
9.5	Learning from Spatial Information	300
9.5.1	Acquiring Spatial Knowledge	301
9.5.2	A Study of User Preference for Different Modes of Communication	304
9.6	Multimodal and Context-Aware Modes of Communication	309

<b>10 The Business of Location-Based Services</b>	<b>313</b>
10.1 Introduction	313
10.2 Emerging Sectors	314
10.2.1 Internet-Based Business Models	314
10.2.2 Implications of Using Mobile Devices	318
10.2.3 Mobile Device-Based Business Models	319
10.2.4 Adoption and Hype Curves	321
10.3 Emerging Products	323
10.4 Standardization Issues	326
10.5 Legal Issues	329
10.5.1 Patents	330
10.5.2 Copyright	332
10.5.3 Liability	334
10.6 Social Issues	337
10.7 Conclusions	344
 <i>Acronyms</i>	 345
 <i>References</i>	 350
 <i>Index</i>	 367



# Mastering GIS Series: Foreword

Since 2001 it has been my privilege to be involved with the John Wiley and Sons book *Geographic Information Systems and Science*. Through its various editions, this book and associated materials has sought to present a state-of-the-art overview of the principles, techniques, analysis methods and management issues that come into play whenever the fundamental question ‘where?’ underpins decision making.

Together this material makes up the organizing concepts of Geographic Information Systems (GIS), which has a rich and varied history in environmental, social, historical and physical sciences. We can think of GIS as the lingua franca that builds upon the common purposes of different academic traditions, but with an additional unique emphasis upon practical problem solving. As such, much of the core of GIS can be thought of as transcending traditional academic disciplinary boundaries, as well as developing common approaches to problem solving amongst practising professionals.

Yet many of the distinctive characteristics, requirements and practices of different applications domains also warrant specialized and detailed treatments. ‘Mastering GIS’ seeks to develop and extend our core understanding of these more specialized issues, in the quest to develop ever more successful applications. Its approach is to develop detailed treatments of the requirements, data sources, analysis methods and management issues that characterize many of the most significant GIS domains.

First and foremost, this series is dedicated to the needs of advanced students of GIS and professionals seeking practical knowledge of niche applications. As such, it is dedicated to making GIS more

## **Foreword**

efficient, effective and safe to use, and to render GIS applications ever more sensitive to the geographic, institutional and societal contexts in which it is applied.

Paul Longley, Series Editor  
Professor of Geographic Information Science  
University of London

# Preface

Location-based services (LBS) are the delivery of data and information services where the content of those services is tailored to the current or some projected location and context of a mobile user. This is a new and fast-growing technology sector incorporating Geographical Information Systems (GIS), wireless technologies, positioning systems and mobile human–computer interaction. Some view LBS as the ‘killer app’ that will propel GIS into the mainstream of quotidian use throughout society. Geo-Information (GI) Engineering is an extension of GIScience into the design of dependably engineered solutions to society’s use of geographical information and underpins applications such as LBS. Making all this possible is the tremendous technological convergence and growing systems interoperability that have taken place over the last decade or so. Most of us now own a mobile phone (if not two or three!) that has more computing power and functionality than a 1980s PC, a digital camera with more mega-pixels than could be dreamed possible in the 1990s, and a high resolution LCD screen and sufficient bandwidth to make it a Web browser, games console and TV all in one! That such mobile devices are a catalyst for LBS is not surprising.

We begin this book by setting the broader social and economic contexts for location-based services, and follow this with an in-depth analysis of a number of technologies that have developed and converged so as to make LBS technically possible. Prime amongst these will be the telecommunications networks servicing mobile devices. We then provide an overview of GIS as a key technology in handling location-based spatial queries and, therefore, at the heart of LBS. However, given that it is now nearly two decades since GIScience emerged from the maturing GIS technologies and their applications, we take a further evolutionary step to consider LBS as an early manifestation of GI Engineering. We then focus on LBS and their architecture with chapters on specific aspects of operational LBS: data, locating the user, using context, spatial queries and



## Preface

communication. Each will consider technical and implementation issues. We finish the book by discussing a number of issues to do with the business of LBS, including a range of business models that providers might adopt as well as some of the ethical, legal and privacy issues that arise from the use of LBS.

Writing this book has been a longer journey than we first expected. LBS are new, heterogeneous technologies and the relevant research is spread across a number of separate disciplines. This research needed collating and synthesizing, the outcome being a definitive state of the art review of current LBS. We ourselves were conducting research that has contributed to this book. All this has meant that we have been able to articulate a new research agenda for LBS. We would of course like to thank the many people with whom we have had discussions on wireless technologies, GIScience and LBS over the last few years. We would also like to sincerely thank each other for the effort, patience and fortitude necessary to write this book. We hope you enjoy it.

Allan Brimicombe  
Chao Li

# Chapter 1

## The Context of Location-Based Services

### 1.1 Introduction

---

This is a book about a new and fast growing technology sector delivering electronic services based primarily on location and geography. This sector is generally referred to as *location-based services* or by its acronym LBS. Because location and geography are at the heart of LBS, this book also has a focus towards *geographical information systems* (GIS), which are at the heart of integrating, managing, querying and visualizing geographical data sets. Whilst some proponents may view LBS narrowly as an application of matured GIS, and whilst we would agree that GIS is one of the underpinning technologies of LBS, they are certainly not synonymous. In fact, we view them as being rather separate technologies: GIS having developed first since the 1960s to occupy one niche and LBS having emerged only recently to fill another. Moreover, LBS have only been made possible due to the maturation and convergence of a whole raft of technologies, such as mobile phones, the Internet, the World Wide Web (the Web) and global positioning system (GPS), which had not been developed at the time when GIS were first developed. Because LBS are underpinned by so many different technologies, in the early chapters of this book both the context and the technological convergence that have made LBS possible are traced. Ordinarily this ground would be covered first before coming to a definition of LBS, such that readers would

understand where it's coming from and where it's leading to. But this approach might also frustrate readers, who would not then have a clear definition of LBS until some considerable way through the book. So we will provide our definition here, out of context as it is, and work towards a firmer understanding of its meaning and subtleties in what follows. Thus:

***Location-based services (LBS) are the delivery of data and information services where the content of those services is tailored to the current or some projected location and context of a mobile user.***

How has society changed in the last quarter of the twentieth century to bring about demand for such services? What trajectory are we on that makes it likely that demand for such services will grow dramatically in the coming decade? This is discussed in the remainder of this chapter under three headings – the ‘information society’, the ‘digital city’ and the ‘new mobility’ – and in doing so will provide the context for looking, in Chapter 2, at the various strands of the technological convergence that are making LBS possible. It could be argued that the information society, the digital city and our increased mobility are just three facets of the same phenomenon: the digitalization of economic and social activity. However, as will be pointed out, cities are where the main economic and social driving forces of an information society reside and if the information society is in large an urban phenomenon, then this particular dimension needs to be explored. Our increased mobility is also pertinent in a number of ways. Not only does this relate to greater social and geographical mobility, but it also relates to informational and intellectual mobility. Flexibility to move across collaborative networks in a more holistic, cross-disciplinary way will be key to innovation and creativity, and hence wealth creation, in an information society. This multi-faceted mobility is a dominant driver for LBS and therefore merits separate discussion.

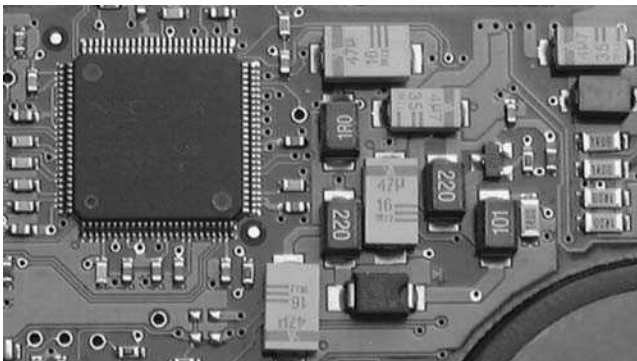
## 1.2 The Information Society

---

The majority of readers would agree that the age in which we now live is as different from the industrializing era of the Victorians as the medieval agricultural society must have seemed different to them. The period from the mid-eighteenth to the early years of the twentieth century is now recognized as being the Industrial Revolution. A radical

and complete transformation of society from an agrarian economy (First Wave) to a modern, manufacturing-based economy (Second Wave) was made possible by the harnessing of power to drive machines coupled with the organization and division of labour (alongside other resources) into units for mass production. For our post-industrial (Third Wave) society to have come into existence, did the latter part of the twentieth century undergo a new revolution – an Information Revolution – or did we merely see an acceleration of trajectories set in motion during the Industrial Revolution? This has been a matter of much debate (e.g. Masuda, 1990; Kumar, 1995; Dyson *et al.*, 1996; Castells, 1996, 1997, 1998, 2001; Leadbeater, 1999; Robins and Webster, 1999), and for a fuller treatment of the discussion than is possible here the reader is referred initially to Webster (2004). Revolution or not, it is widely accepted that we live in an information society that has important differences from an industrial society.

What is it about this post-industrial age that makes it ‘informational’? Most commentators point to the computer as being the core innovation technology and driving force, much as the harnessing of steam power was to the Industrial Revolution. But just as steam on its own was not a revolution (the machines it drove needed to be invented, assembled and used for a purpose), so too is it with computers. But there are two points to be made here. Firstly, the term ‘computer’ commonly refers to that bundle of technologies (screen, keyboard, motherboard, hard drive, mouse) that many of us use on a daily basis in the office or at home for work and recreation. At its heart, though, is the *microprocessor* or ‘chip’ (Figure 1.1). These are



**Figure 1.1** The engine of the information society – chips with everything (photograph by the authors).

## Location-Based Services and Geo-Information Engineering

not just contained in computers but are present in a wide and ever growing range of consumer products, such as mobile phones, DVD players, televisions, toasters and car engines, such that in order for them to work there needs to be a flow of data and the running of software algorithms. Thus it is not just the progressive miniaturization that is key here, but the progressive imbedding of intelligence into just about every device we use. Secondly, what has truly made a difference has been the *networking* of computers and other microprocessor-based devices using modern *information and communication technologies* (ICT) linking computers with computers, computers with individuals and individuals with each other on an unprecedented scale (Figure 1.2). By way of an example, in 2003 an average of 55 million text messages were sent every day from mobile phones in Britain. Not only do these networks ‘produce complex and enduring connections across space and through time between people and things’ (Urry, 2000 p. 192), they even shape our economic and social structures (Castells, 1996).

If First Wave agrarian societies are characterized by working on crop and animal production in nuclear units (family, clan) and Second Wave industrial societies are characterized by working with harnessed power and machines in massed units (factories), then Third Wave informational societies are characterized by working with digital media in collaborative networks. With each successive wave, some aspects of previous waves are rendered obsolete whilst others become



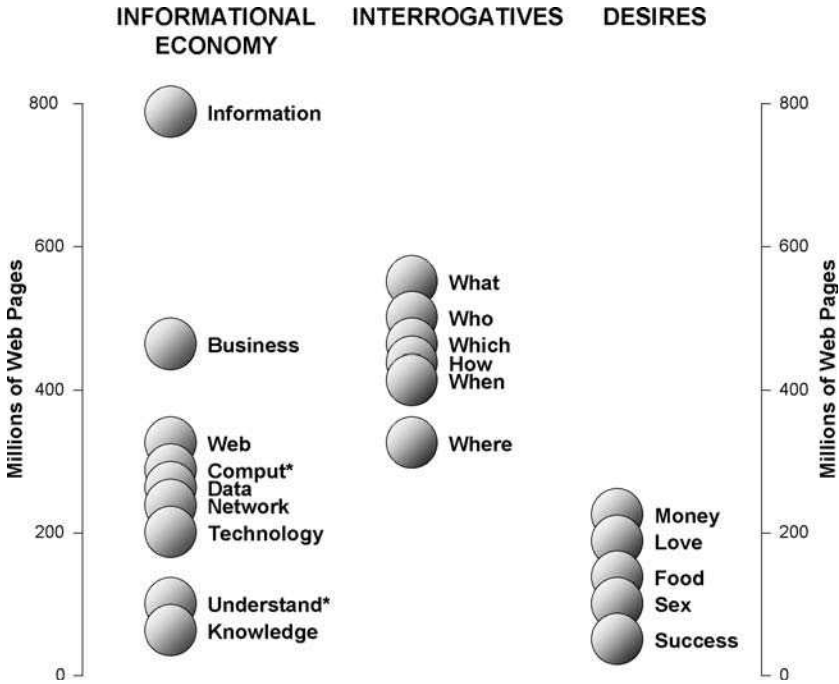
**Figure 1.2** Networking the world – satellite dish farm near Canary Wharf, London (photograph by the authors).

redefined. A progressive example of this latter process would be the industrialization of agriculture through mechanization followed more recently by the incorporation of informational elements that have played a part in both the development of genetically modified crops (knowledge of gene sequences) and in the resurgence of organic farming (knowledge of environmental and nutritional consequences of pesticide use). Castells refers to the Third Wave as both the ‘informational economy’ (1996, p. 66) and the ‘new economy’ (2001, p. 22), that is ‘an economy in which sources of productivity and competitiveness for firms, regions, countries depend, more than ever, on knowledge, information and the technology of their processing, including the technology of management, and the management of technology’ (1997, p. 8).

Knowledge that is actionable (and hence a tradable commodity) is central to the new modes of wealth creation and *cyberspace*, the networked ‘space’ created by ICT, the Internet and the Web, is simultaneously a means of knowledge storage, communication and innovation – all in electronic form. Networks and the ‘space of flows’ (Castells, 1989 p. 348) are the engines of the informational economy; innovation is its fuel. Whilst useful information is certainly present, the anarchy of cyberspace means that it also contains false ideas, propaganda, hype, spin and deviance. It is also huge, growing and immeasurable.

To try and get a snapshot of its size and to get a flavour of its relative content, a range of keywords that represent ‘informational economy’, ‘interrogatives’ and ‘typical desires’ were used in a search engine to see the number of Web pages (just one dimension of cyberspace) that were offered in response. These are given in Figure 1.3. Of the keywords chosen, by far the biggest response is to the keyword ‘information’ with 789 million Web pages – if anybody was to spend only one minute reading each page, it would take him/her some 1500 years to get through them! The disparity in the ranking between ‘information’ and the keyword ‘knowledge’ may merely reflect that the former is commonly used to mean the latter. Of the interrogatives, the question ‘what’ is clearly more important followed by ‘who’, whilst ‘where’ trails by more than 100 million Web pages. This may have repercussions for LBS in that location may be of less interest in the information society – cyberspace after all acts to shrink, even annihilate, space and time. On the other hand, this may only reflect that geographical analyses and communication of information about location have traditionally been the least easy to effect, something that is rapidly changing, for example, with Web-based GIS; but more on this

## Location-Based Services and Geo-Information Engineering



**Figure 1.3** A Web survey of responses to keywords in a search engine (\* indicates variants of the term, e.g. Comput\* includes computer, computers, computing. Source: [www.altavista.com](http://www.altavista.com) on 16 August 2004).

in later chapters. Perhaps the best barometer of cyberspace is to compare the informational economy keywords with those of our typical desires. Keywords for these desires are considerably down the ranking by comparison, though between them of course ‘money’ leads ‘love’ by a small margin!

As an indication of the rate of growth of cyberspace, the numbers of Web pages reported in Figure 1.3 have risen 10-fold from when this mini survey was first run five years ago. However, not all that can be found in cyberspace is necessarily useful. Knowledge utility in terms of wealth creation in the informational economy is time-bound. Of highest utility is *customized* knowledge, that is the right source(s) of information combined, presented and communicated using the right software at precisely the moment it is required (Dyson *et al.*, 1996). Compare this with the definition of LBS given above and it can be seen that LBS are precisely Third Wave products and typical of what creates wealth in an informational economy.



Will we look back and see the information society as a revolutionary change? It may be too early to tell, but information, innovation and creativity have always been the sources of increased productivity and economic growth. Wiener, the originator of cybernetics, put it thus: 'to live effectively is to live with adequate information' (1968 p. 19). What has accelerated in a dramatic way is the rate of knowledge production and the rapidity of its exploitation. There are three aspects to this. Firstly, there are the collaborative networks in which communication over any distance to any number of actors has become almost instantaneous. These have increased research productivity and shortened the time in which the knowledge thus produced is translated into commercial products (Leadbeater, 1999). Secondly, and as part of this, what was a longstanding distinction between the production of knowledge and its communication (Bell, 1980) has merged into an integrated process. Thirdly, governments have assumed a leading role in promoting the information society (Kumar, 1995). In the United Kingdom, e-government has been actively promoted with ambitious targets: 25% of services available on-line by 2002, 100% by 2005 (<http://www.ukonline.gov.uk>), which in the end proved to be too ambitious! High level 'e-government champions' were appointed in each local authority to ensure compliance. 'Citizens will expect to reach the services they want at times and in places that are convenient to them' ([www.ukonline.gov.uk](http://www.ukonline.gov.uk)). Notwithstanding this, it is seen as a means of modernizing government, making it accessible and achieving IT-related efficiency gains (i.e. holding or reducing costs). As part of this agenda, the government has also been promoting the use of computers and IT literacy through schools, universities and public libraries in a bid to assure accessibility and to create a desire amongst citizens to engage with the information society.

But the rise of the information society has had a dark side: inequalities and social exclusion, after nearly a century of decline, have been steadily on the rise over the last quarter century (Kumar, 1995; Castells, 1998; Leadbeater, 1999). Inequalities arise where individuals or social groups, relative to each other, have differential abilities to appropriate wealth (income and assets). Social exclusion, on the other hand, occurs where individuals or groups, regardless of absolute levels of wealth, find themselves systematically barred from access to, for example, average levels of services for health, education and housing or facing above average levels of risk to being victims of crime.

Both inequalities and social exclusion can be difficult to measure and track over time, but in London, for example, since 1984 incomes



have become more polarized, poverty has intensified, mortality differentials have increased and the poorest areas have seen the sharpest rises in violent crime (London Research Centre, 1999). There are a number of trends that have contributed to this. In the informational economy, businesses need to be flexible because the life cycle of products has been shortened and because sources of innovation are less likely to be internal but derived from collaborative networks. This has translated into many businesses downsizing, flattening of the management structure, outsourcing and subcontracting. Labour, in turn, has had to change with rapid growth in self-employment, part-time work and temporary work, particularly for women (Castells, 1997).

The information society is very uneven geographically. It is very strong in general in North America, Europe, urban India, urban China, Southeast Asia and Australia. It is weakest in most of Africa and rural Asia and parts of Latin America. But within those countries that are strongly engaged, there exist geographical pockets, such as the South Bronx in New York and Tower Hamlets in London, that have low levels of engagement (Castells, 1997). Thus the divide is not along the traditional opposition of First World and Third World, leading Castells (1998 p. 337) to conjecture the emergence of a 'Fourth World', which represents 'segments of societies, areas of cities, regions and entire countries' that are slipping into the margins of the informational economy. This is popularly framed as a 'digital divide' (NTIA, 2000). Hull (2003) points out that in the United Kingdom only 3% of the poorer households are on-line as compared with 48% of affluent households – digital restructuring is thus tending to deepen inequalities. Any divides caused by possession of hardware are likely to close due to impending saturation of not just personal computers (PCs) but more importantly of Internet-enabled mobile telephones. Barriers to engagement then will still be to some extent financial (the cost of being on-line, particularly from mobile devices), but are more likely to form along lines of digital skills and usage opportunities (Dijk and Hacker, 2003) which will most probably correlate with levels of educational attainment.

To sum up this section, the rise of the informational economy has provided new means of wealth creation but is still rooted in mass consumption and price utility. Information inevitably has a 'sell-by date', a finite time in which its utility can generate revenue through innovation. When it becomes 'common knowledge' its price can become minimal, but if still above the marginal cost of production (which for digital products can be close to zero) it can still generate

wealth if accessed by large numbers of customers. For knowledge-workers making a living out of the new economy, what is the half-life of their ability to produce new ideas? How long before individuals are consigned to the margins as ‘burnt out’, or indeed as the ‘never glowed’? Lifelong learning and continuing professional development have become rooted in society. But it will be collaborative networking that will allow us to maximize our exposure to sentient knowledge and nascent ideas and, therefore, maximize our chances over a longer period (hopefully a complete working life) for coming up with new ideas, noticing opportunities and participating in innovation.

### 1.3 The Digital City

---

Like the information society, the digital city is a topic of considerable debate of which only a brief cross-section can be provided here. For millennia built environments have been created – from the time we learnt to clear land for agriculture, build villages for shelter and harness water for irrigation. Today there are very few completely natural environments left; such is our impact on our planet. We have been creating cities around the specialization of human activity from the time we could create consistent (though not always reliable) agricultural food surpluses. Cities have always been focuses of information for trade, governance (both internally and of their hinterlands) and social organization for wealth creation. The geographical location of cities has always maximized accessibility to people, goods, capital and, of course, information. From this, cities have in turn created hegemonic power. The growth of cities, both in number and size worldwide, is testimony to their endurance as the dominant means of social and economic organization regardless of political ideology.

Second Wave industrialization, through its benefits to life expectancy and its needs for massed labour and mass consumption, served to accelerate city growth and the spread of urbanization. In the first half of the nineteenth century, city populations were growing at rate of 12% per decade; by the second half of the nineteenth century this had risen to 26% per decade (Carter, 1981). In the last half of the twentieth century the growth rate had reached 46% per decade. United Nations projections (in Fox, 1984) show that the number of metropolitan areas worldwide with a population in excess of five million inhabitants will have risen from 34 in 1984 to 93 in 2025, with some

11 of these metropolitan areas likely to have populations in excess of 30 million. By 2025 there are expected to be 5.1 billion city dwellers, nearly two-thirds of world population. Cities will be huge, networked and, above all, complex.

The Second Wave industrialization, which for some two centuries had been the driving force of city building, has within a generation suffered substantial absolute and relative decline in the face of serviced-based industries (Wheeler *et al.*, 2000). The so-called ‘death of distance’ induced by the wholesale adoption of modern ICT at business and personal levels, and which would remove the need for agglomeration of businesses and employment, was considered to herald the death of cities. Graham and Marvin (1996) and Graham (1998) have summarized a number of competing conceptualizations of the changing relationship between space, place and information technology:

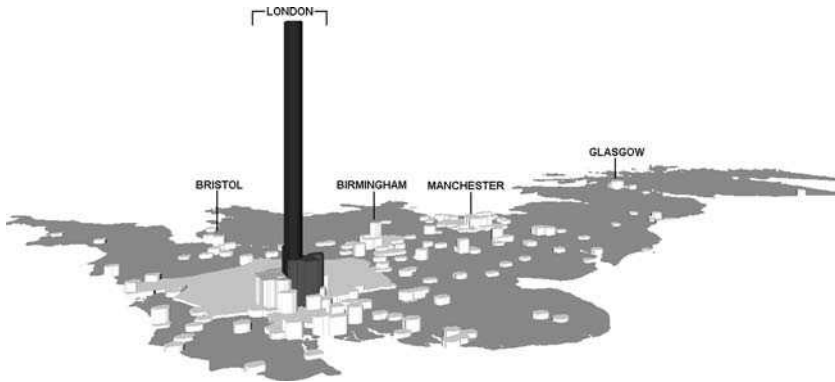
**Technological determinism.** In this conceptualization, ICT and the so-called ‘information superhighway’ would direct change in some simple deterministic way. The main driver would be the massive shrinking or even death of distance which would liberate human life from the frictional effects of distance. Since agglomeration into cities has been the traditional means of overcoming these frictional effects, the removal of distance as a constraint to human interaction would lead to areal uniformity – a sort of veneer of tele-cottages – and the dissolution of cities. In this scenario we would become tele-workers, tele-shoppers and indulge in tele-socializing using networked virtual reality technologies that would provide all the richness of face-to-face activities.

**Co-evolution of geographical and electronic spaces.** Under this conceptualization the technological determinism argument has considerably exaggerated the degree to which ICT can substitute for face-to-face and other place-based activities. Instead, electronic networks will co-evolve with physical spaces and social processes to produce new synergies and new urban forms. Thus residents may be able to tap into digital resources to enhance their place-based experience of the city. Indeed, LBS may have a key role to play here. Such a scenario may intensify city development in symbiosis with the growth of telecommunication networks and the acceleration of their transmission rates. Cities, far from shrinking, will in fact continually be recast due to the complex evolving relations between capital, technology and space.

**Recombination through actor-networks.** This view relies heavily on actor-network theory, in which absolutes have little meaning and much depends on the relational process of agency. Thus networks are not fixed absolutes but are continually recombined at any moment by the actors using them and thus the specific social contexts and power struggles within which usage takes place. For digital networks, however, many of the actors are themselves algorithmically responding machines and this is likely to blur the boundaries between humans and machines to produce a cyborgian conceptualization of reality. These will act to continually produce new forms of human interaction, organization and control. However, one clear consequence of the ever increasing rate of information production and dissemination appears to be an increasing demand for face-to-face contact as the most effective means of resolving information glut. The agglomeration advantage of cities for actor-networks seems set to endure.

Looking around our cities today we can see elements of all three conceptualizations at play, though it seems that technological determinism has the least power to explain current city processes. From the introduction of the telephone through to the construction of optic fibre networks, cities have seen preferential development of telecommunications infrastructure. Indeed ‘... cyberspace is, in fact, a predominantly metropolitan phenomenon which is developing *out* of old cities’ (Graham, 1998 p. 173, original emphasis). Cities represent lucrative concentrations of demand where revenues from ICT can be maximized with the least outlay in infrastructure. Deregulation of telecommunications in the 1980s and 1990s led to concentrations of providers in cities where they could ‘cherry pick’ Third Wave information-intensive markets. Cities that long benefited from excellent physical access are now the beneficiaries of excellent ICT infrastructures (Niles and Hanson, 2003) and are also centres, for example, of Internet content provision because this is where the expertise has accumulated (Kellerman, 2000; Castells, 2001).

Measuring this phenomenon is not easy, but the degree of clustering of the telecommunications industry in Britain as an urban phenomenon is illustrated in Figure 1.4. As a statistical measure of the clustering, the median distance from any one company to its nearest neighbour was found to be 778 metres – just a few street blocks away. The pre-eminence of London in Figure 1.4 as a centre for telecommunications reflects its position as a global city. International networks



**Figure 1.4** Density clustering of the telecommunications industry in Britain: a largely urban phenomenon with highest concentrations in and around London (density per square kilometre using Geo-ProZone spatial clustering technique – see Brimicombe, 2006; base data compiled by the authors from industry sources, 2001).

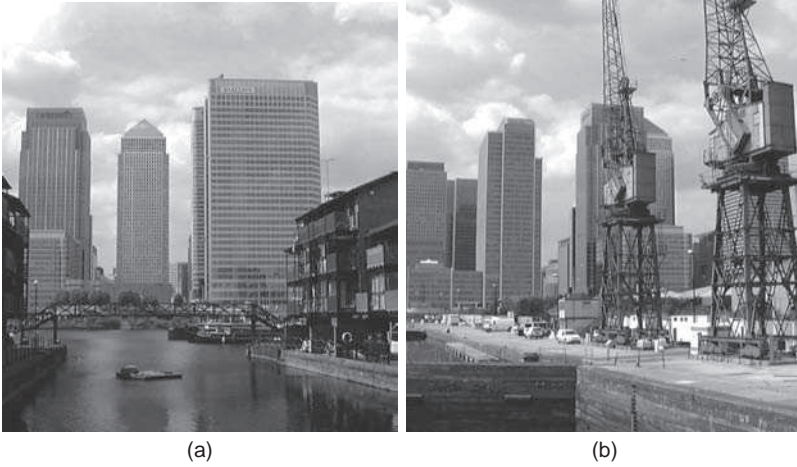
use cities as nodes due to their potential to maximize traffic and revenues. Reed's Law (in Rheingold, 2002) holds that the value of a network grows exponentially to the number of users, in other words, the social and business networks that are created are always far larger than the physical network infrastructure. Consequently, those nodes (cities) that most intensively use ICT and knowledge-based industries for wealth creation have emerged as global cities: their reach and ability to control and influence at a distance dramatically increases their power. For cities, 'being digital' is the key to wealth and status (Negroponte, 1995). Furthermore, to be a key node in the network of cities is to be at the heart of the world economy.

Whilst it is clear that ICT and other modes of communication can substitute for personal movement, they can also complement and stimulate it (Hall, 1999). Face-to-face exchange of information remains important. Whilst collaborative networks are becoming the basic units of innovation and production (Leadbeater, 1999), ICT cannot provide a complete substitute for face-to-face contact, particularly where building a relationship of trust is important. Given the ease of availability of information and the propensity towards overload, face-to-face activities such as listening, advising, brainstorming and negotiating, either in the formal setting of an office or in the informal setting of a restaurant or bar (or even on a golf course), remain an important means of resolving issues and finding new synergies. Castells (1989, 2001) refers to this as the 'milieu of innovation', that

is the physical concentration of clusters of people and companies networking together, both electronically and socially, to produce innovation by synergy. The term 'milieu' does not just express 'environment', it also means 'middle' – expressing the notion that the best place to be is at a node where there is maximum network accessibility to people, information services and resources. The physical accessibility afforded by cities remains an important factor and it is still the case that collaborative networks are most easily established in cities. Thus, grounded (place-based) social relations endure as an essential ingredient in the information society (Niles and Hanson, 2003).

The new economy, as well as enhancing the power of cities, has also brought important changes in structure and complexity. Whilst the value of face-to-face contact endures, not all activities require it. Thus back offices, given the 'death of distance' over networks, can be just about anywhere where there are good ICT infrastructure and suitable personnel. Thus call centres for UK companies are sited from Scotland to India; similarly, document processing for banking, insurance, law and publishing. Product development in areas such as software, graphics and CAD-based design can similarly be networked around the world. These types of activities have left the traditional city cores for cheaper premises and lower wages, often in Third World countries. In these cases ICT have led to dispersal. On the other hand, ICT have promoted the agglomeration of corporate headquarters, R&D and high-end service functions (such as financing, marketing, protecting intellectual property rights) so as to reduce the time and cost of innovation and hence compete effectively (Wheeler *et al.*, 2000). New industries concentrate alongside these to take advantage both of advanced ICT infrastructures and the availability of financial and other services. These agglomerations, however, are not necessarily sited in the traditional central business district (Hall, 1999) but have tended to form new centres. Thus, in London concentrations of knowledge-based industries are to be found in the West End (formerly predominantly high class residential) and in Canary Wharf (formerly docks). Indeed, Canary Wharf has grown to become London's second central business district (Figure 1.5). This trend has led to the creation of 'edge cities' and has changed the overall city structure from monocentric to polycentric (Hall, 1999). As well as city structure, ICT are also transforming public spaces within our cities (Moss and Townsend, 2000). Our experience of public spaces is already frequently augmented by simultaneous use of mobile phones and whilst in public spaces we tend to be under surveillance by CCTV and other





**Figure 1.5** An urban focus of the Third Wave economy – Canary Wharf, London: (a) new tower blocks containing typical knowledge-based industries and waterfront housing for its affluent workers; (b) with preserved vestiges of its Second Wave heritage in the foreground (photographs by the authors).

technologies (discussed further in the next section). Even the architectural facades in high accessibility areas such as Times Square in New York, Shibuya and Ginza in Tokyo and Piccadilly Circus in London are increasingly being designed as communicators of information and digital augmentations of place-based experience (Figure 1.6).



**Figure 1.6** Informational city; informational facades: (a) Ginza, Tokyo (photograph by the authors); (b) Times Square, New York (photograph by Y. Li, used with permission).

Finally, the term ‘digital city’, in addition to its common language meaning of a digitally wired (and wireless) city, has recently come to have a more specific meaning: that of a city that is entirely digital or virtual (Batty, 2001; Couclelis, 2004). Such a digital city is defined by Couclelis as ‘a comprehensive, Web-based representation, or reproduction, of several aspects or functions of a specific real city, open to nonexperts’ (2004, p. 5). She considers that a digital city could provide the core for the ultimate LBS as well as for participatory governance and decision support, virtual and augmented reality and wearable computing interfaces. Some Internet service providers (ISP) in the United States have already been developing extensive networks of city Web sites to deliver locally targeted information (e.g. <http://www.digitalcity.com>). Digital cities are now both real and virtual; they are also both visible as place-based experiences and invisible (Batty, 1990) as more and more functions of the city become reliant on automation, digital networks and microprocessors.

### 1.4 The New Mobility

---

Flows of information and capital produce flows of people. We have seen in the previous section how ICT have allowed the dispersal of some activities and the agglomeration of others – a reshaping of cities and our working and leisure patterns – in which face-to-face contact, though augmented by devices such as the mobile phone, still plays a vital part in the flow of information, ideas and innovation. In a naïve sense then, we travel more; we have become geographically more mobile. Second Wave industrialization created landscapes shaped by the internal combustion engine and, more particularly, by the private automobile. The same wave of deregulation in the 1980s and 1990s that facilitated an explosive growth in telecommunications, also brought more competitive banking, credit facilities and money transfers, widened the competition in air travel and brought on-line shopping and booking of just about anything. Since then the cost of international travel has fallen in real terms and it has never been easier to organize one’s own travel itineraries and pay for them using credit cards. ICT have also allowed us to ‘break loose’. Traditionally, to travel and be away was to be out of touch. Now mobile phones allow us to be instantly reachable, allow us to constantly stay in touch even during the most routine of commutes (Figure 1.7).





**Figure 1.7** On the move, keeping in touch: commuters on the Tokyo underground (photograph by the authors).

Beyond the naïve sense of increased, massed travel (and remembering that travel itself is nothing new), we are benefiting from other forms of mobility that result from the information society. First amongst these, and linked to notions of physical travel, is imaginative or virtual travel (Urry, 2000). This can be over the Internet at the PC, at the cinema (where computer-generated graphics are increasingly used to create fantasy worlds) and on the television. How many people around the world were able to watch and *experience* from their offices and living rooms in real-time (or very soon afterwards), the unfolding horrors of the World Trade Center in New York in September 2001 or of the Middle School No. 1 in Beslan, North Ossetia, in September 2004? Whole TV channels (satellite, cable and terrestrial digital) are dedicated 24/7 to travel programmes and others to world news – where it happens, as it happens. Web-cams and so-called ‘reality TV’ of the Big Brother genre have proved immensely popular, as if experiencing our own individual, physical life somehow needs to be augmented by digital voyeurism.

In another sense, the new mobility is also a social mobility in which traditional class structures and the barriers they imposed have been dismantled. Access to information has also meant that traditional gender-based structures that result, for example, in inequalities of jobs and pay and the notorious ‘glass ceiling’ in the promotion of women employees are also being chipped away.

With e-mail and texting, we are more able to maintain large, spatially separated networks that can be key to social and other

mobilities. E-mail and texting have become hugely popular activities because they are 'personalised, spontaneous and interactive; the content of a particular message is usually tailored to the recipient . . . they sustain ongoing dialogues and relationships' (quoted in Niles and Hanson, 2003 p. 39 from Keisler *et al.*, 1997). This has expanded into on-line social networks such as Facebook and MySpace and in the creation of on-line content through blogging. Second Life combines the new mobility with aspects of the digital city. One's own persona (or indeed an alternative, even re-gendered persona) can be projected through an avatar to have status, travel and even acquire wealth in virtual spaces.

We also have enhanced mobility as consumers of products and services. Travel does not limit us to what can be consumed locally; credit and debit cards have allowed us to detach spending from proximity to bank branches and, as already stated, with credit cards we are able to buy just about anything on-line or over the (mobile) phone. Consumer mobility, on the one hand, has resulted in increased competition and new markets and, on the other hand, has resulted in increased opportunities to sway choice through advertising, junk mail and spam e-mail. In yet another way, we have become more informationally and intellectually mobile (Dogan and Pahre, 1990). In one sense, information resources have become much easier to access in abundance (Figure 1.3) and it has become much easier to seek out for oneself, say, expertise that was once the sole purview of chartered or licensed professionals. In another sense, intellectual mobility is necessary for innovation, in other words, it is often at the cross-disciplinary frontiers of several disciplines that innovative ideas emerge. This type of intellectual frontier hopping is not just a one-off event for knowledge workers, but is a repeated necessity.

It is clear that the new mobility is both horizontal and vertical; in the words of Castells *et al.*, (2006) we have become a 'mobile network society'. The mobility of people together with the mobility of information, images, capital, risks and objects can produce creativity, innovation, wealth creation and freedoms. But it can also produce changes of shaping and reshaping of our environment, systems, networks and flows in chaotic and unpredictable ways (Graham and Marvin, 1996). Furthermore, mobility depends very much on knowing 'destinations' and these destinations may not be fixed points. For example, our ability to access on-line information and services is contingent on knowing what is available on-line and being able to navigate our way there (Niles and Hanson, 2003). Even then, how

often do we try and access a Web page that is no longer there? It is not surprising then that a range of 'wayfinding' tools has emerged, some of which have become indispensable. Internet search engines such as Yahoo, Alta Vista and Google now constitute one of the fastest growing industries as they help users sift and navigate through the billions of Web pages to where they want to be. On-line mapping services such as MapQuest and Google Maps have also become big business as users seek out maps and navigation instructions on how to reach geographical destinations.

It should be remembered, however, that wayfinding tools may not always be impartial with regard to what information is being provided. One example is 'pay for placement', which is active on most search engines, whereby certain pages are always positioned, by subscription, high up on relevant listings. This makes it more difficult to find highly specialized or local information (Niles and Hanson, 2003) and is a subtle control on wayfinding choices and hence mobility. But this should not come as unexpected since control has always been at the heart of information (Kumar, 1995). Indeed, the complexity and uncertainty induced by the new mobilities has intensified the urge (some would say the need) to control (Robins and Webster, 1999). Thus shadowing the growth of the information society has been the Surveillance Society. Whilst surveillance itself is not new, it is the combination of sensors, networks, databases, software and alarms that has made possible automated digital surveillance. 'Advances in the technologies of sensing and recording have enabled a massive growth in the monitoring of individuals and groups ...' (Graham and Wood, 2003); surveillance is now everywhere and in cities can be considered as ubiquitous. Thrift and French (2002) cite the example of the UK Audit Commission which found it almost impossible to identify all the digital surveillance systems in operation given that so many of them were invisibly embedded into cities. Such attention to surveillance has come about for two main reasons:

**Commodification/privatization of places and services.** The trend of deregulation and privatization of services means that access is no longer free and no longer a right. Access therefore needs to be registered, monitored and controlled. Such access is often based on eligibility and the ability to pay (which itself is determined by surveillance, e.g. credit ratings). The presence of digital

surveillance techniques ‘make possible the widening commodification of urban space and the erection within cities of myriad exclusion boundaries and access controls’ (Graham and Wood, 2003). In London, for example, forms of digital surveillance in operation include CCTV, congestion charging, mobile phone cell tracking (and recording of usage), Oyster cards (for multi-modal public transport), chip and PIN point of sales (using credit and debit cards), loyalty cards, cash machines, offender tagging, face prints, iris scans, DNA profiling, motion detectors, pressure pads, swipe cards, Web logs and Internet cookies. With so much surveillance, invasion of individual privacy has become a topical issue, but one which is subverted by the collective need for personal safety.

**Crime and terrorism.** Crime of all types has risen dramatically over last half century and in the 1970s and 1980s firmly took its place on the political agenda (Garland, 2000). The rising crime rates correlate well with the post-1945 rise of mass consumerism and the transformation to a materialistic and acquisitive society (Reiner, 2000), but have been further fuelled by the rising inequalities and social exclusion discussed above. The inevitable response to the consequent fear of crime has been increased surveillance of our public spaces. Terrorism has also been on the rise. Masuda (1990) considers that whereas international conflict and the wars that result are problems of industrialization, acts of individual and group terrorism are problems of an informational society. According to Ruthven (2004), fundamentalism (often linked to terrorism) is a consequence of spreading secularism and intrusion by TV, film, video, news, tourists and outside political pressure that threatens the continuation of traditionalism. In this sense, fundamentalist terrorism is at once a response to an invasive information society yet using its very infrastructure (mobile telephones, collaborative networking, instantaneous media broadcasting and so on) both to plan and perpetrate acts of terrorism and to ensure these acts have maximum media exposure and hence impact worldwide.

This is not an up-beat note on which to end our rapid tour of the information society, the digital city and the new mobility, but these are all dimensions of the context in which LBS are emerging to fill a niche for data and information that is tailored to one’s circumstances and needs – as it is needed, where it is needed. Whether this be travel information, finding a restaurant, dealing with e-mails whilst there, knowing whether contacts or friends happen to be nearby, receiving a

warning of traffic delays or simply getting an update on security alert status at the airport where one is going to be, these (and many more) are all facets of our complicated lives where LBS are potentially useful.

### 1.5 Outline of Following Chapters

---

Having set the broader social and economic contexts for location-based services, Chapter 2 will move forward with an in-depth analysis of a number of technologies that have developed and converged over the last quarter century so as to make LBS technically possible. Prime amongst these will be the telecommunications networks servicing mobile devices. In the last few years mobile phones have achieved the status of a programmable device for voice, text, images, video, sound, data, FM and digital radio, and most recently for digital TV. It is rapidly becoming indistinguishable in function from a hand-held computer in addition to its status as an indispensable item of dress code and social standing. Chapter 3 provides an overview of geographical information systems as a key technology in handling location-based spatial queries and therefore at the heart of LBS. However, given that it is now well over a decade that geo-information science (GIScience) emerged from the maturing GIS technologies and their applications, we take a further evolutionary step in considering LBS as an early manifestation of *geo-information engineering*, that is the design and construction of reliable (consumer) products based on the sound application of the relevant science and technologies. Chapter 4 looks at the short history of LBS, its initial drivers, its architecture, regional differences (Unites States, Europe, Japan) and some key content that has emerged to date. This chapter provides a more informed discussion of the definition of LBS. There then follow four chapters (Chapters 5–9), each on a specific aspect of operational LBS: data, locating the user, using context, spatial queries and communication. Each considers technical and implementation issues. The book finishes in Chapter 10 by discussing a number of issues to do with the business of LBS, including a range of business models that providers might adopt as well as some of the ethical, legal and privacy issues that arise from LBS and similar technologies.

# Chapter 2

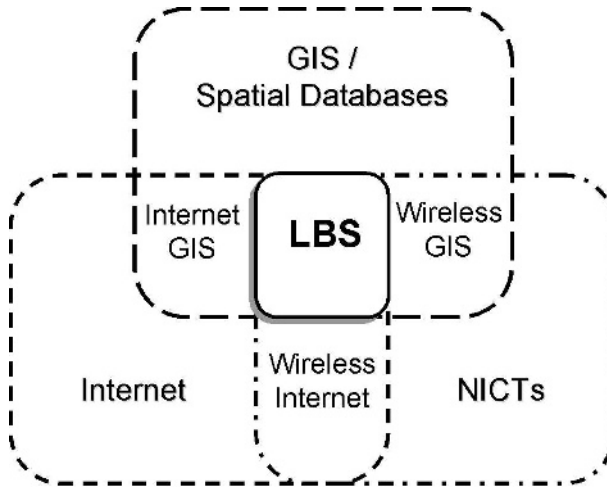
## Technological Convergence

### Towards Location-Based Services

#### 2.1 Introduction

---

This chapter covers a number of technologies which, through their continued development, have combined to make location-based services (LBS) possible. LBS are heterogeneous technologies and in order to understand the development and convergence of the technological strands that come together in LBS, a framework is used (Figure 2.1) which shows LBS at the intersection of geographical information systems (GIS) and other spatial technologies, the Internet and the Web, and new information and communication technologies (NICTs). In this chapter the focus is mainly on aspects of the Internet and NICTs strands. An in-depth focus on the GIS strand is given in Chapter 3. This will then have paved the way for an informed tour of LBS in Chapter 4.



**Figure 2.1** The convergence of technologies towards LBS (from Brimicombe, 2008).

## 2.2 The Internet and World Wide Web

---

### 2.2.1 The Basics of the Internet

The launch of the first artificial satellite, Sputnik, by the Soviet Union in 1957 is widely regarded as the trigger for the development of the necessary technologies to form ARPANET (Castells, 1989; National Research Council, 1997), the precursor of the Internet. The ARPANET network was developed through DARPA (The Defense Advanced Research Projects Agency, USA) and went on line in 1969. ARPANET provided the ability to transfer data and for hosts to communicate on the same network. The first e-mail programme was created by Ray Tomlinson, a Principal Scientist at BBN Technology, in 1972. To enable different computer networks to interconnect and communicate with each other, a new protocol called Transmission Control Protocol/Internet Protocol (TCP/IP) was developed. The term 'Internet' was first used in 1974. The TCP/IP protocol was later used on ARPANET by the US Department of Defense. From the mid 1970s to the early 1980s, a number of networks were developed: SATNET, which linked the United States and Europe using INTELSAT satellites; USENET, created as a decentralized news group network; BITNET, developed by IBM as a 'store and forward' network for e-mail and listservers; and CSNET, created by the

US National Science Foundation for institutions without ARPANET access. The Domain Name System (DNS) for the Internet was also created in the early 1980s to provide an easy way for people to access other servers by using domain names instead of IP addresses. ARPANET was divided into two networks in 1984, one for military needs and another for advanced research. In the late 1980s, a new upgraded network, NSFNET (National Science Foundation Network), was developed with a speed of 1.5 Mbps. In 1989, countries such as Germany, the United Kingdom, The Netherlands, Italy, Australia, New Zealand, Israel, Japan and Mexico joined the Internet. In the early 1990s, networks started to adopt the concept of the T3 and 45 Mbps lines, and the original 50 Kbps lines of ARPANET went out of service. Also in the early 1990s, a hypertext system was implemented by Tim Berners-Lee at CERN in Geneva, Switzerland, to provide efficient information access. CERN released the World Wide Web (WWW or simply the Web) in 1992. In the United States, a new network named the National Research and Education Network (NREN) was developed by the National Science Foundation (NSF) for high speed networking research in the early 1990s. The Internet Society was chartered in 1992. Specific Internet services were provided by InterNIC created in 1993, and a graphical user interface for the Web, 'Mosaic for X', was released in the same year. Mosaic was the first popular graphic Web browser; the advent of browsers accelerating the growth in Internet usage. Many new networks were added to the NSF backbone, and the number of hosts increased dramatically. In 1995, direct access to the NSF backbone was discontinued, and instead Internet connections began to be sold by the companies (acting as service providers) contracted by the NSF. Fees for domain names also started to be imposed. Since 1996, most of the Internet traffic has been carried out by independent Internet Service Providers (ISPs). The Internet has now become indispensable to governments, businesses and to individuals' social and recreational lives.

The Internet, from a technological perspective, is a communication network comprising of many networks worldwide. The Internet uses packet-switch technology by which data are broken into small packets and each packet is sent individually through a series of switches (known as routers) across the Internet. At the receiver end, the data are reassembled back into their original form when all the packets arrive. A set of protocols is used across the Internet for breaking data into packets, routing data packets across the network and



reassembling them at the receiving end. Transmission Control Protocol (TCP) and Internet Protocol (IP) are currently among the most used protocols on the Internet. TCP is a connection-oriented, end-to-end reliable protocol which supports multi-network applications. TCP defines the procedure by which data are broken into IP packets and sent to the receiving end. When these data packets arrive at the receiving end they can do so at different times via different paths in the network. Thus the packets usually arrive out of sequence and are stored temporarily while the rest of the packets arrive. Some of the packets can even be damaged or go missing. TCP defines procedure at the receiving end for reassembling the data packets back to their original data form. When some packets are missing or damaged, a retransmission request will be sent for these packets to be resent. Damaged packets are then deleted. When data are broken into packets, each of the packets is put into separated IP ‘envelopes’. The address information of sender and receiver is contained in the headers of these IP ‘envelopes’. The address information is the same for all IP packets that belong to the same piece of data, so that all of them can be sent to the same destination to be reassembled.

When people use the Internet, they can connect into it from a local area network (LAN), such as their work place or university network. They can also connect to the Internet from home via a broadband connection or a device (such as a modem) provided by an Internet Service Provider (ISP). The LAN and ISP then connect to the Internet using hardware such as routers or bridges. Routers send user requests or data to destinations on the Internet. When users send e-mails, transfer data via FTP (File Transfer Protocol) or access Web sites from their local computer, the requests or data reach the relevant servers via a series of routers on the Internet. Servers can then process requests and send back information across the Internet.

### **2.2.2 World Wide Web**

The Web started at the early 1990s and has been widely used for accessing, exchanging, processing and disseminating information and services to the public. The Web works on a client-server model. The client software, known as a Web browser, works on a user’s local computer; the server software operates on a computer elsewhere on the Internet. When users request information via a Web browser, the

browser sends the URL (Uniform Resource Locator) request using HTTP (Hypertext Transfer Protocol) through the Internet. The request is sent to the destination server by routers. The Web server receives the request using the HTTP protocol, finds the requested documents on its mass storage and sends the data back to the client Web browser. The information required is then displayed on the user's computer screen via the Web browser. One of essentials of the Internet and the Web is that information and services reside on networks and flow across the network between people and organizations. The volume of information and types of services available over the Internet continue to grow dramatically. The number of people and time spent on-line has correspondingly increased. The Internet has become an integral part of society. With the development of wireless mobile networks and near ubiquitous use of mobile devices, information and services on the Internet can be accessed and used by people anytime on the move.

### **2.2.2.1 The Continued Evolution of the Web**

The ease with which the Internet and the Web can be accessed has created an unstoppable movement for the exchange and processing of information. People's use of the Internet and the Web is diverse: to access and make use of information, find and contact people, create and share content, order and purchase goods. Increasingly, the Web is not only just an on-line information resource, but also an on-line means for people to carry out many more activities, such as working, learning, shopping, entertaining, socializing and networking. There are international communities formed on-line with mutual interests for research, business, hobbies, buying and selling, and (more recently) for social networking. Such community building can be used for marketing and sales, enabling more communication with customers in order to offer tailored services and improve customer relationships. It can also be used for increasing productivity in business by sharing information, managing knowledge and stimulating innovative ideas. Such community building on-line has been used in people's social and leisure life. There are ranges of software and tools for such Web activities, such as Wikis, Blogs and various so called 'social book-marking' tools. The contents of the Web are dynamic with more user participation. People can create integrated hybrids of new forms of information, services or applications by reusing distributed information and data in what is often called *mashups*. A mashup is a Web site

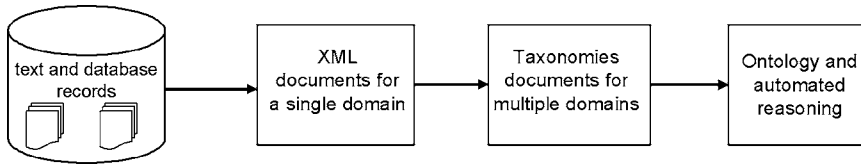
or an application that combines content from more than one source on the Web into an integrated experience.

All of these changes in the Web have led to what is often referred to as *Web 2.0*. Up to now, there is no firm consensus on a single definition of Web 2.0 or on what constitutes Web 2.0. But, the characteristics of Web 2.0 are generally recognized as collaboration, user involvement with user generated content. Web 2.0 shows the potential of ‘social networking’, improved collaboration in business and user participation with user-generated information that is related to themselves and feedback about information and services. Web 2.0 is sometimes viewed as a fuller realization of the Web’s potential.

### 2.2.2.2 Semantic Web

‘The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web – a Web of data that can be processed directly or indirectly by machines’ (Berners-Lee, 1999). Whatever the use of the Web, data on the Web need to be processed by machines for people’s usage. Take a Web search engine as an example (see also Section 5.3.4). Search engines are some of the most valuable tools employed in the Web. However, everyone might have experienced to some degree problems related to their use. Such problems can be: search results obtained can have high recall with low precision or low/no recall; results are highly sensitive to the vocabulary used for searches; search results only list individual Web pages (Antonioni and van Harmelen, 2004). The main obstacle to information retrieval centres on the meaning or semantics of Web content that is currently not machine-accessible. Using intelligent techniques to represent machine-accessible Web content by meaning is the initiative of the Semantic Web.

The Semantic Web aims to achieve semantic interoperability of data in order to achieve application independence, improved search facilities and improved machine inference on the Web (Daconta *et al.*, 2003). Under the Semantic Web concept, data on the Web have their meaning processed by computer. Data in such a form are often referred to as *smart data* and the Semantic Web can be considered as an information ecosystem of application independent smart data that can be machine processed over the Internet. Creating a Web of smart data can be considered as the progression of creating data with increasing intelligence. The basic stages in this progression are illustrated in Figure 2.2. At first, most data belong to specific applications, in other



**Figure 2.2** The progression from data to smart data (based on Daconta *et al.*, 2003).

words, documents and data are very much application dependent. In the second stage, data can be used between applications if they fall within a single domain. Use of the XML (eXtensible Markup Language) standard within a single knowledge domain can be seen as one such example. In the third stage, data are related and compiled from different domains and classified into a hierarchical taxonomy by using the relationships between categories in the taxonomy. At this stage, data are now ‘smart’ enough in certain degrees to be easily searched and combined with data from different domains. In the final stage, the ontologies and rules stage, data are described with more reliable relationships and formalisms that allow logical calculations to be made using ‘semantic algebra’ (not dissimilar to the relational algebra in Section 8.7.1). The combination and recombination of data can be realized at a more atomic level, and the analysis of data can be carried out at a high level of granularity (detail). At this stage documents on the Web need no longer be treated as individual entities but as part of an integrated and analysable whole. This will have important implications for LBS.

To realize the aims and functionalities of the Semantic Web, a number of concepts and technologies need to be deployed, including explicit metadata, ontologies, logic and intelligent agents (Antoniou and van Harmelen, 2004). These are described in principle as follows.

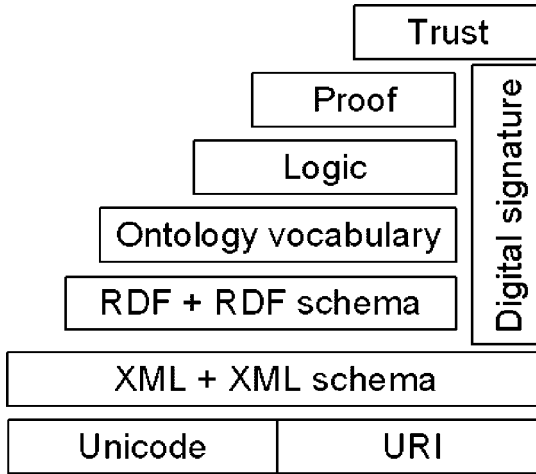
- **Metadata** can be simply described as data about data, which depict the meaning (semantics) of data. Most of the current Web pages are written only for presenting content to readers. In the Semantic Web, there will need to be information describing document content. Such information will enable machines to process the content and would therefore be regarded as machine-processable metadata. Currently, XML is starting to be adopted to create Web page metadata for this purpose. See also Section 5.4.3 for metadata in an LBS data context.

- **Ontology**, originated from philosophy, has been used in the Computer Science area with its specific technical meaning attached. Employed in the Semantic Web, ontology can be defined as an explicit and formal specification of a conceptualization (Gruber, 1995; Studer *et al.*, 2006). Ontology specifies a domain of discourse (Section 3.5 covers ontology in GIScience). Ontology usually comprises an agreed list of terms and the relationships between them (Antoniou and van Harmelen, 2004). These terms refer to important concepts (classes) within each domain. The relationship between the terms can be the description of a hierarchy of classes, a property between classes, a restriction, a disjointedness statement or a logical relationship between objects. If element names are understood differently cross different systems or domains, a machine will not be able to reach a conclusion that the meaning of these two elements is the same or similar. Therefore, ontology is needed in the Semantic Web to help organize and navigate content in a Web site. Hence it will improve Web search accuracy and overcome some of ambiguity caused by current keyword search methods. It can deploy information generalization/specialization in Web content searching depending on search results achieved. Currently, the most used ontology languages for the Web include XML, XML Schema, Resource Description Framework (RDF), RDF Schema and Web Ontology Language (OWL). XML provides an open, standard syntax and verbose descriptions for the meaning of data. It has a standard syntax for metadata, and a standard structure for both documents and data. It can create application-independent documents and data. XML Schema is a language for constraining the structure of XML documents, which can restrict XML documents to a certain structure and a specific vocabulary. RDF is a data model for describing objects (referred to as resources) and the relations between them. RDF is an XML-based language. RDF Schema allows users to create their own vocabulary for describing properties and classes of RDF resources. OWL is a Web ontology language that provides semantic expressivity for developing ontologies on the Web.
- **Logic** can offer explanations for which a series of inference steps can be retraced back from the conclusions. Logic used in

the Semantic Web provides explanations that are necessary for activities between Web agents (discussed below). For instance, while one agent draws certain conclusions from a set of search criteria, another agent might query how such conclusions were derived. Explanation can then be given by tracing back to source Web content and the inference rules used to reach such conclusions.

- **Agents**, often described as pieces of software working autonomously and proactively, can be used to perform various tasks in collecting and organizing information according to user requirements. For example, agents on the Semantic Web can seek information from different Web sources, compare them and select particular information according to user requests and preferences. All the concepts and technologies just discussed can be implemented using agent concept and technologies on the Semantic Web, such as: using metadata to identify and extract information from various Web sources; deploying ontologies to assist Web searching, interpreting retrieved information and communicating with other agents; using logic to draw conclusions by processing retrieved information. A discussion of agent technologies in GIS is provided in Section 3.3.2.

The development of the Semantic Web proceeds in a layered approach following two principles. One is the downward compatibility by which higher layers should be able to interpret and use information written at lower levels. The other principle is the upward partial understanding by which lower layers should take at least partial advantage of information at higher levels. The layered approach is illustrated in Figure 2.3. Unicode provides a unique number for every character in a character set, which is independent of the platform, the program and the language used. Unicode is an industry standard designed to allow text and symbols from all the writing systems in the world to be consistently represented and manipulated by computers. A Uniform Resource Identifier (URI) is a compact string of characters used to identify or name a resource. The main purpose of this identification is to enable interaction of resources over a network (typically over the Internet) using specific protocols. In this layered approach, XML acts as the syntactic foundation layer. Built on top of XML are RDF and RDF Schema. RDF is a basic data model, and RDF Schema provides modelling



**Figure 2.3** The layered approach of the Semantic Web (based on Antoniou and van Harmelen, 2004).

primitives for organizing Web content into hierarchies. RFD Schema can be seen as a primitive RDF-based language for writing ontologies. In order to represent more complex relationships between Web objects, more powerful ontology languages (the Ontology vocabulary layer) will be needed to build a higher layer. Above that, there is the Logic layer that further enhances the ontology languages and offers the ability to create application-specific declarative knowledge. The Proof layer involves deductive processes, proof representation and proof validation. At the top is the Trust layer that is based on recommendations by trusted agents, certification agencies, consumer organizations, or on rating scheme.

‘The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation’ (Berners-Lee *et al.*, 2001). Semantic Web represents a set of concepts and technologies which can work not just for the Web but also for information sharing and automated administration, decision support and business development. It can be used for several crucial issues faced by current information technology architectures and which are also relevant to LBS applications, such as:

- **information overload:** there is increasingly fast growth in the availability of information and data for users;



- **stovepipe systems:** most components are hardwired, which leads to problems in information sharing between different systems;
- **poor content aggregation:** this recurring problem is caused by putting together information from disparate sources.

The Web has had an enormous impact on the way people communicate and interact with each other, the way business is operated and the way information is disseminated and retrieved. The concepts and techniques of the Semantic Web should improve and evolve the current Web and make it more useful.

## 2.3 NEW INFORMATION AND COMMUNICATION TECHNOLOGIES

---

The term New Information and Communication Technologies (NICTs) is used in this book to refer to the currently evolving types of information and communication technologies (ICTs) with a strong emphasis on mobility and location awareness. These include technologies related to mobile wireless telecommunication networks, other wireless networks, small wireless mobile devices and positioning technologies, particularly those integrated into mobile devices. NICTs have been widely used in people's daily social and business lives, and in mobile situations. Wireless has increasingly become a ubiquitous way of sending and receiving data and information, through mobile phones, Personal Digital Assistants (PDAs) and other devices, at locations such as offices, homes and in public spaces. A range of standards and systems for wireless telecommunications has been developed for mobile voice telephony and for data communication. Furthermore, wireless networks have been used more and more as a replacement for wired networks within homes, offices and in-building settings through the deployment of Wireless Local Area Networks (WLANs). Moreover, the Bluetooth standard allows wireless connections for communication between appliances within a personal workspace. In this section, those aspects of NICTs which contribute to the technological setting in LBS are introduced. Mobile wireless telecommunication technologies are introduced in Section 2.3.1 and other short range wireless networks in Section 2.3.2. The development of mobile devices is discussed in Section 2.3.3. Location-awareness technology, which is an important part of NICTs and particularly for LBS, is dealt with in depth in Chapter 6.



### 2.3.1 Mobile Wireless Telecommunication Technologies

The ability of radio to provide communication by wireless telegraphy was first demonstrated by Guglielmo Marconi for use on British naval ships. Marconi's first transatlantic wireless transmission was achieved in 1903 (Figure 2.4). The first public mobile telephone service was introduced in the United States after World War II. In the mid 1950s and 1960s, automatic channel trunking was introduced and implemented under the label IMTS (Improved Mobile Telephone Service). AT&T proposed the concept of a cellular mobile system to the US Federal Communications Commission (FCC) in 1968. In 1983, the FCC finally allocated 666 cellular duplex channels for the Advanced Mobile Phone System (AMPS). All these systems were analogue with a relatively simple Frequency Modulation (FM) technology. The efficiency of the usage of frequency resource was low, and the capacity was not enough to deal with many calls and deliver advanced services.



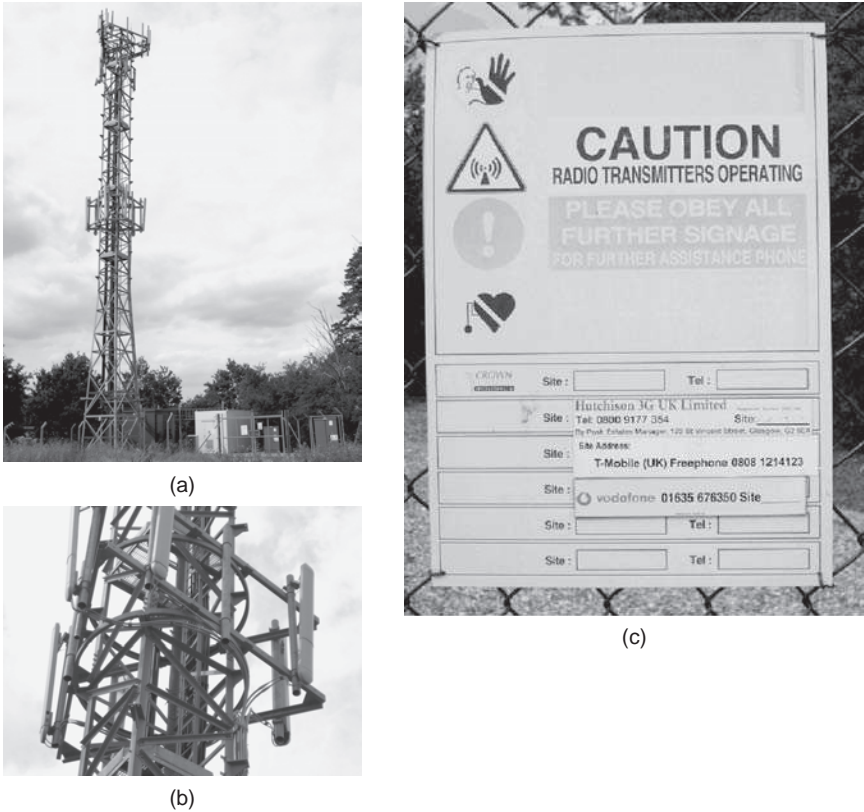
**Figure 2.4** Plaque at Marconi Beach, Cape Cod, USA, commemorating the first transatlantic wireless transmission in 1903 (photograph by the authors).

In the late 1980s, the first US Digital Cellular (USDC) system hardware was installed in major US cities. A cellular system based on CDMA (discussed below) was developed by Qualcomm Inc. and standardized by the Telecommunications Industry Association (TIA) as an Interim Standard. In early 1995, new Personal Communication Service (PCS) licenses in the 1800/1900 MHz band range were auctioned by the US government to wireless providers. In Europe, there was also the development of wireless systems from analogue (such as E-TACS, see below) to digital (such as Global System for Mobile communications – GSM). The analogue systems have been commonly referred to as the first generation (1G) of wireless systems whilst the second generation (2G) are often referred to as the earliest group of digital systems. There is an interim set of technologies referred to as 2.5G before the full implementation of the third generation (3G) systems. All these are discussed later in this section.

### *Basic Concepts of Wireless Mobile Telecommunication Networks*

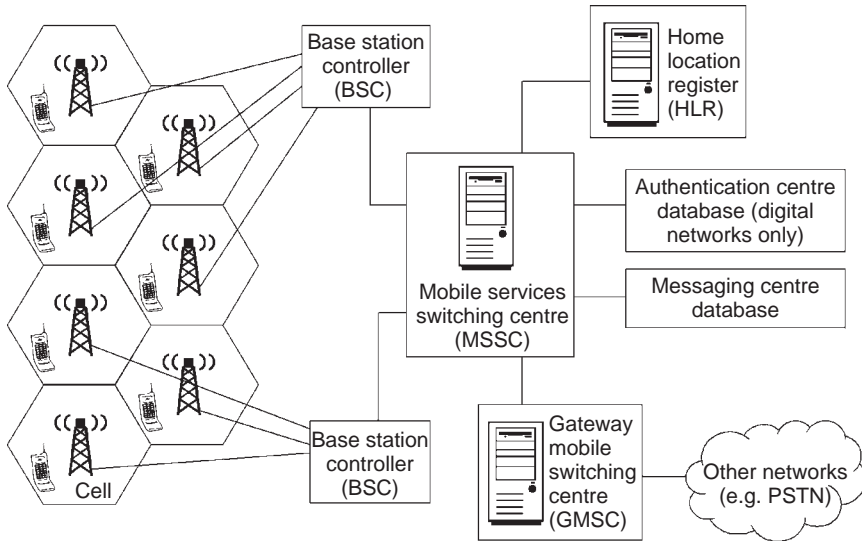
Wireless mobile telecommunication networks are known in the USA as cellular networks. Here the term wireless mobile telecommunication network is used. Shortened terms such as wireless mobile network or just mobile network are also used interchangeably. The basic concept of a wireless mobile network is a network of cells. Within each cell, there is a Base Transceiver Station (BTS) containing equipment for radio transmission and reception. A BTS provides the communication with mobile phones currently within its cell. A BTS is also known as a Base Station (BS); it consists of antennas, amplifier, receiver, transmitter, hardware and software for signal communication (both sending and receiving). A photograph of a BS is shown in Figure 2.5(a) along with a photograph of BS antennas in Figure 2.5(b) and a notice board showing the co-usage of the BS tower in Figure 2.5(c). The coverage of a cell depends on a number of factors, such as the transmitter power of both the BS and mobile phones, the height of antennas on the BS and the topography of the area.

The basic principle of a wireless mobile network is illustrated in Figure 2.6. Mobile phones in each cell communicate through the cell's BS. A number of such BSs are connected to a Base Station Controller (BSC) that has logic software to manage these BSs and manage call hand-off between cells. Each BSC is connected to a Mobile Service Switching Centre (MSSC) that manages the mobile subscribers' calls,



**Figure 2.5** A base transceiver station (BTS), also known as base station (BS): (a) base station; (b) directional antennas; (c) notice indicating that the tower is co-used by several providers (photographs by the authors).

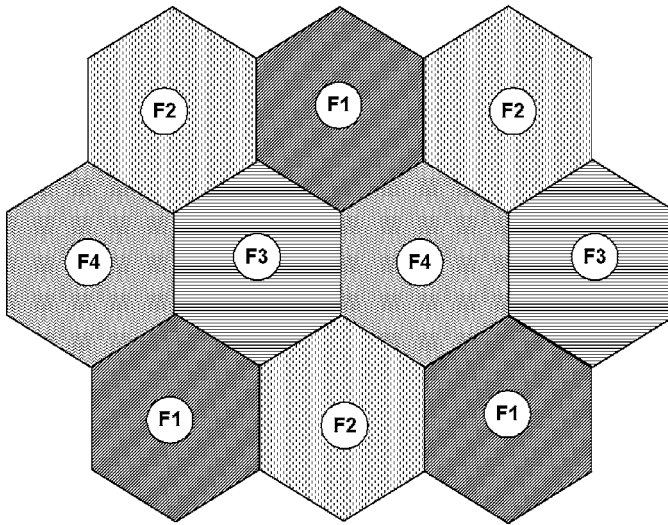
such as routing calls in their cluster of cells and giving instructions to BSs. There can be a number of MSSCs in a mobile network. MSSCs also connect to a number of databases that assist in the running of networks. One of these databases is the subscriber database, called the Home Location Register (HLR), aimed at dealing with the mobility of network subscribers. There can be one or more HLRs, which also store geographic tracking data as subscribers move around the network coverage. Other databases connected with MSSCs are an authentication database (only for digital networks) which authenticates mobile subscribers, and a messaging centre database that routes Short Message Service (SMS) messages to mobile phones. There is one Gateway



**Figure 2.6** Basic principle of a mobile network.

Mobile Switching Centre (GMSC) in each mobile network. MSSCs are connected to this GMSC in order to route calls to other networks. MSSCs can thus be connected with other mobile networks, the Internet or conventional telephone networks such as the Public Switched Telephone Network (PSTN).

One of the fundamental elements of wireless mobile networks is the concept and implementation of frequency re-use, which is also known as the cellular concept. The principle of frequency re-use is that a number of frequencies (channels) are allocated to a cell and the same frequencies are also re-used in other far away cells in the system without interference. This increases the capacity for each geographic area given a limited spectrum allocation without the necessity for major technological changes. The concept of frequency re-use is regarded as a key means of overcoming congestion within a network. The basic concept of frequency re-use is to install a number of small cells with low power transmitters to replace a large cell with high power transmitter for a service coverage area. The BS of each small cell is allocated a portion of available channels in the whole network. The BSs of adjacent cells are assigned different channels. Thus, the interference between adjacent BSs and their mobile users is minimized. The other channels available in a network are allocated to neighbouring BSs; this



**Figure 2.7** The principle of frequency re-use – cellular concept (see text for explanation).

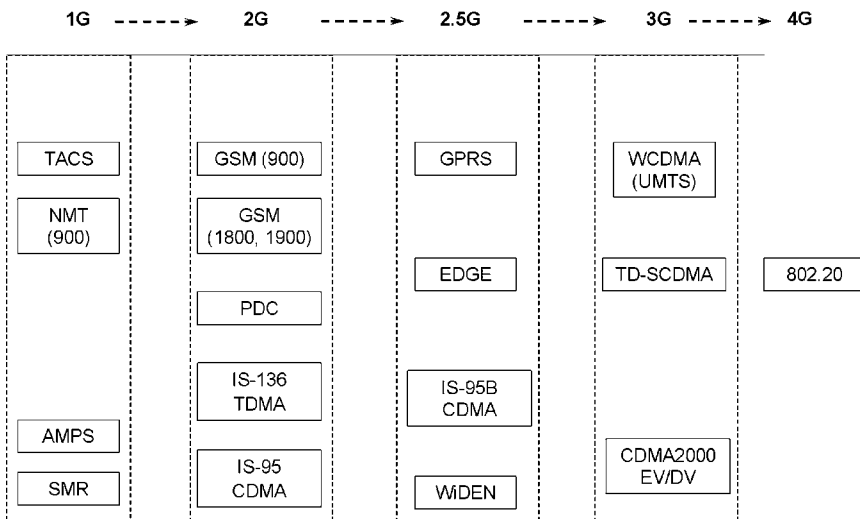
principle is illustrated in Figure 2.7. BSs and their channels can be systematically located, and the available channels can be distributed over geographic regions. Therefore, available channels can be re-used many times, providing the interference level between BSs having the same channels is sufficiently low. This frequency re-use concept increases the capability of total call handling whilst conserving the spectrum.

A mobile network is usually represented as adjoining hexagons. These hexagons are the approximation of the circular coverage areas of BSs. A network with hexagon representation can show a coverage area without overlaps or gaps. Each hexagon is referred to as a *cell* and is given a unique identifier. The identity of a cell is known as Cell-ID. When a mobile device is within range of a BS, it will be registered to the cell covered by that BS. With the concept of frequency re-use just discussed, each cell uses different radio frequencies to reduce possible interference with adjacent cells. A large geographical service area is divided into many cells, and the same frequencies are assigned to multiple non-adjacent cells (Figure 2.7). When a subscriber (mobile phone user) moves from one area to another, the registration and any calls in progress are handed over from one cell to another without interruption.

### *The Development of Wireless Mobile Networks: From 1G to 4G*

Wireless mobile telecommunication networks started to be developed in 1970s. For over three decades, wireless mobile networks have been developed from the first generation (1G) analogue networks, to the second generation (2G) digital networks, to current 2.5G and third generation (3G) networks that offer higher data transmission rates and the capability of both voice and data communication. Implementation of fourth generation (4G) networks is already on the way. In this section, the basic principles, functionality and technologies of these networks are described. A summary of the main technologies that will be described is given in Figure 2.8.

**First Generation (1G)** wireless mobile networks were analogue and circuit switched. They operated in the 800/960 MHz frequency band, with data transfer rates up to 9.6 Kbps. The main technologies used were the Frequency Modulation (FM) analogue modulation technique and the Frequency Division Multiple Access (FDMA) technology. The use of the term analogue originates from the principle that the modulation of a carrier wave is analogous to the fluctuations of the voice itself. First generation networks included the Advanced Mobile Phone System (AMPS) developed in North America, the



**Figure 2.8** The development of wireless mobile telecommunications from 1G to 4G.



Total-Access Communications System (TACS) developed in the United Kingdom, Nordisk Mobile Telefoni (NMT, Nordic Mobile Telephony) in North Europe, the Extended-TACS (E-TACS) and the Japanese-TACS (J-TACS). AMPS was proposed by AT&T in 1971 and tested in 1978, and wireless systems based on the AMPS standard were installed throughout North America in the early 1980s. Originally the AMPS operated in the 800 MHz frequency band using 30 KHz wide channels (Muller, 2003). The TACS was deployed in the mid 1980s operating in the 890–960 MHz frequency range using 25 KHz bandwidths. Although 1G networks only carried voice traffic, they were seen as a great advance and success in mobile communication. The limitation of analogue technology is that its signal could be degraded by a number of factors, including terrain, weather and volume of traffic. The analogue signal could also be intercepted easily. The 1G mobile phones were less secure and it was easy to have interference where the signal was weak. In addition, the transmission of data doesn't perform well across analogue networks.

**The second generation (2G)** networks were first introduced in the early 1990s, developed from the 1G analogue networks. Second generation networks are digital. Digital modulation formats and multiple access techniques (TDMA and CDMA, see below for details) are used in 2G networks. The main benefits of 2G networks over 1G networks are: increased capacity resulting from at least a threefold increase in spectrum efficiency; reduced capital infrastructure costs and per subscriber cost; reduced mobile fraud through better security and different levels of encryption. The primary service of 2G networks is still voice communication. However, with digital techniques, 2G networks support high bit rate voice, limited data communications and applications such as Short Message Services (SMS), Multimedia Message Services (MMS) and games.

The data transfer rates for 2G networks range from 9.6 to 28.8 Kbps, with the most usual rate at 14.4 Kbps. The modulation techniques used to produce digital signals are more complicated than those used in analogue signals. GMSK (Gaussian Minimum Shift Keying) is one such technique to transmit digital wireless signals with high power efficiency (Sampei, 2002). In wireless mobile networks, there is limited frequency spectrum resource so that there are a limited number of channels in a given area for use. Multiple access and duplexing techniques have been deployed in order to maximize the usage of the frequency spectrum resource. In 2G networks, Code Division Multiple

Access (CDMA) and Time Division Multiple Access (TDMA) are the main access techniques used; both are techniques that allow multiple user access to a BS.

Code Division Multiple Access (CDMA) is based on spread-spectrum technology. In spread spectrum technology, the carrier waves used are designed to have a much wider bandwidth than point-to-point communication requires at the same data speed, thus providing the potential for a number of users to share a bandwidth simultaneously. By using CDMA all users share the same carrier frequency and may transmit simultaneously. Users are assigned their own pseudorandom codeword or spreading code and these are used in a time correlation to separate out the calls made by different users. At the transmitter, the spreading codes are used to process each conversation (between two matched users) and distribute the signals over the available bandwidth. For any one conversation, the other simultaneous conversations are reduced to background noise. Such frequency sharing can increase the capacity by 8- to 15-fold over analogue networks. Two technological developments made the usage of CDMA in commercial applications possible. One was the development of high density digital integrated circuits at low cost, which resulted in small size, lightweight mobile phones at relatively low cost. The other was the ability of mobile phones to regulate their transmitter power to optimize signal quality during multiple access communication.

The Time Division Multiple Access (TDMA) technique divides the radio spectrum into time slots to increase the capacity. Within each time slot, only one user at a time is permitted to transmit or receive. No other conversations can access this slot until the user's call is finished or until that original call is handed off to a different slot by the network. In the case of TDMA used in one of the 2G networks (Global System for Mobile communications – GSM), the 200 KHz spectrum is divided into eight time slots (also known as channels). Users are assigned their own time slot into which voice or data are inserted for transmission via synchronized timed bursts. The bursts are reassembled at the receiving end and appear to provide a continuous, smooth communication because the process is very fast. The networks using the TDMA technique transmit data in a non-continuous way for any user. The technique of dynamic time slot allocation can be used to improve TDMA, by preventing wasting unused bandwidth caused by silences in conversations. The use of dynamic allocation can increase bandwidth efficiency a further twofold. Different versions of the TDMA technique are used



in commercial digital mobile telecommunication networks such as Digital American Mobile Phone Service (D-AMPS) and GSM.

There are several CDMA-based and TDMA-based 2G networks. Three popular TDMA-based networks are: GSM, which is widely used in Europe and Asia; Interim Standard 136 (IS-136), which is popular in North America, South America and Australia; and Pacific Digital Cellular (PDC) that is a Japanese standard for TDMA (Rappaport, 2002). One CDMA-based network, Interim Standard 95 (IS-95, also known as cdmaOne), is popular in Korea, Japan, China, South America and Australia. The general principles of GSM networks are now examined as a way of further exploring 2G networks.

The Global System for Mobile communications (GSM) was firstly developed in Europe in 1982 and entered commercial service in 1992, with the intention that users should be able to use their mobile services across national boundaries. The GSM is circuit-switched and provides voice and data services with a data rate speed up to 14.4 Kbps. The GSM has a bandwidth of 200 KHz and has been deployed at several frequencies: 900, 1800 and 1900 MHz bands, resulting in networks being named GSM(900), GSM(1800) and GSM(1900). A GSM network usually has the following components: Mobile Station (MS, usually known as mobile phones), Base Station Subsystem (BSS) and Operations and Support System (OSS). An MS consists of a handset (mobile phone) and a Subscriber Identity Module (often known as a SIM card). A SIM card provides personal mobility, including information such as the identity of a subscriber, subscriber authentication and service information, so that a subscriber can access services through a handset into which the SIM card has been inserted regardless the handset's location. BSS includes both Base Transceiver Stations (BTS) and a Base Station Controller (BSC). A BTS contains radio transceivers which define a cell and provide the radio interface with handsets. A BSC is connected with one or a number of BTSs and also provides the connection to a Mobile Service Switching Centre (MSSC). The functions of BSCs include radio resource management, mobility management for subscribers and some operational functions for the overall network. An OSS provides the functionality needed to manage mobile subscribers such as registration, authentication, location updating, hand-offs and call routing to roaming subscribers. The basic components in an OSS include an MSSC, Visitor Location Registration (VLR) and Home Location Register (HLR). An MSSC is the 'switch' node which provides the functionality of controlling call setup, call routing as well as functions provided by a standard

telecommunication switch. VLR is a database containing roaming subscribers' information where they are in the coverage area of an MSSC. The HLR is a database with subscribers' data such as the details of user subscribed services.

There are two other popular TDMA-based networks. One is the digital version of the AMPS (D-AMPS), now known as IS-136, which evolved from the 1G network AMPS. IS-136 provided 10–15 times more channel capacity than the AMPS and supported new feature services such as data communication, voice mail, call waiting/diversion, voice encryption and calling line identification. IS-136 could be added onto the existing AMPS infrastructure. IS-136 could also be implemented with CDMA to increase channel capacity by up to 20 times and provided a comparable range of services. However, it required an entirely new network infrastructure if implemented with CDMA. Another network is the Pacific Digital Cellular (PDC), which was widely used in Japan. It operated at 800 and 1500 MHz frequencies. IS-136 and PDC standards were abandoned by several major carriers, such as AT&T Wireless and Cingular in the United States and NTT in Japan in 2001, in order to proceed to 3G standards based on the GSM TDMA platform. Most carriers throughout the world have been adopting a 3G standard based on either the GSM or CDMA. Therefore, there are two main 3G wireless technologies: one based on the GSM and another on CDMA.

**The 2.5G** networks evolved from 2G networks and are regarded as an interim step from 2G towards 3G networks. The standards representing 2.5G technology allow existing 2G equipment and software to be modified and upgraded to support higher data transmission rates. 2.5G technology has a packet-switched extension developed for 2G mobile radio standards, which means that data are sent in small portions (packets), and each packet is sent completely independent of each other. Packet-switched data are primarily used to offer the ability of high speed Internet access from mobile handheld devices. The packet-switched data support for 2G GSM is the 2.5G General Packet Radio Service (GPRS). 2.5G networks can have data transmission rates up to 384 Kbps and can have always-on capability. Furthermore, 2.5G uses the IP to provide fast access to data networks. 2.5G technologies also support Wireless Application Protocol (WAP), which enables the use of the Internet via a mobile device and allows Web pages to be designed in a compressed format purposely for mobile devices with small screens. WAP is discussed in more details in Section 2.5.2.

## Location-Based Services and Geo-Information Engineering

Depending on the applications, 2.5G/3G mobile phones can be used on 2.5G networks and support WAP, SMS, MMS, mobile games and search directories. There are two main 2.5G networks, which will be introduced here: GPRS and Enhanced Data Rates for GSM Evolution (EDGE). GPRS and EDGE provide a 2.5G migration path to the next generation 3G wireless mobile networks.

GPRS is an enhancement to GSM type networks and allows users to access the mobile Internet and local networks from mobile phones and other handheld devices. GPRS, an important step in the evolution to 3G mobile networks, is a packet-switching technology and has 'always on' data connection. It combines TDMA time slots to provide high data transfer speeds. The speed in theory can be up to 171 Kbps but in practice the speed achieved is much lower at 40 or 53 Kbps (Smith and Meyer, 2004). Combining TDMA time slots in GPRS means that a mobile device can have access to more than one time slot, although GPRS has the same basic radio interface as GSM. Furthermore, GPRS uses multiple channel coding schemes that are different from GSM. GPRS also has a lower cost than circuit-switched services, because GPRS uses shared communication channels. The advantage of the GPRS is in conserving radio resources and reducing reliance on traditional circuit-switched network elements. It also allows the sharing of the same radio resources among all mobile devices in a cell to relieve capacity impacts. The main benefit of GPRS is the integration of higher throughput packet data and wireless mobile networks, which supports advanced data services more efficiently and enables mobile Internet applications. In order to carry out such advanced services and applications, a GPRS-enabled mobile device is needed. Moreover, the middleware required for adapting applications to the slower speed networks is not needed, which makes applications available to users less complicated. The improved quality of data services, in terms of reliability, response time and other features, makes applications more appealing to mobile users.

EDGE is a more advanced upgrade to GSM. EDGE technology builds on GPRS. It aims to enhance the data throughput capabilities of a GSM/GPRS network. EDGE uses the 200 KHz channel and an eight time slot structure that is the same as used by GSM and GPRS. However, 8 Phase Shift Keying (8 PSK) modulation is introduced for EDGE in addition to the 0.3 Gaussian Minimum Shift Keying (GMSK) used in GSM to generate digital signals. EDGE offers higher bandwidth efficiency to

allow more user data in the same 200 KHz channel. It increases data transfer speed by using existing TDMA networks. There are no further requirements for new spectrum and for major changes to the networks, although the addition of new software and hardware is needed for existing BSs. EDGE supports data transmission rates up to 384 Kbps for a single dedicated user on a single GSM channel (Rappaport, 2002). The typical services, which can be provided through EDGE networks, include multimedia messaging, Web browsing, enhanced short messages, wireless imaging with instant pictures, video services, document and information sharing, surveillance, voice over the Internet and broadcasting.

IS-95B is another interim solution to evolve networks from 2G to 3G. IS-95B is a CDMA 2.5G solution, which is based on IS-95 (cdmaOne) 2G networks. It provides high speed packet and circuit-switched data access on a common CDMA radio channel.

**The third generation (3G)** of networks provides high speed data transmissions of at least 144 Kbps and further up to 2 Mbps, which can support fast mobile Internet access, multimedia applications such as full-motion video, and video conferencing over mobile communication networks. Under the 3G concept, it is aimed to achieve voice and data convergence with seamless and global radio coverage by integrating a number of wireless technologies at a higher level. This could lead to expanded coverage and new service capabilities. The third generation is expected to include satellite and terrestrial systems for both fixed and mobile users. The 3G initiative aims to provide a set of specifications to achieve the interoperability between all national and regional networks and products. Therefore, the incompatibility between regional 2G networks can be eliminated in the evolution from 2G to 3G. Third generation networks then will provide mobile users with greater interoperability.

However, currently there is no single unified standard for 3G networks, and multiple standards have been established. For TDMA-based networks (such as GSM, IS-136 and PDC), deploying GPRS is the first step to true 3G services. As just discussed, the GPRS enhancement to GSM can support data transmissions with high speed. The upgraded networks with EDGE-compliant software can boost data transmission rates to as much as 384 Kbps and provide more 3G-type services. For these TDMA-based networks, the 3G evolution is to move to the Universal Mobile Telecommunication System (UMTS), also known as Wideband CDMA (W-CDMA). UMTS is

based on the network essentials of GSM and the merged version of GSM and IS-136 through EDGE. For CDMA-based networks, it takes a different technology path to 3G, from IS-95 (cdmaOne) to CDMA2000. Furthermore, there is another 3G standard developed in China: Time Division-Synchronous Code Division Multiple Access (TD-SCDMA). In general, the main 3G networks are: UMTS, CDMA2000 and TD-SCDMA.

UMTS (W-CDMA) has been developed as a standard for 3G wireless telecommunication and was submitted by the European Telecommunications Standards Institute (ETSI) to the International Telecommunication Union (ITU)'s International Mobile Telecommunications 2000 (IMT-2000) to be considered as a world standard. UMTS makes use of UMTS Terrestrial Radio Access Network (UTRAN) as the basis for a global terrestrial radio access network. UMTS is a 3G standard for the evolution of GSM (2G) to 3G in Europe, Japan and the United States. It is also a principal alternative being discussed in Asia. It provides a high capacity upgrade path for GSM, whilst assuring backward compatibility with the TDMA-based 2G and 2.5G networks. UMTS uses one 5 MHz channel for both voice and data. Its data transmission rate capability is at least 144 Kbps for applications with full mobility in all environments, and reaches 384 Kbps for applications with limited mobility in the macro- and micro-cellular environments. The data transmission rate can be up to 2.048 Mbps for applications with low mobility or if a user is stationary (Muller, 2003). UMTS supports high speed multimedia services over mobile networks, such as high quality data services, full-motion video, mobile Internet access and broadcast-type services. Users are able to be connected to these services anytime, anywhere from their wireless mobile devices. UMTS has been designed for 'always-on' packet-based wireless services. Therefore, it could support various devices sharing the same wireless mobile network and be connected to mobile Internet regardless of location. Furthermore, UMTS has been developed for both wide area mobile cellular coverage and indoor cordless applications. Although UMTS has backward compatibility for all 2G and 2.5G equipment and applications, it does require a change of the RF equipment at each BS due to a wider radio bandwidth.

CDMA2000 is a different technology path to 3G for the CDMA-based 2G and 2.5G networks. The CDMA2000 standard, recognized by the ITU in 2001 (Rappaport, 2002), offers mobile networks the capability of mobile Internet access with high data transmission rates. The evolution to 3G for CDMA-based 2G and 2.5G

networks is gradual within existing systems. It maintains backward compatibility with existing IS-95 (cdmaOne) and IS-95B equipment. CDMA2000 supports an instantaneous data transmission rates up to 307 Kbps in packet mode, and typical data transmission rates up to 144 Kbps in mobility (depending on the velocity of users), the number of users and the propagation conditions. Where channels are dedicated to data users, the data transmission rate can be up to 2.4 Mbps. The data speed achieved is sufficient for Web browsing, e-mail access and mobile Internet applications. The technologies in CDMA2000 allow the same spectrum, bandwidth, framework and RF equipment to remain in use at each BS when the 3G upgrades are introduced. CDMA2000 is regarded as providing a seamless and less expensive upgrade path for CMDA-based networks.

TD-SCDMA has been developed in China and was submitted to IMT-2000 as a 3G standard proposal in 1998 and adopted by ITU as one of the 3G options in late 1999. TD-SCDMA is based on the existing core GSM infrastructure, adopting smart antennas, spatial filtering and joint detection techniques to gain more spectrum efficiency. It provides an evolution path to 3G through adding high data transmission rate equipment at each GSM BS.

To recap, the data transmission speed of 2.5/3G networks is determined and influenced by a number of elements, including the equipment and software in a wireless mobile network, the speed at which users are moving and the distance of users from the nearest BS. Therefore, the stated data transmission speed of 2.5/3G services is sometimes not reached in the reality of the operating environment.

**The fourth generation (4G) networks.** Although 3G aims to develop a unique standard, in the end multiple standards are being implemented, as discussed above. With the expansion of people's needs for information and services it is already clear that 3G cannot satisfy all the demands that will be made of it. 4G is viewed as the next generation of wireless communications beyond 3G. The noticeable trends in the networks beyond 3G will be high data transmission speed, high mobility and seamless roaming ('anywhere' and 'anytime'). Fourth generation will be a fully IP-based integrated network, which aims to integrate local area network techniques into wide area networks, and to combine both wired and wireless technologies. Fourth generation will provide users with different types of services with higher data transmission speed, larger capacity and high security. Fourth generation is expected to have data transmission rates of 100 Mbps to 1 Gbps.



## Location-Based Services and Geo-Information Engineering

There is not yet a unified definition on 4G. Currently, a number of major features and objectives are recognized as being characteristics of 4G. They include:

- high network capacity and high data transmission rate (100 Mbps to 1 Gbps);
- seamless roaming with smooth hand-off across different networks;
- the ability to provide high quality multimedia services;
- the ability to integrate wired and wireless communication networks and be interoperable with existing wireless networks;
- self-adaptive resource allocation with strong self-organization and high flexibility;
- a comprehensive IP, packet-switched and spectrum efficient network;
- the development of broadband-based concepts.

The main technologies for 4G include Orthogonal Frequency Division Multiplexing (OFDM), Multiple-Input-Multiple-Output (MIMO) technology, adaptive radio interface, fixed relay networks (FRNs) multi-mode protocol for the cooperative relaying concept. In research and laboratory experiments, 1 Gbps packet transmission speeds using 4G mobile communication equipment have been achieved. A down-link speed of 1 Gbps was achieved in a laboratory experiment by NTT DoCoMo (Japan's biggest mobile communications carrier). Technology of Variable-Spreading-Factor Spread OFDM (VSF-Spread OFDM) was used to facilitate downlink connections at a very high speed, and MIMO technology was used to increase wireless bandwidth and range by sending information over multiple paths in order to carry more information. Furthermore, four antennas were employed, with each transmitting 250 Mbps streams of data to increase the amount of data throughput. There has been the development of networks based on the 4G concept in Europe and United States. The operation of 4G is expected to be after 2010.

The Institute of Electrical and Electronic Engineers (IEEE) specification IEEE 802.20 is the standard on which 4G technology will be based. The IEEE 802.20 specification is also known as Mobile Broadband Wireless Access (MBWA). A MBWA working group was established in December 2002 to develop the specification for truly worldwide interoperable multi-vendor mobile broadband wireless networks that are always on, low cost and ubiquitous. The IEEE 802.20

specification is designed to operate in licensed bands below 3.5 GHz, use packet architecture and offer more ubiquitous services with mobile IP for interoperable mobile broadband wireless networks. It is designed to provide mobile data to a Wide Area Network (WAN) which can be integrated with wireless mobile networks. The IEEE 802.20 specification aims at IP roaming and hand-off at a data transmission rate of 1 Mbps or more, optimized for IP data transport (full mobility up to speeds of 250 km/h) and low latency. Such mobile broadband wireless network is also known as Mobile-Fi.

### **2.3.2 Other Wireless Networks**

In this section, different types of wireless networks are described. Two of them are: Wireless Local Area Networks (WLANs) and Bluetooth networks. These networks generally do not require a license for their use of spectrum and provide ad hoc high data transmission rate connections deployed by individuals. In the late 1980s, the FCC provided license-free bands for low power spread spectrum devices in the Industrial Scientific and Medical (ISM) bands of 902–928 MHz, 2.4–2.483 GHz and 5.725–5.825 GHz under FCC regulations Part 15. The purpose of these license-free allocations is to encourage the development of individual wireless local area networks and other short range, low power devices for data communication in workplaces.

#### **2.3.2.1 Wireless Local Area Networks (WLANs)**

WLANs are the wireless networks used as a replacement for wires within buildings, such as in office settings and at home. They have become popular and been fast growing from the late 1990s due to the use of the Internet along with portable laptop computers. WLANs originally aimed to connect computers (including laptops and PDAs) within a hundred to a few hundreds metres with speeds of 11–54 Mbps. By using WLANs, users can share data and applications, access peripheral equipment (e.g. printers) wirelessly. Users can also carry out other tasks, such as accessing the Internet and exchanging e-mails, in the same way as they do on LANs.

A universal wireless networking standard known as Wireless Fidelity, WiFi, is used in wireless local area networks. WiFi, instead of being a single standard, refers to a set of standards based on the 802.11 networking protocol: Institute of Electrical and Electronic Engineers



(IEEE) 802.11. The IEEE 802.11 networking protocol has three standards: 802.11a; 802.11b; and 802.11g (Gralla, 2006). It became an interoperability standard for WLAN manufacturers in 1997, providing 2 Mbps user data transmission rates (1 Mbps in noisy conditions). In 1999, IEEE 802.11b was standardized, using direct sequence spread-spectrum (DSSS) technology. IEEE 802.11b standard designated devices can provide user data transmission rates of 5.5 Mbps (11 Mbps in theory) in the 2.4 GHz band within a 100 metre range. The IEEE 802.11b standard implemented with some proprietary extensions can offer applications with higher data transmission rates of up to 22 Mbps. Another standard, IEEE 802.11a, works by transmitting high speed digital data over a radio wave, using Orthogonal Frequency Division Multiplexing (OFDM) technology. IEEE 802.11a standard designated devices can provide user data transmission rates of up to 30 Mbps (54 Mbps in theory) in the 5 GHz band within a range of 300 metres. The IEEE 802.11g standard is developed in both the 2.4 and 5 GHz bands and supports roaming capabilities and dual-band public WLANs. Both IEEE 802.11b and IEEE 802.11g can work in the 2.4 GHz spectrum.

The main advantage of WiFi is the ability to establish a relatively simple wireless communication network as an extension of the wired network. Such a wireless network can be implemented using an existing network infrastructure with minimal changes to the system. Users with WiFi-enabled mobile devices can keep connected to the network while roaming throughout the areas where there are available Access Points (APs). These APs are located strategically and are physically connected into the wired network. Users can also run the same network applications as if they are on an Ethernet LAN (Muller, 2003).

The High Performance Radio Local Area Network (HIPERLAN) standard was developed in the mid 1990s in Europe, providing a similar capability to IEEE 802.11. HIPERLAN uses the 5.2 GHz and the 17.1 GHz bands and provides asynchronous user data transmission rates between 1 and 20 Mbps. It can operate in vehicles travelling at speeds of 35 km/hour with up to 20 Mbps throughput within a range of 50 m. HIPERLAN is designed to provide individual WLANs for computer communication. HIPERLAN/2 is the next generation of the European WLAN standard. It operates in the 5 GHz band and can provide user data transmission rates up to 54 Mbps. With the increasing wireless data transmission rates and convergence of worldwide standards, new applications for WLANs

have been emerging. However, because of the unlicensed spectrum used by WLAN, there could be saturated WLAN deployment in certain areas without appropriate management in frequency planning, radio engineering and installation of WLAN.

An Access Point (AP), also known as router, is a key component in a WiFi network. It acts as a BS or bridge between wireless local area networks and a wired network such as Ethernet or the Internet. An AP has a radio transmitter, a receiver and an interface to other networks (i.e. Ethernet, or the Internet). APs can be connected with a corporate local area network (LAN), usually an existing Ethernet-based LAN. Various types of WiFi-enabled devices in the communication range can then be connected to the LAN wirelessly via these APs. APs can also connect to a modem, and then enable WiFi-enabled devices within range to access the Internet wirelessly. WiFi-enabled devices are those installed with a WiFi adapter able to communicate with APs in a WiFi network. In a WiFi network, WiFi-enabled devices search for an AP by sending out probe request frames and receiving a probe response from a nearby AP. If there is more than one AP in the range, an AP is selected based on its signal strength and low error rate. When the link is established, data in packets can be transmitted between the device and the AP. One AP can communicate with a number of devices. There are many APs in a WiFi network and devices can move from one AP to another. In addition, the WiFi standard allows WiFi-enabled devices to communicate with each other directly, which establishes what is known as a peer-to-peer network.

Another type of WiFi network, usually created in a public place, provides people with the ability to connect to the Internet wirelessly while on the move. Such WiFi networks are often known as 'Hot Spots'. Hot Spots can be viewed as WiFi networks open to the public. There are thousands of such Hot Spots located in airports, coffee shops, restaurants, hotels and a range of public places.

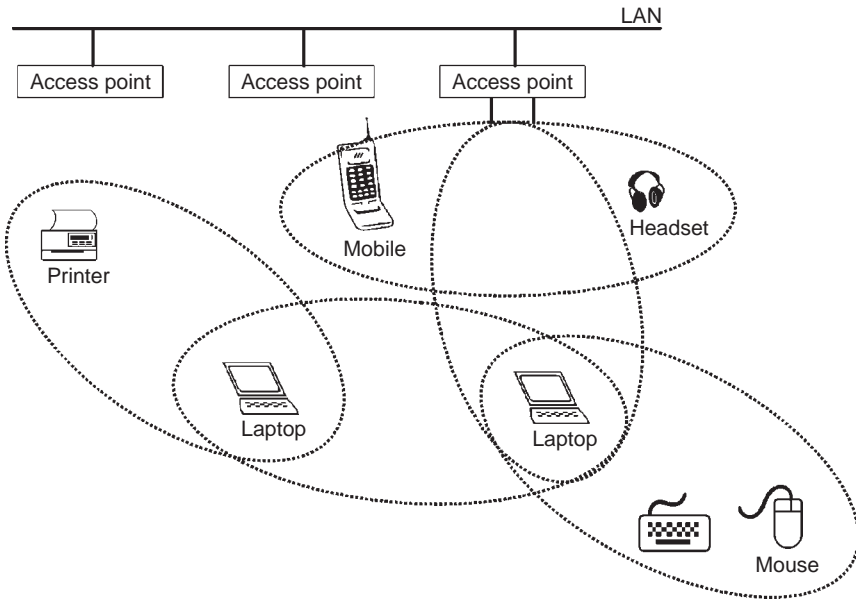
### **2.3.2.2 Bluetooth Technology**

Bluetooth technology provides an ad hoc approach for a low power, short range wireless connection for voice and data transmission between various devices within a nominal 10 metre range. Bluetooth technology was initiated by Ericsson Mobile Communications in 1994, intending to investigate the feasibility of a low power, low cost radio interface between a variety of fixed and portable appliances without cables (Muller, 2003). The aim of Bluetooth is to unify the connectivity

of various devices within the personal workspace of individuals. Bluetooth was named after King Harald Blåtand II (Bluetooth in English) who united Denmark and Norway in the tenth century. By using Bluetooth technology, communication can be established ad hoc and instantaneously between a wide variety of fixed and mobile devices such as mobile phones, PDAs, laptops, desktop computers as well as peripheral devices which are Bluetooth-enabled. Within the communication range, the connection between Bluetooth devices can be made easily without specific set-up, and such a connection is running in the background and always on. In addition, the connected devices can be in motion.

Bluetooth is an open standard, which operates in the globally available unlicensed 2.4 GHz ISM radio frequency band. Hence, Bluetooth-enabled devices can be used anywhere in the world. However, different countries have allocated various channels for Bluetooth operation. The 2.4 GHz ISM band is available for Bluetooth use in the United States and most European countries. The data transmission rate of Bluetooth is 30–400 Kbps and 1 Mbps at aggregated level. The transmission power is 1 milliwatt. Bluetooth technology is incorporated into mobile devices by either installing tiny, inexpensive, short range transceivers directly into existing component boards or by inserting an adapter device (such as a PC card) into a laptop computer. Bluetooth-enabled devices are potentially the least expensive wireless technology to implement. Up to eight devices, one master and seven slaves, can communicate with one another in a so-called piconet at distances of up to 10 meters. When Bluetooth devices are in the communication range of each other, a service discovery procedure takes place. This procedure involves the exchange of messages for awareness of each other's service and feature capabilities. A connection can then be established between two or more Bluetooth devices. The protocol used by Bluetooth is a combination of circuit and packet-switching, which enables voice and data communication. By use of a technique called 'frequency hopping', Bluetooth can achieve high security.

The advantages of Bluetooth technology are that it is convenient, reliable, easy to use as well as low cost. In addition, it doesn't require a clear line of sight between devices. The connection doesn't need to be set up, is run in the background and is always on. Bluetooth-enabled devices can communicate with each other while within range without the need to open an application. The Bluetooth specification is widely used in wireless communication and networking between PCs,



**Figure 2.9** An illustration of a PAN using Bluetooth and WiFi connections.

mobile phones and other mobile devices. Compared with the WiFi described above, WiFi networks need more set up, whilst Bluetooth technology is more suitable for ad hoc networking of mobile devices but with a lower speed for transferring data and shorter ranges. Along with WiFi, Bluetooth supports what is called Personal Area Networks (PANs) that interconnect pocket PCs, PDAs, mobile phones and other appliances (Figure 2.9). PANs can enable the collaborative communication between users, their appliances and their environment.

### **2.3.2.3 Worldwide Interoperability for Microwave Access (WiMax)**

WiMax is a broadband wireless standard. WiMax networks can be used over much greater distances than WiFi. They can cover a metropolitan area, as much as tens of square kilometres, depending on the number of users. Through WiMax networks, wireless Internet services are made available by Wireless Internet Service Providers (WISPs) at high speed – much faster than WiFi. In these networks, WiMax towers, which can broadcast about 30 miles, are connected to WISPs through a high speed wired connection. A high speed connection can be also

established through a line of sight link between a WiMax tower and a WISP, and between WiMax towers themselves. Thus, a WiMax network can have a large coverage. The wireless connection between a WiMax tower and WiMax users can be implemented either through line-of-sight transmission directly, or through a non-line-of-sight connection such as WiFi. The line-of-sight connection has higher speed and is more stable. WiMax technology is starting to be built into laptops and other mobile devices.

The IEEE 802.16 standard defines the specification for the wireless metropolitan area network, known as Wireless MAN (WMAN). Nonmobile WiMax using the IEEE 802.16d standard has been around for a while but mobile WiMax has taken longer, due to the need for countries to assign licences and telecommunications companies to build infrastructure. IEEE 802.16e is a new radio access method for wireless broadband, and for providing IP mobility access in the licensed frequency bands of 2–6 GHz. Thus, IEEE 802.16e is designed for WMAN whilst IEEE 802.11 (WiFi) is only for local area networks. Although there is an overlap between IEEE 802.16e and IEEE 802.20 (discussed in Section 2.3.1), IEEE 802.16e is different from IEEE 802.20. IEEE 802.16e is designed as an extension of IEEE 802.16a network, and to provide a bridge between wireless local area networks (WLANs) and Wide Area Networks (WANs).

### 2.3.3 Wireless Mobile Devices

Wireless mobile devices can include mobile phones, PDAs with wireless connection, smart phones and other handheld devices such as tablet PCs. With the development of technologies in the areas of communication, hardware and software, more and more features are likely to be integrated into a single wireless mobile device, even though there will always be devices with specific functionality and purpose. Such combined feature devices can be used for making mobile phone calls, taking messages, being on-line with mobile Internet, as well as for personal information management, enterprise application and LBS applications.

The mobile phone (cellular phone) is currently one of the most widely used mobile devices. The number of mobile phone subscribers worldwide rose from 23 500 in 1980 to over 2.2 billion in 2005 (Comer and Wickle, 2008). This compares to 1.3 billion subscribers to landlines in 2005. The current five-year growth rate for mobile phone

subscribers is 298%. The development of mobile phone usage has been part of the development of mobile wireless telecommunication networks (cellular networks) described in Section 2.3.1. Mobile phones started in the 1970s. Used with 1G networks in the 1980s, these 1G phones were only used for voice traffic, and signals were easily affected by interference. They were also less secure and had poor battery life. In the 1990s, 2G mobile phones started with the introduction of a range of 2G networks (Figure 2.8). Second generation phones are digital, with functionality still mainly focusing on voice communication but with high rate and good quality with roaming ability. Second generation phones also support limited data communication and applications, such as the popular SMS, and games. Security has been improved.

2.5G mobile phones have all the functions of 2G phones plus the capability of always-on data connection. This allows access to mobile Internet through Wireless Application Protocol (WAP, Section 2.5.2) with much improved data transmission rates. Charging for phone usage on data communication is usually based on the data volume rather than the time spent. Currently, 3G phones are in use, with the initial aims of ‘any time, any place and anywhere’ usage both for voice and high speed data services. The 3G phones have features such as access to mobile Internet, video-telephony and other new services. However, so far, 3G phones are still mainly being used for voice communication and SMS. Even though there is development work towards the next generation of 4G phones, there is no unified definition of 4G. There will certainly be more convergence in 4G with the aim of delivering users a range of seamless services with high data transmission rate across a range of access technologies, including mobile networks, WiMax and WLANs (Sections 2.3.1 and 2.3.2). Although current mobile phones are mainly used for voice and SMS, they support a number of services through voice and data communication, including voice, messaging, e-mail, digital photos, videos, music, mobile Internet, TV and m-commerce (Chapter 10).

Personal Digital Assistants (PDAs) are another type of commonly used mobile device. They are generally regarded as handheld computers with operating system and a range of application software. With the wireless technologies added to PDAs, they can have communication capabilities for the Internet, e-mail, messaging and voice mail. PDAs integrated with GPS technologies enable the devices to be location-aware, and hence offer users the ability to access location-related information and services. In addition, PDAs can provide

multimedia functions such as taking digital photos, playing music and recording voice. The main components of a PDA include: operating systems, memory, battery, colour display screen normally with non-glare and backlighting for different light conditions, on-screen keyboards using a stylus for inputting text, handwriting recognition and optional external foldout full size keyboards. For PDAs with communication ability, they can have conventional fax/modem cards and a cradle for connecting to a PC or LAN to synchronize and transfer data. For wireless connection, PDAs can have built-in Bluetooth technology and infrared (IR) connection. Furthermore, connection can be made through a mobile phone (usually through Bluetooth) and built-in capability to communicate with wireless mobile networks. PDAs are still mainly data orientated, but they are increasingly used connected to networks.

As can be seen from these two types of widely used mobile devices – mobile phones and PDAs – handheld wireless mobile devices tend to be the devices with more convergent technology and with the capability of providing users a range of services anytime, everywhere. More functions are integrated into a single device. Mobile devices play an important role in the technology convergence and in convergence of services. A mobile phone is not just a device for making phone calls in a conventional way. It can be used as a personal organizer, a device for entertaining (music, TV, multimedia services and games), a device for accessing mobile Internet and a device for using services provided by LBS. The mobile phone itself is a convergence of technologies as its electronic hardware derives from semiconductor technology, its phone components from traditional telephone technology and its operating system and application frameworks from computer technology. On the other hand, with integrated technologies, a PDA is not just a data-oriented computer type of handheld device, but can be a versatile mobile device which also provides all the features just described.

The boundary between types of mobile devices is increasingly blurred. Mobile devices continue to have more integrated functionality, such as from mobile phones, PDAs, Web browsers, cameras and other devices. In addition, a wide range of applications from software providers can be run on the devices. For the wireless communication connection, mobile devices can be connected to desk-top computers and servers by USB link, infrared, Bluetooth and WiFi. The data can be transferred and synchronized between them. Take the example of smartphones. Although there is no unified definition of smartphones, they are generally considered as devices for integrating mobile



computing and mobile communications. Smartphones, beyond conventional mobile phones, have an operating system, enhanced data processing capability and the ability to support applications developed by third party developers. The functionality of such smartphones includes voice, messaging, e-mail and Web browsing, as well as further services such as a personal organizer, lightweight PC applications, entertainment (e.g. music, photo, video, etc.) and LBS applications with integrated positioning function. Smartphones might have a keyboard or an icon driven screen, and camera included.

Current versions of the BlackBerry can be seen as this type of mobile device, which includes functions for e-mails, a built-in mobile phone and the ability of accessing corporate information and working with enterprise level programmes. Also the iPhone launched by Apple in 2007 sees itself as a reinvention of mobile phones, providing all the above smartphone features with a touch-screen casing. From another perspective, conventional landline telephones are converging with wireless technologies. i-mode phones used in Japan have been providing many of these functions over the years (Section 2.5.2). In 2007, a new wireless phone service called 'BT Fusion' (from British Telecom in the United Kingdom) was introduced with new handset devices. Such devices can enable users to connect into the wireless Internet through a network of WiFi Hot Spots (BT Hot Spot zones) in Britain. With this service, the location of users can be identified whether at home or on the move. Other convergences are components such as data cards and wireless USB sticks used for connecting between mobile phones and laptops/PCs to increase the voice and data capability for both phones and computers.

A number of key trends in the convergence of devices, both in components of hardware and software, have been highlighted by Walker *et al.* (2007). The first is the growth of high level operating systems (OS). Currently, only the high end of smartphone devices are equipped with these OS, such as Symbian, Microsoft Window Mobile or Linux. This is expected to grow in the future as complex, multimedia types of services expect to run on reusable software components supported by high level operating systems. The second is the continued existence of devices with specialized functionality despite the trend of integrating more functions within mobile devices. The third is the battery life of mobile devices, which will become an even greater challenge with increasing power demands in devices with multifunctions and multimedia applications. Battery technologies, such as fuel cells and inductive charging, are considered to be future solutions



whilst dual batteries are seen as a temporary solution. The fourth trend would be the increasing storage capacity of mobile devices, such as high capacity flash memory and hard disks in mobile phones and external memory cards with higher capacity. In addition, storage technologies are developing whilst the cost of memory is reducing. The fifth trend is the use of WiFi in converged mobile devices, which provides solutions for co-existence between multiple radio interfaces for better performance at the edge of the range. Also, the power requirements for WiFi networking are lower. Finally, there will be much enhanced services particularly with rich content. Such content can also be downloaded and saved using fast occasional access via WiFi. The convergence of technologies will have a further impact in this area.

Wireless mobile devices have only started to be widely used by the general public since the 1990s and increasingly have become a necessity in people's lives. People in developing countries are adopting mobile phones as a cheap, reliable best chance to gain phone services (Comer and Wikle, 2008), whilst people in developed countries are using them as multifunction devices. Therefore, wireless mobile devices are still rapidly evolving in terms of their performance, functionality, capability, size and even shape. With the fast advance of technologies and the availability of multimedia services, the development and use of wireless mobile devices will continue with greater convergence and more potential functionality and capability, also with the focus of being simple to use, application centred and fit for purpose.

## 2.4 GEOGRAPHICAL INFORMATION SYSTEMS

---

As identified in Figure 2.1, GIS are a key technological strand upon which LBS are built. Because mobility and location are at the heart of LBS, there is a need to reach an in-depth understanding of GIS; this is done in Chapter 3. Without pre-empting those discussions here, it is, nevertheless, necessary to provide some orientation for the convergence of the Internet and GIS as Internet GIS and of NICTs and GIS as Wireless GIS; these are discussed in Section 2.5.3. GIS have developed from the 1960s onwards as the principal means of integrating, managing, querying and visualizing geographical data sets. GIS software is typically large, such that full-functionality GIS cannot currently be installed on mobile devices (though some reduced

functionality versions are available for PDAs). Geographical data sets are also typically large and transactions using them are usually long. Their use in mobile situations is currently constrained. However, the technological developments discussed above and the convergences that are discussed below are opening a door for GIS to move towards more ubiquitous use through applications such as LBS that take advantage of the new technological environment and the demand for services arising from the information society and people's increasing mobility.

## 2.5 Convergence of Technologies

---

The convergence of technologies has been a trend in many areas of mobile technologies as shown in the development and usage of wireless mobile devices (Section 2.3.3). The convergence is also a key aspect in the technologies that constitute location-based services (LBS). In the following sections, some aspects of the convergence amongst a range of wireless technologies, the Internet and GIS that lead to LBS are discussed.

### 2.5.1 Convergence of Wireless Technologies

There is a growth in the convergence of various wireless technologies to provide overall better services across different networks by taking advantage of various technologies. The following are a couple of examples that demonstrate such convergence.

The convergence of wireless mobile networks (particular high speed 3G) with WiFi can bring benefits in services provided. With agreement, WiFi can be integrated directly into the wireless mobile networks, such as GSM (2G), GPRS (2.5G), EDGE (2.5G) and 3G networks. For 3G networks, one of the initiatives is to provide users with interoperability, expanded coverage and new service capabilities. Thus, 3G networks can offer high speed data services via a radio access network that can be a stand-alone network or can be working with other networks. When integrating mobile networks with WiFi, the capability of WiFi provides higher data transmission rates whilst mobility of mobile networks provides much greater roaming and voice services on a global basis (Smith and Meyer, 2004). Such convergence also provides wide area coverage for applications which need constant

access to services, such as e-mail, organizer and large volume data access. WiFi networks are accessible in convenient on the move locations through Hot Spots. The vision of 4G can be the possibility of overall convergence. For the detail of converging mobile networks and WiFi, readers can refer to Smith and Meyer (2004).

3G networks can also be integrated with Bluetooth technology. Bluetooth wireless connectivity not only allows devices to be conveniently moved within a personal space, but also provides collaborative communication between individuals, their appliances and surrounding environments. The Bluetooth technology can support potential uses of the Personal Area Networks (PANs) with the connection to the Internet (Rappaport, 2002). The converging of 3G networks and Bluetooth technology allows services to be provided with a wide area connectivity as well as local intercommunication, which could be difficult to achieve by either technology working separately (Muller, 2000). An application example can be as follows: a 3G mobile phone receives an e-mail that has a file attachment as a data transmission; using integrated Bluetooth, the e-mail is forwarded to the user's laptop for storage; the user who is on the move, when in an airport or train station, can approach a Bluetooth services kiosk and use the 3G/Bluetooth mobile phone to print out the attachment (such as to show a client on arrival).

### **2.5.2 Wireless Mobile Internet**

Wireless mobile Internet, often referred to as mobile Internet, is an outcome of the convergence of wireless communication technologies and the Internet. Mobile Internet is one of the major driving forces behind new developments in the area of telecommunication. Mobile Internet can be viewed as the Internet that can be accessed and used by people through their mobile device anytime and anywhere.

2.5G and 3G mobile networks provide the possibility of the wireless mobile Internet. 2.5G networks, such as GPRS, can be employed for non-real-time Internet usage, including the retrieval of e-mail, faxes, and asymmetric Web browsing. It works better in situations where data are downloaded, rather than for uploading data onto the Internet. Third generation mobile networks start to allow unparalleled wireless access. It will be possible to have the Internet access at high data speed of multi-megabit with ubiquitous 'always on' access, as well as communication using Voice over Internet Protocol (VoIP).

Third generation technologies can enable users to access the Internet through their mobile devices to carry out interactive Web activities, receive live music, watch digital TV, as well as having the ability of voice and data access with multiple parties at the same time. Users should be able to do this anytime, anywhere, on the move, at home or in office.

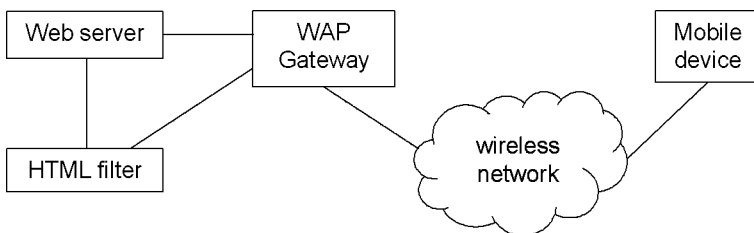
There are a number of issues on accessing the Internet wirelessly in mobile situations using small handheld mobile devices. The standard network protocols, IP and particularly TCP (Section 2.2.1), have been developed as connection-oriented and end-to-end protocols for the Internet with fixed line access and nearly error-free links. TCP is designed with the presumption that 99% of the packet losses are caused by network congestion rather than by network link failure (Taferner and Bonek, 2002). Therefore, the performance of TCP/IP protocols over wireless mobile networks where connection might be lost (such as entering a tunnel) will clearly give rise to problems. In other words, the commonly used Internet protocols, such as HTML, HTTP and TCP, are inefficient over mobile networks when transferring large amounts of data. HTTP and TCP are not optimized for the connection provided by wireless networks which can have signal loss, long latencies and limited bandwidth. Therefore, mobile Internet services provided by these protocols through wireless networks can often be unreliable, slow, costly and difficult to use. Furthermore, from the device aspect, mobile devices used in mobile Internet usually have small screen size and limited capability, so content must be delivered in a 'no frills' format. An HTML written Web page cannot be displayed and navigated efficiently and easily on the small size screen of a mobile handheld device. In addition, bandwidth constraints of mobile services mean that Web content must be optimized for delivery to mobile devices. To get the information in this form, Web sites have to be built differently. Therefore, there has been the need for solutions such as the Wireless Application Protocol (WAP) standard developed particularly for wireless mobile Internet services, and i-mode (similar standard to WAP) that is very popular in Japan.

### **2.5.2.1 Wireless Application Protocol (WAP)**

WAP is an open international standard for applications that use wireless communication. Its principal application is to enable users to access the Internet wirelessly with small handheld mobile devices without a full size browser (i.e. mobile phones, PDAs). WAP is designed to

work with most wireless networks. It can work with mobile networks with relatively low speed. WAP can be built to run on any operating systems, including PalmOS, Window CE, OS/9, FLEXOS and JavaOS. It offers the additional advantage of providing service interoperability between different device families. WAP is one *de facto* standard for providing mobile Internet communication and advanced services on wireless mobile devices.

WAP is a set of protocols and services, including WAP Transaction Protocol (WTP), Wireless Transport Layer Security (WTLS), Wireless Markup Language (WML) and WMLScript. WTP is the transaction layer protocol, an equivalent to TCP/IP, designed for both error-free and packet-loss transactions. WTLS is for transmitting encrypted information for security. WML is a light version of HTML, designed specially for displaying text and graphics on a mobile device with a small screen. WMLScript is a scripting language, like JavaScript on the Web, allowing the interactivity between mobile devices and Web pages. Interactivity is particularly important for small screen devices in using mobile Internet. Mobile Internet sites developed using WML and WMLScript are known as WAP sites. Other mobile Internet sites can be dynamically converted to WML and accessed via a WAP browser. A WAP browser is designed to provide all of the basic services offered by a computer-based Web browser, but is simplified to operate within the restrictions of a mobile phone. A key component in WAP is the WAP gateway. It is a special server to improve the efficiency of signal exchange between the Internet and a wireless network. The requests and information between TCP/IP protocols of the Internet and WAP protocols of mobile devices are translated through the WAP gateway. The WAP gateway is situated between a wireless network and the Internet (Figure 2.10). If a Web server can provide WML written content,



**Figure 2.10** An illustration of the basic WAP architecture.

the WAP gateway sends information from the Web server directly to a mobile device via a wireless network. If a Web server can only provide normal HTML format pages, a filter will be used to translate HTML into WML, and then send them. When a mobile device is used to access mobile Internet through WAP, the device has to have a micro-browser installed to use WAP.

In general, WAP uses binary transmission for better data compression. It is optimized to suit the communication of wireless mobile networks which often have long latency, low-to-medium bandwidth and intermittent coverage. WAP can work across a wide variety of wireless communications, using both IP and other optimized protocols. With WML and WMLScript languages, content of mobile Internet can be optimized to be used on a small screen, and to be suitable for navigating Web pages with one hand without a keyboard. WAP content Web pages can comprise both text and graphics. WAP enables mobile Internet to offer interactive data communication and applications through a range of mobile devices, such as e-mailing by a mobile device, tracking of stock market information, updating news headlines, listening/downloading music, watching video/TV and so on.

### **2.5.2.2 i-Mode**

i-mode is the Japanese mobile platform which offers wireless mobile Internet services. The i-mode system has been developed by the Japanese mobile firm NTT DoCoMo, and the i-mode services are also provided by the company. It has its own proprietary wireless data service and Internet microbrowser technology. i-mode provides wireless access to Internet services as well as offering the ability to send and receive e-mails through its iMail service. i-mode can support colour graphics and interactive Web page browsing using 2G mobile networks with a data transmission rate of 9.6 Kbps. i-mode provides various services from horoscopes, games, music downloads and movie listings, to news, instant messaging, banking, stock information (and more) via mobile Internet. i-mode started in 1998 and had around 47 million subscribers in June 2006 in Japan. Outside Japan, network providers such as O2 have been offering i-mode in the United Kingdom since 2004 which gives users access to mobile Internet and the ability to send and receive e-mails.

In order to use i-mode services, special i-mode mobile phones are required; they are designed specially for i-mode with its own Web microbrowser. i-mode is an 'always on' service, which means that

i-mode phones are always connected and ready for using the services. The i-mode phones are connected to the Internet through an i-mode gateway that deals with the information and requests between the phones and the Internet. E-mails and other information are sent from the i-mode gateway to i-mode phones. The Web sites that i-mode subscribers can read and use are specially built using compact HTML (cHTML) standard designed for i-mode services. i-mode phones cannot display the contents of WAP pages. The Web sites built by i-mode are generally very small in size, so that they can be downloaded to i-mode phones in a timely manner at a speed of 9.6 Kbps (see also Section 4.5). i-mode is planned to migrate to higher-speed broadband with the move to 3G networks (Gralla, 2006).

### 2.5.3 Internet GIS and Wireless GIS

Internet GIS is often regarded as GIS functionalities that are available and used by remote users via the Internet (Pandey *et al.*, 2000). Internet GIS is also often termed as Web-based GIS and on-line GIS. There are two basic approaches for implementing Internet GIS: the server-side approach and the client-side approach. Internet GIS based on the server-side approach has the GIS software residing on the server with the user at the client side accessing GIS functionality over the Internet via a Web browser. The data processing and analysis carried out by the GIS software are mainly performed on the server-side after receiving the request sent by the user via the Internet. The results are then sent back to the user. The major benefit of the server-side approach is the unified centralized hardware, software and data which can be easily maintained and updated. The disadvantage would be the slow processing and response time of the whole system when the site has to deal with too many users simultaneously.

Internet GIS based on the client-side approach offers users at the client side the possibility to carry out GIS data processing and analysis locally (with Internet connection) by downloading simple GIS functions and the required data from the server via the Internet. In this approach, it will lessen the burden on the GIS server side and on data traffic on the Internet. However, there is also less centralized management of software and data, and less advanced GIS functionalities available to users. The approach taken to implement Internet GIS depends on a range of factors such as application types, main requirements of targeted users, user volume, network speed and provider's capability.



For more details, readers are referred to Pandey *et al.* (2000). In Section 3.2 two providers of GIS-driven Internet mapping are compared and in Section 10.5.1 a patent for delivering GIS-based mapping over the Internet is examined. Further developments of Internet GIS (and also Wireless GIS) are likely to move towards interoperable distributed component GIS using agent-based technologies (Section 3.6). This would make GIS very light and portable with components of GIS downloaded over the Internet on demand to mobile devices. Internet GIS also holds out potential for using Grid computing.

Wireless mobile GIS, often shortened to Wireless GIS, is the result of the convergence of wireless mobile technologies and GIS with real-time and mobility characteristics. Here a differentiation is made between wireless GIS and mobile GIS. Mobile GIS, usually integrated with GPS technology, has been used in the GIS area for some time. Mobile GIS is regarded as GIS used in the field, often through handheld devices such as pocket PCs and PC tablets for work in the field or away from office settings. But it is important to note that mobile GIS used by a handheld device is not connected to networks. Its applications are mainly for spatial data collection and maintenance in the field, as well as processing data and managing services. Mobile GIS usually has most of the GIS software and necessary data installed on the terminals in order to operate as a stand-alone GIS whilst in mobile situations. Because there is no real-time wireless connection, data in mobile handheld devices have to be synchronized with the main GIS facility and databases on the server through fixed line connection. Wireless GIS integrates wireless mobile networks (discussed in Section 2.3.1) and GIS to offer wireless connectivity. This enables wireless GIS to have the real-time capability of interacting with external software and remotely accessing and managing data (Braun, 2003).

Wireless GIS provides a real-time data, software and application access through mobile devices in mobile situations. Therefore, using wireless GIS does not require fixed line data synchronization and extra applications installed on each mobile device. It can also reduce duplication in data processing. Wireless GIS extends the functionality of GIS with its portability, usability and flexibility (Drummond *et al.*, 2007). The portability of wireless GIS takes advantage of wireless networks and devices. The wireless technologies provide the possibility of exchanging and analysing spatial information in a real geographic world in real-time. Wireless GIS provides the potential for users to employ GIS in more mobile and diverse situations and for a range of applications. Despite the advantages brought by wireless GIS, there are concerns and



issues about it. The issue about bandwidth and data transfer speed of wireless mobile networks directly influences the usage of wireless GIS, as many GIS applications often deal with large data volumes. The development of technologies could solve some of these problems. However, the fast development in wireless mobile network and mobile device technologies and changes in these industries (Sections 2.3.1 and 2.3.3) also raise concerns for the reliability and consistency of the wireless network infrastructure and the devices on which wireless GIS relies on to build, implement and deliver applications.

Wireless GIS integrates numerous components associated with mobile telecommunications, hardware, software, security, data and database (Braun, 2003). It is the essence of converged technologies. Wireless mobile telecommunication is a key element in wireless GIS. Such wireless mobile networks provide connection between wireless GIS users and the corporate network where the main GIS software, data and applications reside. Some of those wireless networks, such as WiFi, described in Section 2.3.2 could be used in wireless GIS. Therefore, the capability of various wireless networks described in Sections 2.3.1 and 2.3.2 can have a direct effect on the implementation and usage of wireless GIS. Another element closely related to wireless GIS is mobile handheld devices. With the fast development in this area, as discussed in Section 2.3.3, mobile devices are continuously evolving and converging. The suitability of mobile devices for wireless GIS can be dependent on the requirements of the particular application (e.g. real-time data collection, or navigation) which might emphasize more on processing and displaying capability or on the mobility and flexibility. Also related to devices used in wireless GIS can be battery consumption, processing capability and screen quality. Software in wireless GIS can be run from a server with mobile devices as thin clients. Some of software can also reside in mobile devices with wireless connection for real-time data transfer. Data in wireless GIS can be readily exchanged via wireless connection. Importantly, only data needed should be accessed and transferred both for security and for reducing unnecessary data load on a wireless connection.

### **2.5.4 Towards LBS**

The evolution and convergence of a number of key technology strands which, when brought together, provide the heterogeneous technological foundation for LBS (Figure 2.1), have been examined in this

chapter. It is important to understand these strands and their current trajectories as they fundamentally inform the remaining chapters, which examine in detail how LBS work. By definition, LBS are the delivery of data and information services tailored to the location and context of a mobile user.

The level of technological convergence that underscores LBS cannot be over-emphasized and is encapsulated in the following advertisement for a smartphone:

Sat Nav with step-by-step directions for life's journeys.

Multi-media player for entertainment en route.

Facebook and Instant Messenger for socialising on the move.

Camera with video capture for sharing the memories.

Business grade e-mail and web browser for everything else.

(<http://www.blackberry.co.uk/8110>)



# Chapter 3

## GIS and Geo-Information Engineering

### 3.1 Introduction

---

In Chapter 1, in defining location-based services (LBS), it was identified that their focus was on information and geographical location. In Chapter 2 it was seen how the development and convergence of technologies has set the stage for LBS in terms of their technical feasibility. Geographical information systems (GIS) were identified as one of the underlying technologies that are specific to organizing, handling and visualizing data that describe geographical locations. Clearly, for such a technology to exist there must be an underlying science which provides the theoretical basis that informs our understanding, design and application of the technology. This is generally referred to as GIScience. The maturity of GIScience/GI Technology and the accumulated experience of successes and failures in their application are paving the way for GI Engineering, that is dependably engineered solutions to society's use of geographical information. LBS as they mature will be, in our view, products of GI Engineering.

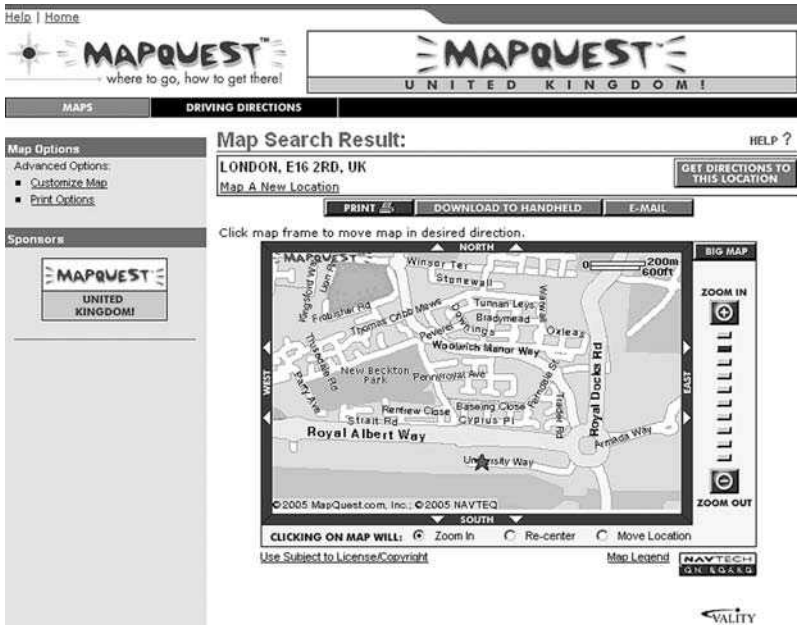
In this chapter we will introduce how data about geographical locations are managed, analysed and visualized using GIS, what are the issues of GIScience that need concern us and hence the important design elements in engineered applications. This chapter will necessarily be selective as there is not the space here to cover all aspects of GIS. For a short but well structured overview there is Chapters 2 and 3 of Brimicombe (2003); for whole textbooks there are Burrough and

McDonnell (1998), Chrisman (1997), Heywood *et al.* (2005), Longley *et al.* (2005) and Wilson and Fotheringham (2008); for an advanced text that is close to the subject of this book there is Peng and Tsou (2003). For a reader who is already very familiar with GIS, reading Section 3.3.2 and Section 3.6 and then moving on to the next chapter is possible.

### 3.2 Where is . . . ? How Do I Get there?

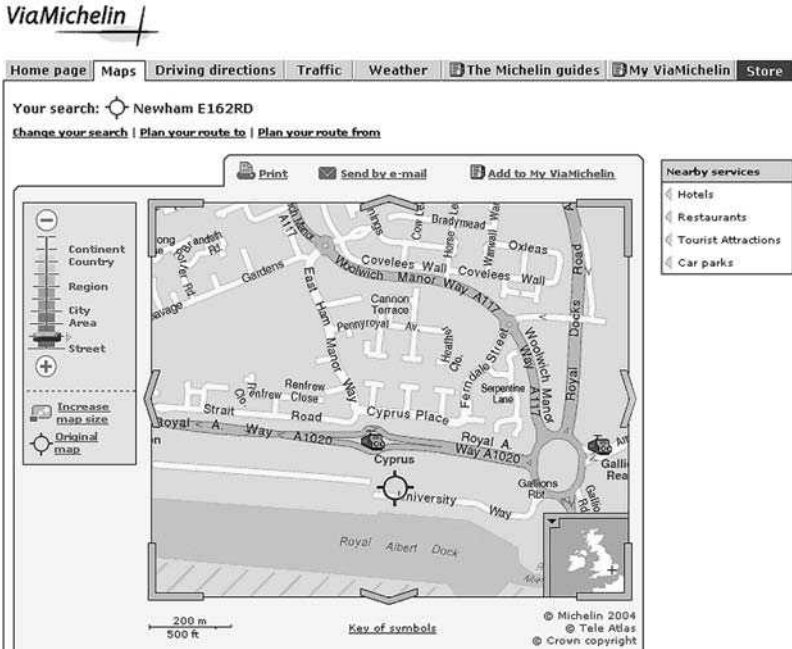
---

The less we know a place, the more help we need in understanding how to get there and finding our way around. Of course, we could always treat every trip to somewhere new as an exploration of *terra incognita*, but more often than not we don't have the time to indulge ourselves in such protracted adventures. We want information, preferably accurate information. We could ask a friend or a person on the side of the street for directions, and depending on your desired mode(s) of transport you would get a list of sequential instructions on how to get there. How correct or easy to follow these instructions might be is another matter! If your experience is anything like ours, then the quality of information garnered in this way has much to be desired. It is not necessarily that it is wrong; the person who you have asked might be able to make the journey with no trouble at all and not get lost once. They have a cognitive understanding of the route and relevant mode(s) of transport inside their head (often referred to as a 'cognitive map'), but it is in the communication of that knowledge and subsequent storing and interpretation of the instructions inside our own head that more often than not gets us lost. That is why we consult maps. For centuries, maps have been a way of recording and communicating relatively objective truths about geographical location. The issue of 'objectivity' will be returned to later, but suffice to say that if a map has been produced by some mapping authority, whether it is an A to Z of London or an Ordnance Survey map, and providing it has a reasonably recent publication date, we generally trust it as a source of dependable information. These days we are just as likely to consult the Internet, perhaps even more so than we would an atlas or printed map sheet on the grounds of both convenience and cost. Two examples of world-wide mapping available on the Internet are Mapquest (<http://www.mapquest.co.uk>) and ViaMichelin (<http://www.viamichelin.co.uk>).



**Figure 3.1** Web-based delivery of maps – MapQuest (Copyright 2005 MapQuest.com Inc.; Copyright 2005 Navteq; accessed 11 July 2006).

The maps in Figure 3.1 and Figure 3.2 were retrieved using the postcode for the University of East London, Docklands Campus (E16 2RD). Whilst road and district names can also be used to get maps from these services, postcodes (zip codes) have become in many countries around the world a powerful form of detailed geographical referencing of places. When we look at these two maps they are obviously of the same place: the same road configuration, the same body of water at the bottom of the map and both have a sign to show where the postcode is located. There are also similarities in the look and feel of the interface that allows interaction with the maps. The user can zoom in or out using the bar tool provided at the side; the map can be scrolled up or down or sideways by clicking on the edge of the map; the overall size of the map can be changed and in both cases the map can be printed or e-mailed to oneself and/or someone else. And yet the maps themselves look quite different in appearance. They have different colour schemes, the labelling of the roads and other features is different in the detail, MapQuest includes a park, ViaMichelin includes the stations for the light rail transit system. So, on the one



**Figure 3.2** Web-based delivery of maps – ViaMichelin (Copyright 2004 Michelin; Copyright Tele Atlas; Crown Copyright; accessed 11 July 2006).

hand, commonalities are present because we expect maps to have a certain content (the roads as a minimum) and there are certain common things we would like to do with maps such as pan, zoom and print them out. The latter are what we call system *functionality*. The differences however arise from choices both in the key content that the map provider has deemed to be of likely interest to us and in the style of presentation of that content, which the map provider has also deemed to be an elegant and readable form of communication. These reflect both the *data model* that determines information content and the *cartographic design*, that is, the means of visualization. These are returned to in more detail shortly.

The Mapquest and ViaMichelin Web sites offer another interesting (and useful) piece of functionality: to provide driving directions between two locations. In Figure 3.3 and Figure 3.4 the driving directions from the University of East London Stratford Campus at E15 4LZ to its Docklands Campus at E16 2RD are shown as provided by Mapquest and ViaMichelin respectively. Both routes have roughly the

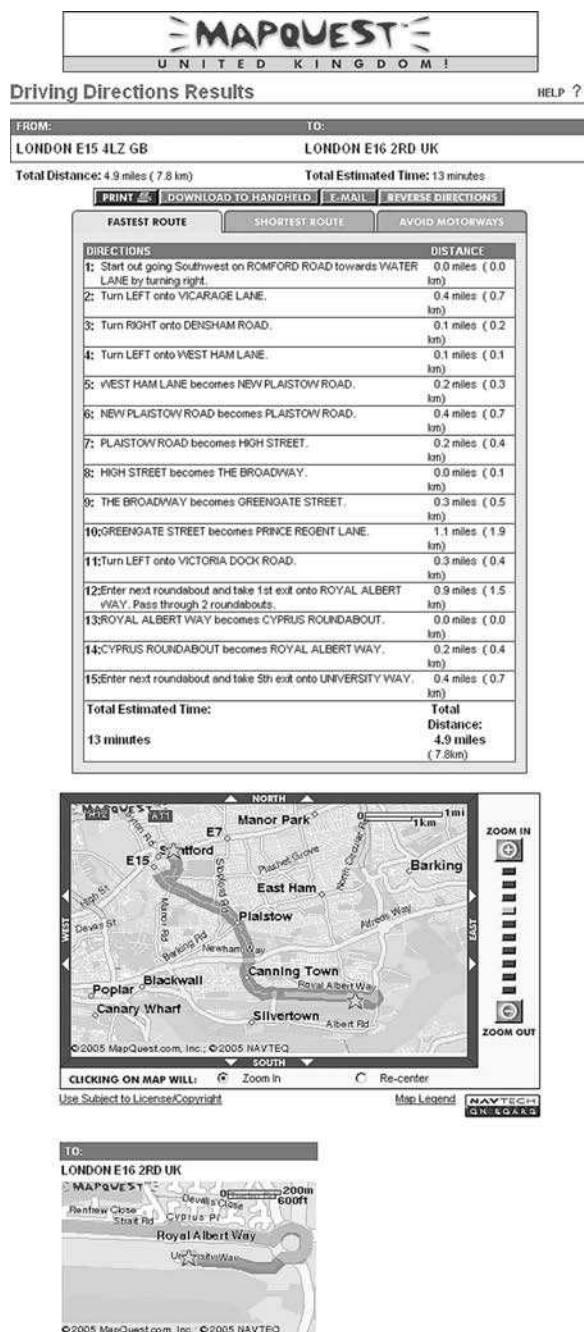
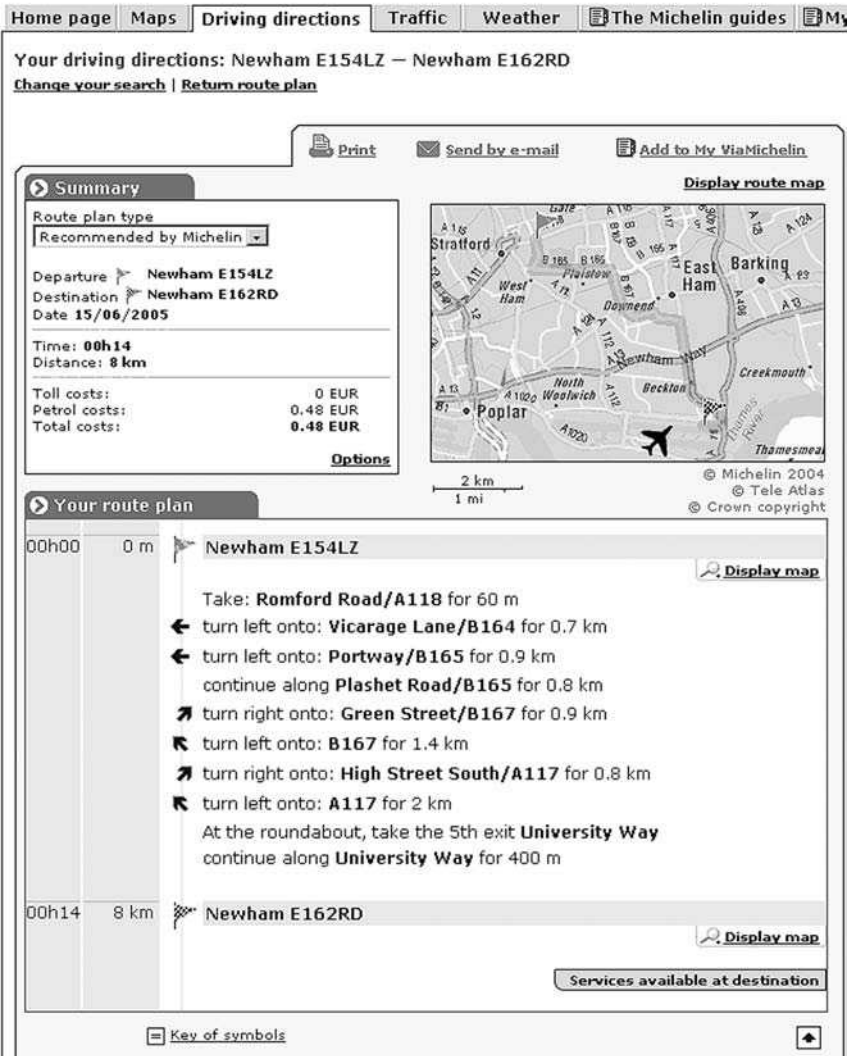


Figure 3.3 Web-based service for driving directions – MapQuest (Copyright 2005 MapQuest.com Inc.; Copyright 2005 Navteq; accessed 11 July 2006).





**Figure 3.4** Web-based service for driving directions – ViaMichelin (Copyright 2004 Michelin; Copyright Tele Atlas; Crown Copyright; accessed 11 July 2006).

same length (within 200 m) and would take roughly the same time (within a minute of each other), but the routes themselves are different. The differences arise because the *algorithm* that underlies the functionality will have been programmed with different objectives. These might

be to give the shortest route, the shortest travel time, giving the least number of directions or avoiding minor roads. From both Figures 3.3 and 3.4 it is evident that the user can make some choices for different algorithms by clicking the appropriate tab or selecting from a pull-down menu.

Internet sites such as Mapquest and ViaMichelin are supported by GIS technologies which organize the data and provide the necessary functionality and their underlying algorithms including the delivery of responses to queries through an Internet browser. GIS can come in many forms over a range of platforms from personal digital assistants (PDAs) to mainframe. Here the focus will be on the common aspects in order to give a broad understanding of GIS. To help illustrate many of the points, Microsoft MapPoint will be used because of its easy accessibility, low cost, ability to seamlessly integrate with other Microsoft Office products and because it comes with a wealth of embedded data (which varies depending on whether it is the North American or European version). It is a good starting point for anyone who is unfamiliar with GIS and would like to give it a go. An excellent manual (and 60-day trial software) is provided by Holtgrewe and Freeze (2002).

### 3.3 Defining GIS

---

GIS are often thought of as just being the software resident in a computer. Certainly the software is very important. Most GIS software, however, like spreadsheets and databases, comes with no data and is essentially a toolkit ready for use. So an equally important aspect of GIS are the data – what do they represent and what is their quality, who provides them and at what cost, who will update them and how often. Involved in the production of data are people and organizations, and standards devised by professional and regulating bodies. Also, it needs to be remembered that using GIS is not a game of solitaire – they are used to solve problems and the communication of those solutions through their outputs can be used to generate revenues, contribute to good governance and to understand and manage a whole range of physical and social phenomena. So GIS are not just tool-boxes, they are a way of working. Within this broader concept of GIS there is a broad consensus of what GIS are but no single form of words that provides a universal definition of GIS that is going to satisfy

everyone. This is hardly surprising when all the different applications that GIS can have are considered. The definition we would like to use comes from (Dueker and Kjerne 1989 p. 7) with our modifications in brackets:

a system of hardware, software, data, people, organizations and institutional arrangements for collecting, storing, analyzing, [visualizing] and disseminating [spatial] information about areas of the earth.

### 3.3.1 An Historical Perspective

So, where did this all begin? GIS began in North America in the mid-1960s when professionals and academics from a number of disciplines became excited over the prospect of handling and analysing spatial data in a digital environment. The focal points were in the Harvard Graduate School of Design, the Canada Land Inventory and in the US Census Bureau. Each of these made important contributions towards laying the foundations of the GIS industry seen today. The first digital mapping package, called SYMAP, was operational in 1964 and was created in the Laboratory for Computer Graphics and Spatial Analysis, part of the Harvard Graduate School of Design. Using line printers, primitive maps could be produced that allowed landscape themes in human and physical phenomena to be visualized so as to recognize spatial similarities or groupings (McHaffie, 2000). Although these maps gave a coarse line by line representation using equally spaced characters and symbols, the leap was to recognize that virtually any spatial phenomenon could be represented in this way and stored digitally. The approach was evolved and by 1971 a package called GRID had been developed with a flexible command-line user interface; by the late 1970s the Laboratory had developed ODYSSEY, a line-based (vector) prototype of today's GIS. Both GRID and ODYSSEY could integrate data about different landscape themes by treating each theme as a separate layer of data. Since then, the use of data layers has remained a fundamental aspect of GIS.

Meanwhile in Canada, also in the mid-1960s, a prototype GIS was being developed to meet the needs of the Canada Land Inventory to map existing land uses and to derive maps of land use potential for agriculture, forestry, wildlife and recreation (Tomlinson, 1984). The development of GIS was seen as the only economic solution to seamlessly mapping the whole land area of Canada, despite what we would now consider the primitiveness of the technology and its high expense

at that time. The system did not become fully operational until 1971. Along the way a number of key developments were realized: optical scanning of existing paper maps as raster (grid) files, conversion of those files to vector (line) data and the use of a database management system to organize spatial data. Eventually some 10 000 digital maps were produced and edge-matched so they could become a seamless map coverage – a very considerable achievement for the 1970s and an important proof-of-concept for large corporate GIS facilities.

Another corporate user at this time was the US Bureau of Census. Its goal was to produce digital map representations of street blocks and census tracts in support of the 1971 census. Digital cartography at the time relied on unstructured collections of points and lines (commonly known as spaghetti files) which, when plotted using appropriate symbology sets, produced comprehensible maps that could be visually interpreted by a user. But if the focus is less on the streets but more on the blocks of buildings that they circumscribe, then there needs to be a way of defining polygons (area features) from the lines and intersection of lines that form their spatial boundary. In other words, the data need to be intelligently structured to explicitly encode the topological (graph-theoretic) relationships of points, lines and polygons. This it did in the implementation of its Dual Independent Map Encoding (DIME) scheme and laid the foundation of topology being a key aspect of GIS data handling. By way of example, the data topology of which road joined onto which other roads was fundamental in working out the driving directions in Figure 3.3 and Figure 3.4 above.

The power of GIS grew through the 1980s and 1990s as the power of computers and their peripherals grew and as the cost of data storage (the price of ever larger hard drives) and hardware in general fell. Thus GIS software was able to migrate from mainframe computers, to UNIX workstations, onto PCs, laptops and palm-size PDAs ... and, of course, soon the mobile phone. Clearly the level of sophistication, functionality and data storage available on a workstation will not be replicated on a mobile phone for some time, but GIS software for PDAs already have much of the sophistication of PC-based software. As with the price of hardware and in line with other software, the cost of buying into GIS has dramatically fallen over the last twenty years with entry level software and bundled data available commercially, to be purchased on-line, for just about £230 for the EU version (US\$299 for the North American version). Needless to say this has widened the use of GIS to a considerable extent into many aspects

of government and corporate operations, though to a much lesser extent in our personal lives. Whilst technological developments strongly facilitated the development of GIS, one key element held back its wider use: the availability and cost of acquiring and storing spatial data. Up until about the mid 1990s this data bottleneck as it was known was a major sticking point, and indeed in Europe generally, official pricing policies for spatial data are still a bone of contention. There were several factors working together that have eventually led to the breaking of that bottleneck and which have opened up an era of data-richness:

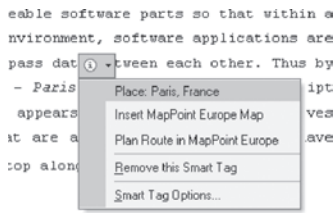
- improved technology and wider use of the global positioning system (GPS), remote sensing and digital photogrammetry to produce a higher degree of automation in data collection of physical landscape features;
- near ubiquitous use of postcodes by governments and businesses as geographical identifiers in data records and transactional databases;
- the advent of data warehousing technologies;
- increased competition in the data provider market;
- more efficient ways of accessing and delivering data on-line.

### 3.3.2 On-Going Development Trends

In the latest phase of development of GIS software, it is possible to discern four important trends that are helping GIS achieve the ubiquity of other software tools such as word processing and spreadsheets.

#### 3.3.2.1 Interoperability

This refers to interchangeable software parts so that within a specific hardware and operating system environment, software applications are able to seamlessly operate together and pass data between each other. Thus by way of example, if I type the city name – *Paris, France* – in this manuscript and I move my cursor to it, a ‘smart tag’ appears which, if I right-click my mouse, gives me options (Figure 3.5(a)) that are available only because I have Microsoft MapPoint installed in my laptop alongside Microsoft Word. Use the ‘smart tag’ and I can get MapPoint to insert a map of Paris into my text (Figure 3.5(b)). In a similar way, I can call up Excel from within Word, work on a spreadsheet and insert the resulting table into my document. Interoperability has very much eased the convergence of once separate software types (word



(a)



(b)

**Figure 3.5** An example of software interoperability within the Windows environment: (a) smart tag offering services; (b) embedded map object via the smart tag (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

processing, spreadsheets, databases, statistics, mapping and so on) to produce an environment in which they work closely together. These software inevitably start to have the same look and feel and follow many of the conventions around interface design such as use of a standardized icon set, common vocabulary within pull-down menus and the ability to read/save a range of file types that other software commonly use. This has vastly increased the accessibility of GIS to lay (nonprofessional) users and is an important step towards ubiquitous use of GIS.

### 3.3.2.2 Open Systems

The GIS vendor community is not large, nor is the GIS university research sector compared with mainstream computer science and informatics. These two groups have come together and formed the Open GIS Consortium (now the Open Geospatial Consortium (OGC); see <http://www.opengeospatial.org>) as a means of fostering interoperability so that the handling and processing of spatial data can become part of the mainstream. The way GIS software had been developed by competing vendors had lead to a lack of interoperability which in the end was going to hinder the growth of the industry. At the same time, with the breaking of the data bottleneck, the growing rate at which spatial data were being collected by a growing number of technologies threatened to overload the vendors and the user community with overflowing tool-boxes of data import, data export and data transformation routines. Thus a cornerstone of Open GIS has been the adoption of a common



model for implementing services that access, manage, manipulate and share spatial data between relevant communities. For example, OGC has promoted the Geography Markup Language (GML) as a version of XML that better handles geographical features in open interoperable communication of spatial data (Huang *et al.*, 2004). Pertinent to LBS is the OGC OpenLS specification (OGC, 2004). Sponsored by key industry players such as ESRI, Oracle and Sun, it establishes an architecture model for middleware to be used in building interoperable LBS applications that operate smoothly with GIS servers (also see Section 10.4). Middleware is the term given to reusable software components that support the rapid development of distributed services and applications. These components sit between the operating system and an application (hence middle) incorporating standardized application program interfaces (APIs) and protocols that allow an application to run across heterogeneous environments. There are hundreds of components available on the open market (see <http://www.componentsource.com>) many supporting conventional business applications but with, as yet, few supporting GIS and even fewer supporting LBS applications. Open and middleware component GIS will be a major development area contributing both to interoperability and ubiquitous applications of GIS.

### 3.3.2.3 Agent-Based Technologies

In the early 1990s, research into artificial intelligence faced a critical issue: how might a large number of actors and experts interact in trying to achieve a common set of goals but not in isolation to each other? At the same time a new paradigm in computer science – object-oriented modelling – provided the means of developing programme code that focused on objects (entities) rather than purely on processes (algorithms) as had previously been the case. This object-oriented approach is the foundation, for example of programming languages such as C++ and Java and of the scripting language that comes with Microsoft Office products (VBA: Visual Basic for Applications), allowing users to automate and customize their particular applications. Agent-based modelling grew out of object-oriented modelling allowing software to have more embedded intelligence and to be mobile across networks. One application of agent-based modelling is simulation using multi-agent systems, which implement a collection of concepts and techniques that allow heterogeneous software (different types of agents or agents possessing a range of characteristics) to work

together in complex interactions. They are an attempt to understand the aggregate behaviour, or emergent patterns, from the behaviour of individual entities which can communicate, sense their environment and interact. Thus an agent is a software entity, situated in an environment and acting within a social structure (Ferber, 2005). It can:

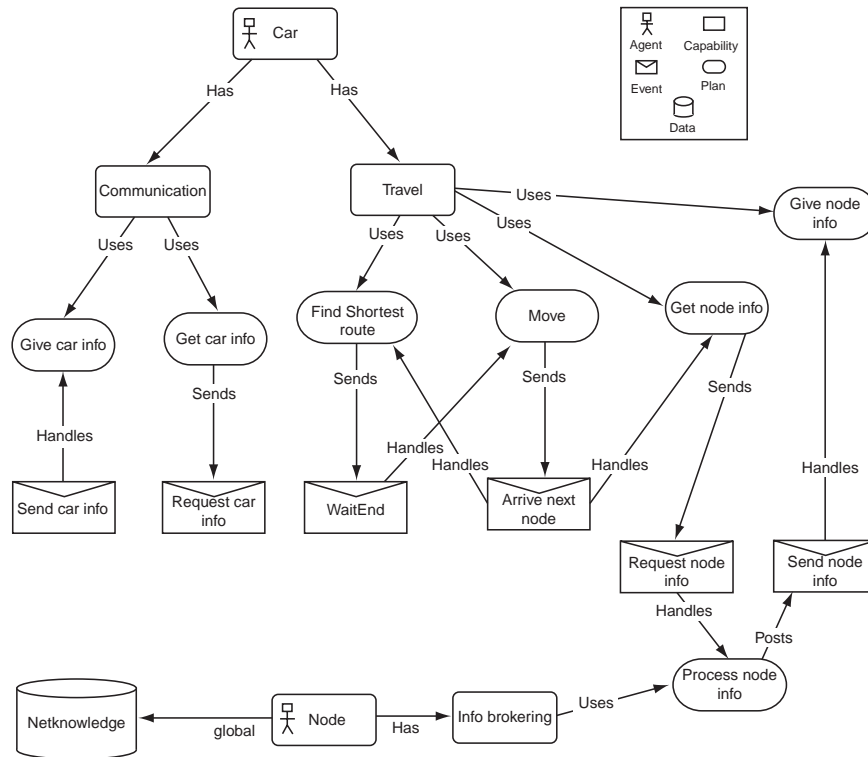
- move within its environment;
- sense environmental states;
- communicate directly or indirectly with other agents;
- be driven by internal tendencies or needs (goals, desires, utility);
- conserve itself and if necessary it can reproduce itself.

Agent technologies are finding widespread use as both distributed component software (see open systems above) and in multi-agent systems used to study the dynamics and emergent behaviour in social, economic and ecological systems. Research has been conducted into using agents and GIS in order to conduct *geosimulations* (Benenson and Torrens, 2004; Albrecht, 2005; Sengupta and Sieber, 2007). Agents are going to be a key component of LBS; an experimental use of multiple agents for regulating vehicle routing along a road network is shown in Figure 3.6.

### **3.3.2.4 Location-Based Services**


The GIS community has long been looking for the ‘killer app’ that will become so widely used in society that GIS will emerge from its niche to achieve ubiquity and thus become part of mainstream IT. Certainly some applications have become widespread and well entrenched, such as the use of geodemographics by marketers and hot spot mapping by the police. But these are still specialist applications and not killer apps. LBS currently hold the greatest promise for a killer app given the uptake there has been for on-line mapping and wayfinding, such as delivered by Mapquest and ViaMichelin and latterly by Google Earth and Google Maps. However, as seen in the last chapter, although there has been a technological convergence that has prepared the ground for LBS, the degree of technological heterogeneity involved in LBS means that it is not just GIS on their own that will determine whether or not LBS are the killer app. The next chapter (Chapter 4) is devoted entirely to discussing the nature of LBS and that discussion is held over until then.



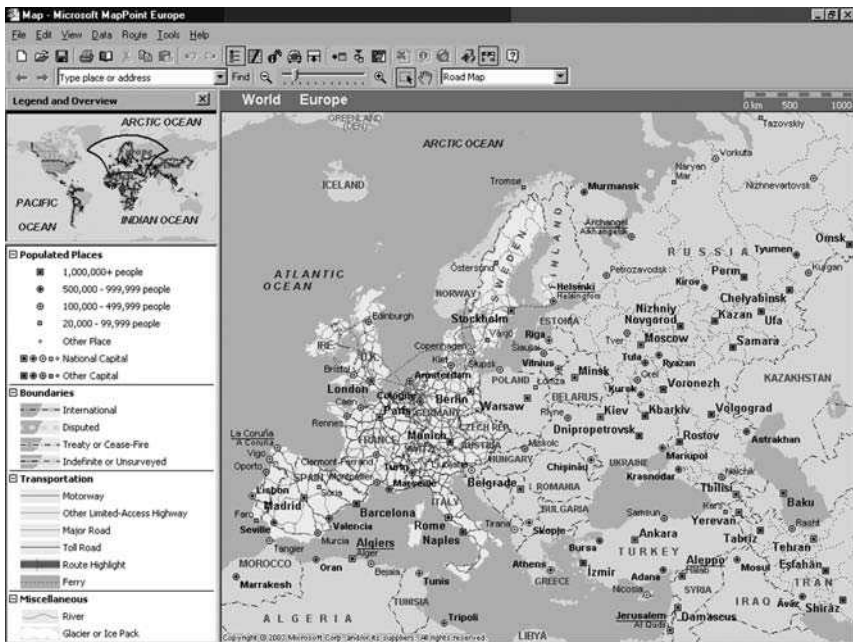


**Figure 3.6** An example of agent modelling for LBS.

### 3.4 Exploring GIS Software

There are a number of GIS software programs being widely used in business, government, utilities and universities, the most popular being MapInfo and ArcGIS. There is also free GIS software (*freegis.org*) and free viewers (e.g. ArcExplorer from [www.esri.com](http://www.esri.com)). To explore here the basic functionality of what GIS software has to offer, it has been decided to use Microsoft MapPoint (<http://www.microsoft.com/mappoint/default.mspx>) as it is both inexpensive, and comes pre-packaged with data. It also illustrates well many of the basic concepts of GIS. Even its icon on the toolbar –  – looks like the sort of pin that might have traditionally been stuck in a wall map. As with any icon, double click it and things start to happen...

The opening screen of Microsoft MapPoint is shown in Figure 3.7 and the most noticeable feature is the map of Western Europe (as this is the Western European Edition of MapPoint) with a scale bar on the top right hand corner. To the left are two smaller



**Figure 3.7** Opening screen of Microsoft MapPoint, Western European Edition (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

windows: the lower one containing a legend, the upper one containing an overview map showing the area covered by the main map in relation to the rest of the world. Three important aspects of map making fundamental to GIS are present: measurement and scale, projection and symbology. Other less obvious principals are also at work to do with the basic functionality of GIS software. These are all discussed in turn below.

### 3.4.1 Measurement and Scale

It should be very obvious that if the main map in Figure 3.7 is to be viewed, then all its features – coastlines, borders, roads and cities – must have been surveyed and digitally stored using a common system of measurement. Because every part of every feature is located on a three-dimensional body (the globe) in theory three coordinates would be needed to fix the position of anything. But since, conventionally, it is assumed that most features we would want to map are on the surface of the globe (rather than in it or above it), a spherical coordinate system with two axes can suffice. For centuries this coordinate system has been latitude and longitude. Arguments over where the prime meridian ( $0^\circ$  longitude) should be located prevailed for a long time but it was fixed in 1884 to pass through Greenwich. At a local level, latitude and longitude are less convenient, particularly for mapping smaller features. Take London as an example: degrees of longitude are positive to the east of the meridian and negative to the west of the meridian, a second of longitude ( $1/3600$ th of a degree) is approximately 30 metres in length whilst a second of latitude is approximately 20 metres in length. This is clearly awkward. So Britain, like most countries, has devised its own plane coordinate system (its National Grid, measured in metres – see <http://www.ordnancesurvey.gov.uk/oswebsite/freefun/geofacts/geo0667.html>), that is nevertheless mathematically keyed into latitude and longitude. If each country's survey data is transformed into latitude and longitude and combined, the map in Figure 3.7 can be constructed.

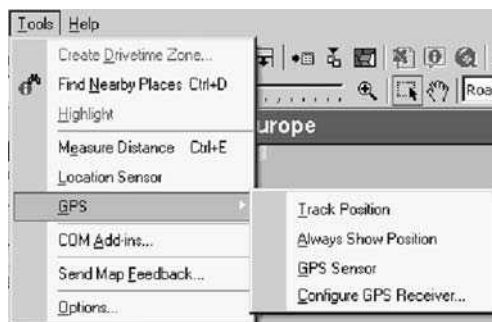
A number of different data types, or ways of measuring things, can be represented using maps. It has become conventional in the social sciences to classify measurements into four classes (Stevens, 1946):

- **nominal** data, where objects are identified by a name (e.g. France, Germany) or fall into a named group (e.g. European Union);

- **ordinal** data, where objects are ranked in some order such as smallest to largest;
- **interval** data, where objects are measured against a scale that is arbitrary in terms of its zero point and interval (e.g. temperature where 0°C is arbitrarily the freezing point of water and 20°C is not twice as hot as 10°C, only 10°C hotter);
- **ratio** data, where objects are measured against an absolute zero and where relative ratios are preserved (e.g. percentages where 0% is absolute and 30% is twice as much as 15%).

Map coordinates themselves may be seen as not sitting well with this classification. On the one hand zero is arbitrary (e.g. Greenwich meridian) and the interval (whether in degrees or metres) preserves relative ratios, yet on the other hand two measurements are required (easting and northing) in order to specify any one point. Properly, reference systems such as map coordinates should be seen as a data type in their own right, and is one reason why there is special GIS software. Another issue concerning data that ought to be stated at this stage is that most data used in GIS have a 'use-by-date' in the sense that many measurements of geographical phenomena are snapshots in time. Since most geographical phenomena are dynamic over varying time scales, the utility of data has a finite life beyond which they are deemed out of date. Of course, wait long enough and they may become historically interesting, but LBS applications certainly focus on data that reflect well the current situation.

MapPoint has functionality for displaying position in real-time using GPS (Figure 3.8). GPS and other position fixing technologies will be discussed in detail in Chapter 6, so will not be dealt

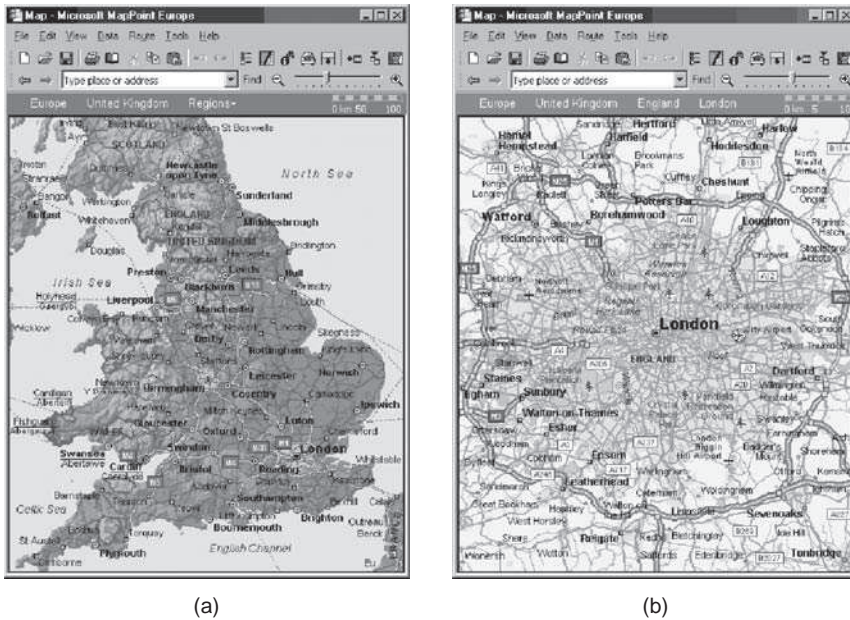


**Figure 3.8** Real-time GPS functionality in Microsoft MapPoint (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

with here. It suffices to say that in ‘track position’ mode, MapPoint reads the GPS determined location every 15 seconds and moves a car icon (🚗) on the map to that location. As the icon approaches the edge of a map, so the map will be automatically re-centred. The coordinate system used by MapPoint for this and other functionality is longitude and latitude so there is no need to transform data to national grid reference systems.

Scale and data resolution are fundamental issues in GIS and yet continue to be problematic (Lam and Quattrochi, 1992; Atkinson and Tate, 2000; Goodchild, 2001). As discussed above, the word scale can be used to refer to measurement type. It can also be used to refer to the extent of some area (e.g. continental drift is a large scale phenomenon). In the context of maps it is the *representative fraction* that tells us that one measure of something on a map is some multiple on the ground. So a scale of 1 : 100 000 indicates that 1 cm on the map is 100 000 cm or 1 km on the ground. Representative fractions tend to be nice round numbers that are easy to deal with in mental arithmetic. This is fine for paper maps which are carefully printed to preserve the designed representative fraction, but what about digital maps that can be zoomed in or out, displayed on different size screens, captured as graphic images and pasted into documents (such as the many figures in this chapter) or squeezed into the display of a mobile phone? Not only would it be necessary to keep track of changes to the representative fraction but we wouldn’t necessarily end up with nice round numbers anymore. In these cases it is much better to use a scale bar to visualize the representative fraction, as illustrated in Figure 3.9. Here the zoom capability of Microsoft MapPoint is illustrated firstly with the whole of England and Wales (which is a zoom in from the whole world) and then into the whole of London. The scale bar to the upper right of the map in Figure 3.9a shows 100 km coming down to only 10 km in Figure 3.9b – notice also the scale bars incorporated into Figure 3.1 and Figure 3.2. The scale bar is integral to interpreting what is being shown on the map in terms of real extents on the ground. Figure 3.11 shows the maximum zoom in the centre of London with the scale bar showing only 100 m.

Data *resolution*, that is the smallest discernable feature within the data, is linked to scale such that the more one zooms into an area, the more detailed features one is able to show. Compare the level of detail shown in Figures 3.9a, 3.9b and 3.11. To include the level of detail shown in Figure 3.11 inside Figure 3.9a would result in such a mass of clutter that the map would be rendered useless for any



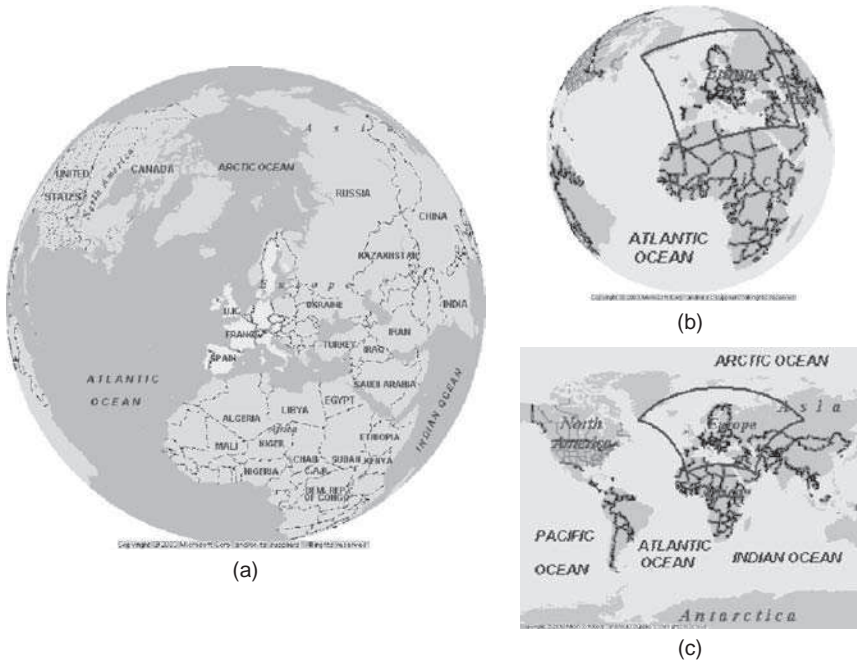
**Figure 3.9** Scale change from (a) England and Wales to (b) London (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

perceivable purpose. This, as will be discussed in later chapters, is a real issue for LBS as there is a tension between providing sufficient detail to small screen devices and yet being able to view a large enough area so as to make decisions about, say, getting from one's current location to some target location.

### 3.4.2 Map Projection

Returning to Figure 3.7 and looking at the small overview window on the upper left, the conventional atlas page appearance of the main map widow has become distorted into a weird, almost fan-like shape. This arises because every map has a *projection*. Our world is round, or almost so (Figure 3.10a), so portraying the whole of it on a flat screen or sheet of paper becomes a challenge. If we imagine peeling the surface off the Earth and laying it flat, not only would it be necessary to tear it in some places but other areas left intact would inevitably crumple up when flattened. The traditional way cartographers solved this problem was to imagine that a light is being shone from the centre





**Figure 3.10** Map projections: (a) globe representation of the world viewed from above Europe; (b) map area of Figure 3.7 shown on a globe; (c) map area of Figure 3.7 when projected using Mercator's Projection onto a flat surface (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

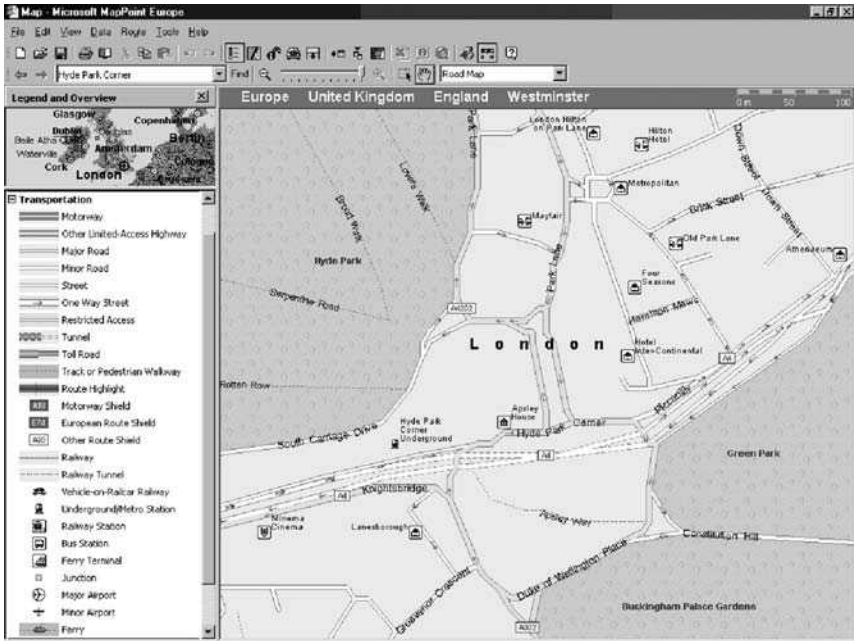
of the Earth and the shapes of the continents were thus projected onto sheets of paper wrapped around the globe as a cylinder or cone, or just placed tangentially at some point on the surface. These map projections are worked out mathematically so that coordinates in longitude and latitude can be transformed into one projection or another. Thus in Figure 3.10b the square area marked, when transformed to the Mercator Projection in Figure 3.10c become a weird fan-shape. Choose a different projection and it is likely that a different shape will result. As one zooms into smaller and smaller areas this progressively becomes less of a problem, as very small areas (such as Figure 3.11) can be treated for all intents and purposes as flat. Nevertheless, projection is a fundamental property of any map and for any datasets to be displayed on the same map they must all be transformed not only to the same coordinate system but also to the same projection. Usually GIS software has functionality to achieve this.

### 3.4.3 Symbology

Symbology refers to the graphical way features are represented on maps. This is very much an issue of design: how to make clear, attractive maps that communicate the necessary information (see for example Tufte (1983) for issues of graphic design and Brown and Feringa (2003) for issues of colour). Interestingly, national differences are clearly discernable in choice of colours and symbols in map making, often *de facto* standardized by major map producers (often the Mapping Authority, but not always) possibly reflecting a synergy with the national psyche. Compare for example a North American style of representation in Figure 3.1 with more of a French style of map design in Figure 3.2. This is important in LBS as it means that there isn't a one-size-fits-all approach to map design unless the wish is to impose a form of cartographic, cultural imperialism on all users. Mapping Authorities, for example, have traditionally imposed on their customers the range of maps, content of maps and design of maps. GIS software, in varying degrees, allows users to design their own maps based on taste and what needs to be communicated (Section 3.4.6). This does not mean that the results are always good! Students often produce garish maps with clashing colours, inappropriately sized symbols and altogether too cluttered. Another consideration for LBS is the use of language: French maps have their legends in French, German maps in German. Not surprising, of course. But if, say, Dr Li travels on holiday through Germany would she, as a non-German speaker, want the results of her LBS queries all to be in German? Of course not, but they may well have maximum utility if they were dual language, knowing that a *motorway* is labelled or signposted as *autobahn* is clearly helpful.

The use of symbology to convey meaning is illustrated in Figure 3.11. The legend to the left shows all the ways transport features are represented: different grades of road, road numbering, railway lines and points of interest such as underground station and bus station. The map also shows additional features such as parks, hotels, a cinema, car parks and an historical monument (Aspley House). Each of these features has its own distinct symbol. Some features are also labelled with their name. Whilst this map may give the impression that this is a relatively uncluttered area, this is in fact one of the busiest and most expensive parts of London where Mayfair and Knightsbridge meet across this narrow corridor between popular parks and the Royal Family's back garden. In GIS, fundamental





**Figure 3.11** Map symbology – representing features around Hyde Park Corner, Central London (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

decisions need to be made about what to include and what to leave out, because any attempt to include every feature in an area inevitably produces very large data sets. This process of deciding what features are to be represented in a database and at what level of detail is called *data modelling*. Thus, whilst the data modellers of the database underpinning Figure 3.11 have decided to show just about every road, very few of the buildings have been included. This reflects one of the main purposes of Microsoft MapPoint, which is route finding and navigation by car from one place to another. It also reduces the size of the database to what can be reasonably loaded into a PC and sold at an economic price. In other words, in this case, it's largely a marketing decision. Other GIS databases are data modelled around other applications and will include a selection of features designed to make it fit for purpose.

Comparing Figures 3.9a, 3.9b and 3.11 illustrates another issue of map creation using GIS, which is the tension between scale, degree to which features are generalized and the way symbology is

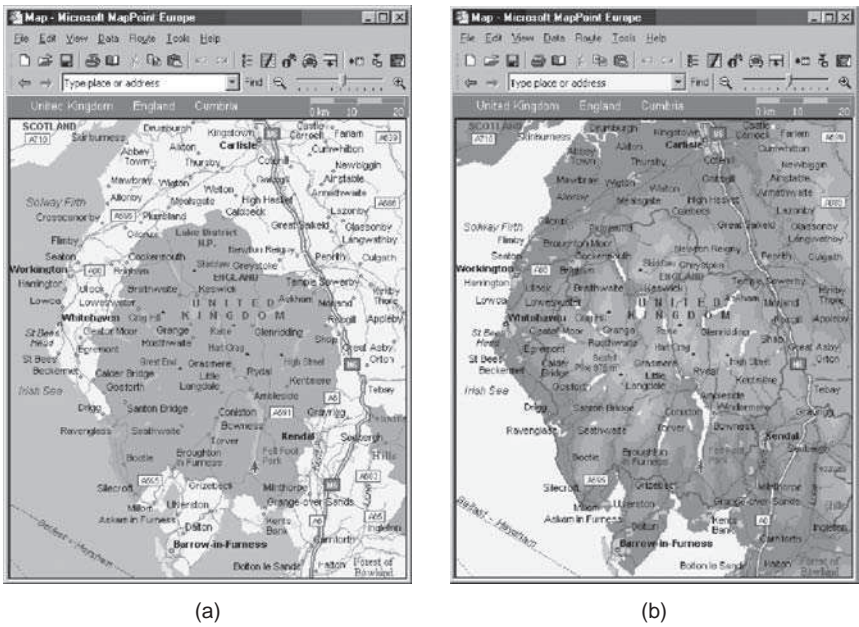
used to give a useful representation. Thus in Figure 3.9a, only the main urban areas and major roads are shown on a topographic backdrop. Compared to Figure 3.9b, the M25 motorway circling London is shown in a very simplified (highly generalized) way in Figure 3.9a. However, whilst the M25 in Figure 3.9b has all the junctions shown, its size has been exaggerated – the width as shown on the map when compared with the scale bar would suggest that it is 1 km wide! This is so as to make the motorway clearly visible amongst all the other features. All roads that are not a motorway are represented as lines. Yet on zooming into Figure 3.11, nearly all the roads have been represented by a double line to give them width. The symbology used to represent features has been changed depending on map scale so as to generalize on zooming out and to increase specification on zooming in.

#### 3.4.4 Data Primitives and Data Layers

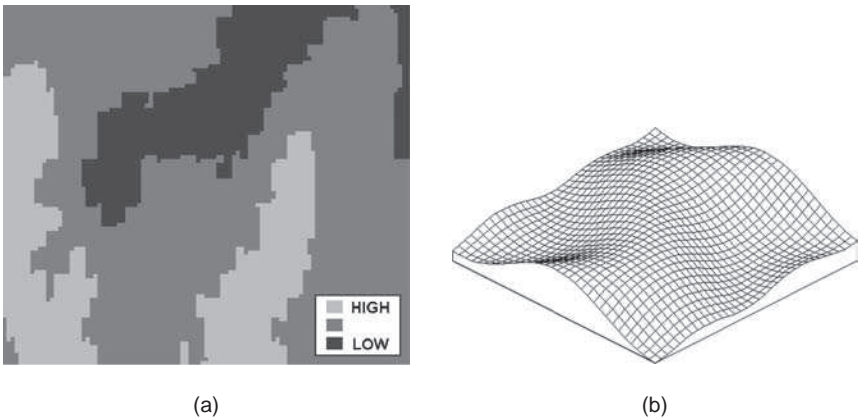
GIS databases are constructed using four spatial data primitives: point, line, polygon and cell/pixel. These can be seen in Figure 3.12 and Figure 3.13. In Figure 3.12a all the towns and villages are shown as points using the symbols  $\square$  or  $\circ$ ; another point feature is mountain peaks shown using the symbol  $\blacktriangle$ . Obvious line features are the roads which link together to form a network. A polygon is a series of lines which joins back onto itself to enclose an area feature. A large polygon in the centre of Figure 3.12a is the shaded area of the Lake District National Park. Points, lines and polygons are used to depict discrete objects, but the choice of which to use will depend largely on scale. Thus a town in Figure 3.12a is represented as a point as this is a small scale map. Zoom in somewhat and it would probably be more instructive to show a town using a polygon that describes its outer limits. Zoom in even further and the town would be shown as a series of street blocks implied by the street network or even to show individual buildings as polygons. Again, a river at small scale would be represented as a line, but at large scale by a polygon showing both its banks and the area occupied by water. Data stored as points, lines and polygons are commonly referred to as *vector*.

The fourth spatial data primitive, the cell, is used to represent continuously varying features. An example of this is the topography shown in Figure 3.12b, which depicts the mountainous landscape of the Lake District National Park. An enlargement in Figure 3.13a

# Location-Based Services and Geo-Information Engineering



**Figure 3.12** Points, lines, polygons and cells: (a) road map view of the Lake District, Cumbria; (b) topographic view (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).



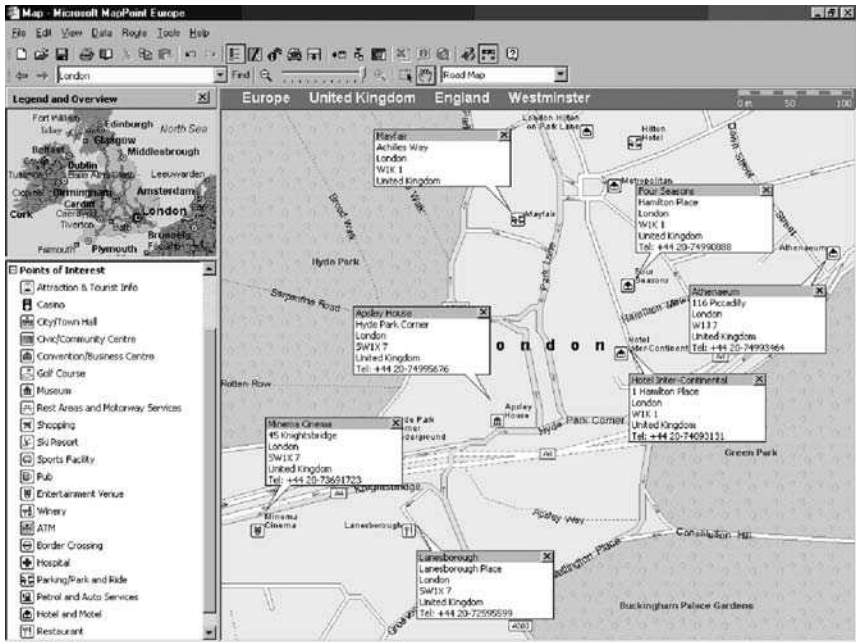
**Figure 3.13** (a) enlargement of topography to reveal its grid cell structure (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved); (b) 3-dimensional view of a grid-based topography (from Brimicombe, 2003).

reveals the cell structure, in this instance a grid, in which each cell is given a height ( $z$ ) value according to the patch of topography it covers. A typical three-dimensional view of grid-based topographical data is given in Figure 3.13b to further illustrate the concept. Data arranged as a grid are commonly referred to as *raster*. However, square grids are not the only shape of cell that can be used. Grids are part of a family of space-filling *tessellations* with other possible shapes including triangles, rectangles, rhombi and hexagons. The issue of how vector and tessellation data are stored efficiently in a database is held over to Chapter 8, when querying of spatial data is discussed.

It is a general principal of GIS data structures that features described using points, lines, polygons and cells are kept separate, that is each feature type is stored separately and are displayed as *layers*. Thus in Figure 3.12a the towns and villages have been stored as a point layer, the roads as a line layer and the National Park polygon in its own layer. In Figure 3.12b the settlement and road layers are again displayed, but this time on top of the topography raster layer. This approach makes life much easier not only when constructing databases but allows for mixing and matching of layers, and features that overlap, in versatile ways for any analysis or map display. It should be noted in relation to Section 3.4.2 above that all layers being displayed together must have the same coordinate system and projection or else they will not superimpose properly.

### 3.4.5 Feature Attributes

Thus far, only the geometric representation of features has been considered. Clearly, from the maps included above, places are named and roads are labelled or numbered. So there are other data attached to the geometric representation of a feature that allow us to know more about it. In GIS these are known as *attribute* data; they are stored in a conventional database and can be linked to any point, line or polygon. Returning to Hyde Park Corner in the centre of London (Figure 3.11), there are a number of hotels and other buildings that have been recorded as point features and represented by icon-type symbols. Figure 3.14 shows the same area with a number of these points of interest (POI) having been clicked to reveal their further attributes. In Microsoft MapPoint the default attributes for POI are the address and telephone number, but the user can edit this on screen and can add their own new POI with relevant attributes.




**Figure 3.14** Points of interest (POI) with their stored attributes (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

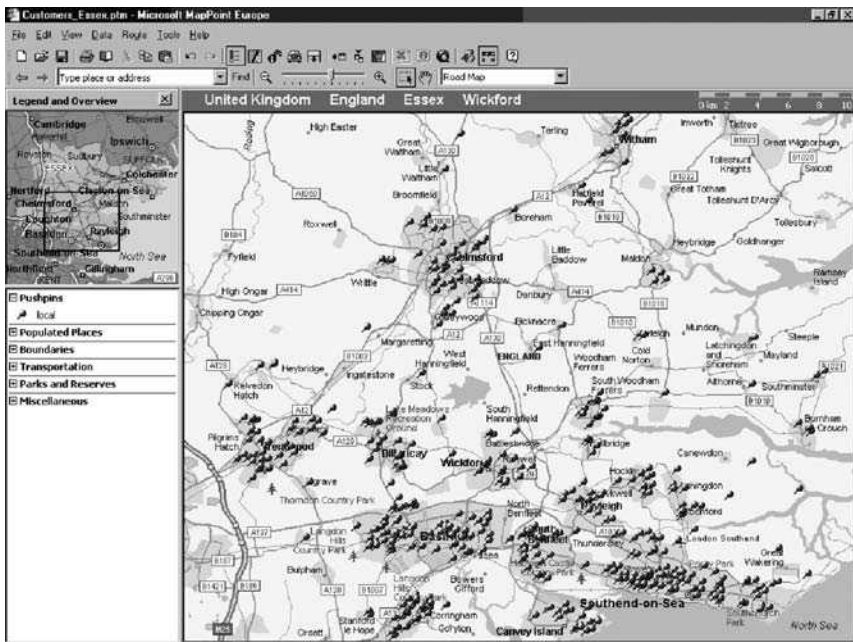

New tables of attributes can also be imported into Microsoft MapPoint from external databases, spreadsheets or text files. Any imported table should include a *primary key*, that is a column containing a unique identifier for each record. Examples of a primary key could be a customer ID, postcodes or names of administrative areas. A second requirement is that at least one column, the *secondary key*, should contain a relevant geographical identifier that will allow the attributes in the table to be joined to the graphical features in the MapPoint database. Examples of these could be city names, district names or postcodes. A secondary key must be able to match with the primary key in the table to be joined with, otherwise no join can be achieved. All columns should be labelled. A small clip from a customer database in spreadsheet format is given in Table 3.1. Customer ID is the primary key and postcode is the secondary key and also acts as a geographical identifier, the other three fields are some attributes for each customer. Once imported into Microsoft MapPoint a join with the internal spatial database can be achieved through the postcode field and x, y coordinates can thus be associated with each customer.



**Table 3.1** An extract of customer data prepared in a spreadsheet for importing into Microsoft MapPoint.

	A	B	C	D	E
1	Customer ID	Product type	Customer type	Spend £	Postcode
2	205	3	2	64	CM0 7AF
3	206	1	1	42	CM0 7BA
4	207	1	2	191	CM0 8DR
5	208	2	2	287	CM0 8ED
6	209	2	1	479	CM0 8TL
7	210	2	2	860	CM1 1RF
8	211	4	1	250	CM1 2EG
9	212	2	1	115	CM1 2SW

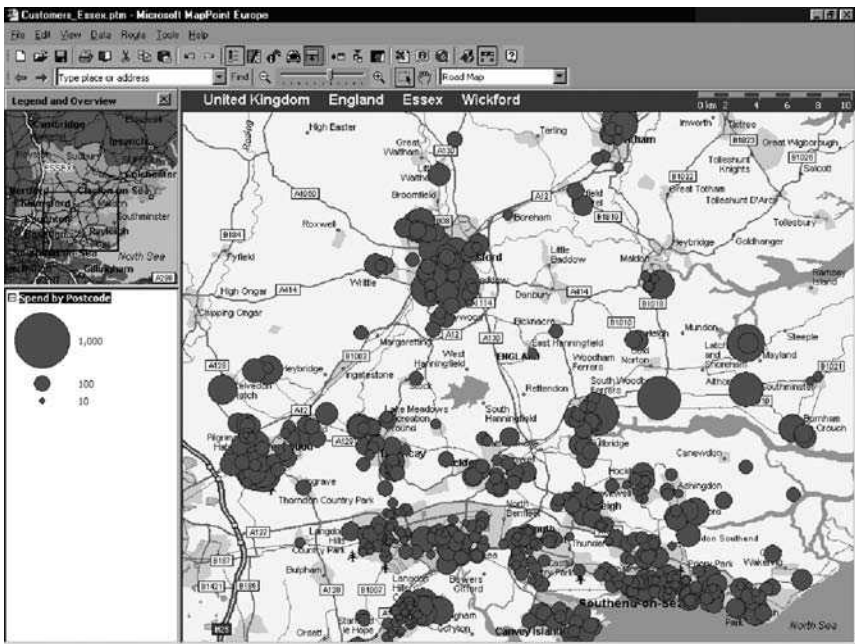
In this way, all the customers can be plotted (Figure 3.15 – each customer is shown with a pushpin  symbol), or, as is shown in Section 3.4.6 any of the attached attributes can be spatially analysed

**Figure 3.15** Postcode locations of customers (from data set in Table 3.1) as shown by pushpins  (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

and mapped. This is one of the great strengths of GIS – large numbers of different attributes can be attached to the same geographical features and analysed in terms of their spatial properties as well as their statistical properties.

### 3.4.6 Creating Thematic Maps

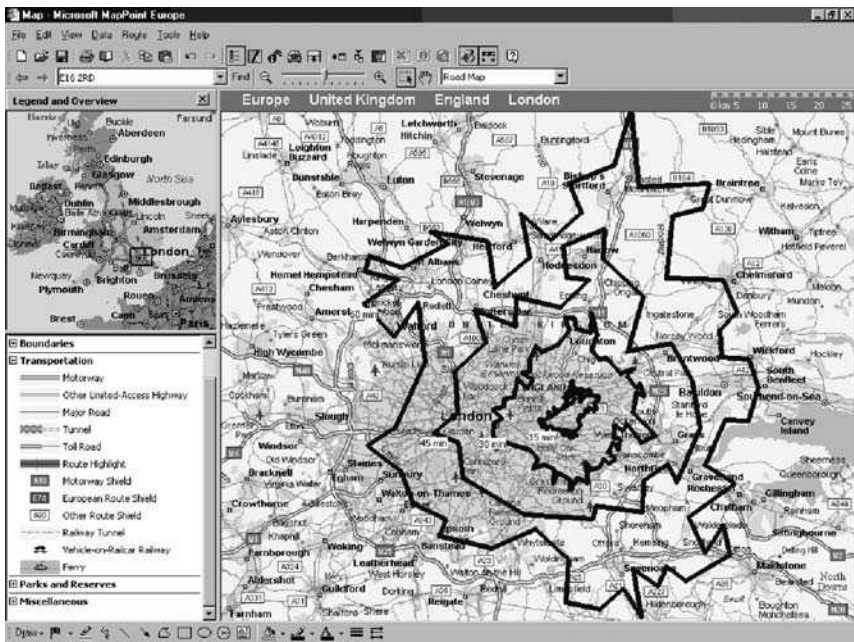
Thematic mapping includes a number of steps for the production, layout, visualization and printing of maps that depict particular themes, more often than not based on attribute data. The production of good thematic maps is both art and science and is by no means a trivial activity. Most GIS software include wizards to assist users through the various steps and choices that need to be made around colours, patterns (or textures), class intervals, symbology and symbol sizes. Numerous examples of thematic maps can be found on the Web, but it is recommended that the reader have a look at <http://www.magic.gov.uk>. By way of an example here, Figure 3.16 shows the customers located in



**Figure 3.16** Thematic map of customer spend shown as proportional circles  
(Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).

Figure 3.15 mapped as proportional circles indicating the amount of spend for each customer. Clusters of low spending customers become obvious, as do the locations of large spending customers.

A particular form of thematic map that is particularly relevant to LBS is route planning and the calculation of drive time zones. Examples of route planning have already been seen in Figure 3.3 and Figure 3.4. In Figure 3.17 an example is given of drive time zones. These are in 15 minute intervals from the University of East London, Dockland Campus. To run these calculations, the road data must be structured in the database as a network or planar graph such that the way roads (as lines) join one another is explicitly modelled through the junctions (as points where roads intersect). This type of data structure is discussed further in Chapter 8 when the details of spatial queries are considered. The search algorithm can identify either the shortest distance route or the route that takes the least travel time. Other variants include having the least number of route changes or preferentially choosing the fastest roads. For least travel time, it is



**Figure 3.17** Fifteen minute drive time zones from University of East London, Docklands Campus (Copyright 2003 Microsoft Corporation and/or its suppliers. All rights reserved).



necessary to know something about the speed at which a vehicle will travel along different classes of road (motorway, major road, minor road). This could be the maximum permissible speed for each class of road but more realistic would be an average speed to account for traffic lights, slowing down at roundabouts and so on. Usually there are default values, but a pop-up menu allows the user to tailor these speeds according to their own preferences based on factors such as type of vehicle (e.g. van vs sports car) and the likelihood of meeting heavy traffic. For the drive time zones, these speed parameters are used to calculate the time taken from a single start point along all parts of the road network emanating from that start point and then connecting up all the points reached for a pre-determined time interval (say every 15 minutes) in order to create zones such as in Figure 3.17.

Most GIS software is capable of being automated or having its functionality extended either through the use of purposely designed scripting languages (e.g. Avenue for ArcView and MapBasic for MapInfo) or by using more generic programming languages such as Visual Basic. Thus MapPoint has its own object model that allows a programmer to access and manipulate the map data so as to avoid having to go through all the steps manually, or allowing a programmer to create new applications including dialogue boxes, imported attribute data and new algorithms. All the examples of functionality seen above can be programmed into macros using the object model. Thus, for example, the statement `object.AddDriveTimeZone (Centre, Time)` can be used to create a drive time zone as in Figure 3.17. Furthermore, by using a Visual Basic editor the MapPoint object model can be accessed and embedded into new applications using the ActiveX Control. This is a way of integrating, for example a spreadsheet with map visualizations of the data within a single custom-made application.

### 3.4.7 Scaling the Applications

GIS software can run on a number of different hardware platforms depending on the context and scale of the applications, where scale in this instance refers to the size of the database(s) and number of users accessing the system. Thus GIS software can run on platforms from PDA and tablet PCs through to 'server farms' (large number of interconnected servers) and super computers. This is not to be confused with viewing maps in generic file formats such as JPEG or GIF, which can be achieved on virtually any technology with a graphic screen. Of course the level of functionality in the GIS software is

scaled down as the platforms become smaller or thinner, but is nevertheless still recognizable as interactively handling spatial data. PDA versions tend to be used for field data collection (the checking of features and attributes) or assisting in the fulfilment of mobile activities such as deliveries. At the other end of the scale is large corporate systems or enterprise GIS. In these systems the attribute data may become very large and subject to frequent updating to reflect the business transactions. In these cases the attribute data may be stored separately from the object geometry in a relational database management system (RDBMS). The GIS and RDBMS then work together through an open database connection (ODBC). Market leaders in enterprise RDBMS such as Oracle have evolved their systems to accommodate spatial data. So for example Oracle Spatial is an extension of the standard product to include an object data type (SDO\_GEOMETRY) that allows points, lines and polygons to be stored as well as providing indexing capabilities (Chapter 8) and query tools that address the spatial dimension of the data. Is Oracle Spatial, then, an enterprise GIS? Not really, as it is only the database component (though very powerful and able to handle millions of records) and relies, for example, on GIS software for the visualization of the map data and the creation of thematic maps. However, in recognition of its strong spatial capabilities, in later chapters this and similar database technologies are referred to as *spatial databases*.

As shown in Section 3.2, Web delivery of maps and basic GIS functionality has become very popular indeed, with applications varying widely from route finding to viewing neighbourhood house prices and crime rates. These client-side applications use Java applets or dynamic HTML to display maps on a browser, albeit with some interactivity for layer switching, zoom and pan, but otherwise quite restricted in their functionality. Server side there is GIS software and/or spatial databases to run the application. The market leaders in GIS software such as ESRI (for ArcGIS; <http://www.esri.com>) and MapInfo (<http://www.mapinfo.com>) have add-on modules for managing Web-based applications. Queries submitted through a browser can be processed on a remote server but this usually requires middleware that services the interaction between GIS software and the client Web browser. Given the generally long transaction time for GIS queries (discussed in Chapter 8) and the limitations of using middleware as well as the processing limitations of Web browsers, the time is ripe for moving GIS software away from large, expensive software packages towards distributed components (Peng and Tsou, 2003).

The Internet has provided unprecedented capabilities for searching the Web and downloading information in the form of Web pages as well as browser functionality in the form of plug-ins and applets. Whilst these plug-ins and applets add to the functionality of a Web browser, it is also possible to download client-side executables that run independently of a Web browser. These are *distributed components* and use agent-based technologies (Section 3.3.2). Although still some way off, specifically packaged aspects of GIS functionality could be made available as distributed components, *GI Services components*, that could be downloaded (and if necessary paid for!) on demand. This would allow a scaling up from accessing static maps over the Web to accessing increased levels of GIS functionality by communicating directly over the Internet with server-side GIS. More importantly from an LBS perspective this approach would open up opportunities for more sophisticated interaction with spatial data from mobile devices. Whilst the GIS industry appears to endorse the move towards mobile GI Services components, current business models have directed the focus towards sharing data from multiple sources rather than towards sharing the software components. Nevertheless, GI Services components are an important and growing area of research.

### 3.5 Issues of GIScience

---

In Section 3.4 the focus was on the key aspects of GIS as technology. GIS are now well established in many application areas for handling spatial data. As with other technologies, in order to progress to mainstream there needs to be an underlying science – a body of rigorously investigated knowledge – that underpins developments and uses of the technology. As with many rapidly developing new technologies, the cart came somewhat before the horse. GIS as technology had already had a thirty year history by the time Goodchild (1990, 1992) launched a debate for the recognition of a coherent GIScience. He put forward two main criteria for its recognition as a particular branch of science:

- that the knowledge domain contained a legitimate set of scientific questions;
- that spatial data were unique from other forms of data and therefore needed special consideration.

Goodchild argued that these two criteria were indeed satisfied, the distinctiveness of GIScience resting on:

- the use of the spatial key  $\{x, y, a_1, a_2, \dots a_n\}$ , where  $\{x, y\}$  define *location* as continuous dimensions and  $\{a_1, a_2, \dots a_n\}$  define the *attributes* of location either as continuous or discrete dimensions;
- the presence of *spatial dependence* between locations in that near things are more likely to be similar than distant things;
- the durability of the spatial *data primitives* of point, line, polygon and cell/pixel that have underwritten the technology and its application in many diverse applications (Burrough, 2000).

The above three points argue more for spatial being special, leaving the legitimate set of scientific questions to be articulated. One way of gauging what these are is to look into the number of academic journals, such as the *International Journal of Geographical Information Science* (since 1986), *Transactions in GIS* (since 1996) and *Journal of Geographical Systems* (since 1999), which focus on publishing the results of research that can be classed as GIScience. A review of their contents pages would illustrate the breadth of scientific questions currently occupying GIScience researchers. There are, however, many other established disciplines where research into aspects of GIScience is taking place within the context of those disciplines and the results are published from time to time in a wide range of journals from *Mathematical Geology* to *Social Sciences & Medicine*. GIScience is not that neatly summarized, though the collection of papers in Duckham *et al.* (2003) give it a plausible go. One important reason is that as questions and issues are either solved or better understood so the research agendas evolve and GIScience, as a body of knowledge, continually moves on. Agendas have been successively articulated by Rhind (1988), Goodchild (1992) and Mark (2003). Agendas specific to GIScience and LBS are to be found in Brimicombe and Li (2006) and Jiang and Yao (2006) – and of course in this book (e.g. Section 4.8).

At the heart of GIScience is the production and dissemination of information, from spatial data, that has utility to society. From this perspective LBS are intertwined with, and a legitimate research area of GIScience. But the process of transforming data into useful information, whilst easy to say, has a number of key aspects which, whilst it is always difficult to subdivide an integrated

knowledge domain into discrete chunks, can be considered to fall under the following headings:

- **Ontology:** this deals with what exists, or in this context, what exists within the knowledge domain of GIScience and how it relates to user groups and society at large. This is not limited to arriving at formalisms that structure terminology and conceptualizations within the knowledge domain but has relevance to the consumption of the information. In other words, ontologies are also at the heart of communicating and sharing information. Supposing we take the object *river*; nearly everybody would say they know what a river is, if you understand the term in English of course. Languages can cause data integration and interoperability problems, but we can use a dictionary or an expert to derive a mapping of terms such that *river* in English is *fleuve* in French. But the French also commonly use the word *rivière* which might lead to confusion or uncertainty. Then again, when is a *river* sufficiently small to be only a *stream*, and smaller still to be only a *rill*? Then again, when does *near* become *far*? It is often at the boundaries of classes that maximum uncertainty arises and the interpretation of classes and their boundaries shouldn't be left to chance. Thus ontology permeates GIScience from establishing data models, through issues of interoperability and uncertainty to the semantics of communicating spatial information. For further reading on this topic, the reader is referred to Visser *et al.* (2002), Smith and Mark (2003) and Uitermark *et al.* (2005).
- **Representation:** this deals with how geographical data and phenomena are encoded and stored in a digital form. The term 'phenomena' implies a move beyond static snapshots to considering the processes of dynamic change. Included within this heading then are: data modelling, that is establishing which features of geographical reality are relevant to an application; issues of scale and the degree of resolution (granularity) that is appropriate to an application; data structures, in which the data are organized as data primitives (Section 8.2), topology (Section 8.3) and attribute data (Section 8.4) and includes the implementation of efficient indexing and retrieval (Section 8.5); how the time dimension is to be represented (Section 8.6) and how boundaries between features and

between processes are to be handled (Burrough and Frank, 1996; Goodchild *et al.*, 2007).

- **Computation:** much of GIS revolves around computational geometry in the way data primitives are handled during data transformation, integration, analysis and visual presentation. GIS requires efficient solutions to the geometric problems encountered such as in generalization (Section 9.3.3) and in ensuring data layers are consistent with each other (*conflation*) so that, for example, roads in one layer do not run through buildings in another. Computational geometry is also relevant to building topological data (Section 8.3) that provide added intelligence to geographical data and help speed spatial queries. Computation is also relevant from an interoperable perspective in which GIS are often used alongside other computation technologies such as databases, spreadsheets, statistical packages, artificial neural nets and multi-agent systems. This is now referred to as *geocomputation* (Longley *et al.*, 1998; Openshaw and Abraham, 2000) and is central to the way GIS are used in technologically heterogeneous applications such as in environmental simulation modelling (Brimicombe, 2003).
- **Cognition:** this concerns the way humans perceive, learn, communicate and reason about geographical information. This is not just how geographical knowledge is learnt and its use in wayfinding (Section 9.5) but also how through our ability to assimilate, synthesize and interpret geographical information we both generate knowledge and are able to make decisions around problems that are inherently spatial. This then extends to how GIS are used to implement spatial decision support systems (SDSS) and participative systems for spatial planning. From a different perspective, cognition is also concerned with how spatial queries are framed – to interrogate spatial data – and the human–computer interaction (HCI) necessary to do so.
- **Uncertainty:** this issue lies at the heart of GIScience with an active research agenda since the early 1980s. This concerns not only the quality of the spatial data used in GIS (Section 5.4) but also the fitness-for-use of the informational products that emanate from GIS analyses. Errors and ambiguities in data will propagate through analyses and may become amplified in the final outputs. Errors and ambiguities in the

information may in turn lead to errors in its use (Section 10.5.3). Important tools in handling uncertainty are the production and use of *metadata* (data about data) to inform data usage and validation of models, including calibration and sensitivity analysis to assess fitness-for-use.

- **Spatial analysis:** there is a class of problem involving mathematical or statistical summarization and manipulation of data in which spatial dependency becomes an important issue. Spatial dependency derives from Tobler's First Law of Geography (1970) in which '... everything is related to everything else, but near things are more related than distant things'. In other words, geographical objects tend not to be independent of nearby objects, or put another way, we tend not to find random physical or anthropogenic landscapes. This leads to a consideration of how spatial dependency is measured through indices of spatial autocorrelation (Cliff and Ord, 1981) and the semi-variogram. Spatial dependency is also present in the modifiable areal unit problem (MAUP; Openshaw and Taylor, 1981), underlies algorithms of spatial and areal interpolation, and in cluster analysis.
- **Visualization:** maps have been the traditional principal means by which geographical information has been portrayed and communicated. Key issues have been scale-related generalization and use of colour and symbology in the design of thematic maps. In a digital environment, maps are still an important medium, but not only are there additional design considerations for the Web and mobile devices (Section 9.3.2), they may also need to be integrated with other types of graphics (graphs, images) and sound to become multimedia presentations. Digital also opens up the possibilities for interactive maps, tactile maps, 3-D representations, augmented reality and virtual reality.
- **Institutions and society:** this concerns the impact that GIS have had on society, from spatial decision support systems and participative GIS, to major applications such as Web-based cartography and LBS. Other issues that come under this heading are business models, legal issues and standards (Chapter 10). There have also been debates around the ethics of GIS (Pickles, 1995), such as the use of GIS in war and as a key technology in building a surveillant society (Section 1.4).



### 3.6 GI Engineering: the Rise of Ubiquitous GIS?

---

GI Engineering can be defined as *the design of dependably engineered solutions to society's use of geographical information*. These solutions naturally build on GIS and GIScience but may also be in conjunction with other technologies and other branches of science. They are likely to be technologically heterogeneous solutions. What sort of things are we talking about here? Perhaps it is better to start with a couple of analogies. Firstly, the mobile phone: as was seen in the last chapter the heterogeneous systems that go to make mobile phones work are complex, let alone the electronics and software in the handset itself. Do we worry ourselves about all these components, does the average person in society need to have studied all aspects of their complexity? Of course not. We switch on the handset and as long as we can 'get a signal' we expect it all to work – seamlessly. And the handsets are designed to be intuitively simple (unlike video recorders were!) so that even children can quickly work out how to use them.

The second example is lifts (elevators) in buildings. We walk in, press a button for a floor, wait until the doors open again and walk out. Again, we expect them just to work – we don't need to be structural engineers in designing the shafts, nor electrical or mechanical engineers in designing the carriage and lifting mechanisms. We need know none of these things to operate and gain utility from a lift. Granted they can breakdown, but if properly serviced by an engineer (whom we rarely meet; we, the users, have little of the specialist knowledge required to carry out such a service) failures are minimal and rarely cause injury.

We take mobile phones and lifts (and many other technologies) for granted in our daily lives. So why not with technologies that use geographical information, or are associated with society's use of geographical information. GIS students are often horrified at this prospect. All that specialist knowledge needed to properly understand and use GIS and spatial data – you couldn't possibly allow the public just to do it for themselves! But if only lift engineers were allowed to use lifts because they were the only ones who truly understood how they operate, the market for lifts (which after all are expensive to install) would be very small indeed. The challenge is to design products that use GIScience and GI Technology that are easy to use and don't particularly require any specialist knowledge in order to be used effectively by the public.



So what will it take to engineer GI into the mainstream with widespread use within society? In a way it's been achieved once before in an analogue era through the production of paper-based atlases, maps and A to Z type street gazetteers. But in a digital age these no longer seem enough, though such products are not entirely obsolete. Firstly, it's no longer enough to know, for example, where a cinema is; we also want to know what's on and when without consulting a number of different sources. In other words we want increasing information richness as well as finer granularity (more detail) that is seamlessly packaged as an easy-to-use one-stop shop. This makes the information package much more dynamic with shorter 'shelf life'. We also seek more customization of information to meet our particular contextual needs. Web developers, for example, have tackled some of these issues through the use of dynamic HTML and hypertext processors whereby customized, updated Web pages can be created in real-time. Finally, we would like to be able receive information wherever we are in either static or mobile situations and, increasingly, to wireless-enabled mobile devices with roaming capability that are easily carried around with us. To meet these digital age demands, what needs to be the focus of the engineering whether achieved through research, borrowing from other disciplines or just sheer pragmatism? Here are some important candidate areas:

- **Evolving ICT:** the miniaturization and convergence trend discussed in Chapter 2 introduces challenges not only of processing capacity on thin, remote clients but also when considering that, conventionally, GIS have been designed to deliver high resolution graphics on large monitors. Even paper maps tended to be large. All this now needs to be delivered to very small screens. GIS are typically large data applications requiring high processing power. We may have to wait until Moore's law (Chapter 2) progresses a bit further to provide the necessary memory and processing power on small enough chips, but fully-fledged LBS may yet have to wait for the bandwidth of 4G wireless technologies in order to have the necessary data throughput to support applications.
- **Scalability:** for ubiquitous applications, GIS will need to be scaleable to potentially millions of users without noticeable degradation of service. Clearly open and interoperable middleware will be important, but key to scalability will be

disaggregating the current approach to full functionality GIS into interoperable distributed component GIS using agent-based technologies. This would make GIS very light and portable without loss of functionality, as individual component functions could be downloaded on demand either automatically by the core component or depending on a user's preference in a component market (i.e. shopping around).

- **Syntax:** this refers to data structures. Current data structures tend to be monolithic and insufficiently flexible for mass usage. Whilst tiles of data might be downloaded – each new tile accessed as needed – each tile has within it the same granularity of data and this can be sub-optimal in terms of response time and the capacity of mobile devices to process. Some research has been carried out into data structures that can support variable granularity (Tsui and Brimicombe, 1997; Worboys and Duckham, 2006) with more detailed data where the focus of user attention is, say, on a particular set of roads that constitute a journey and less detail the further the distance is from that focus of attention. Additional detail in peripheral areas can be augmented on demand if the focus of attention shifts.
- **Data:** as will be discussed in Chapter 5, spatial data are but one source of information and the handling of spatial data will need to be seamless with other types of data. GIS will have to be integral to a multimedia approach to information delivery. There will also be advantages in being able to spatially reference fuzzy geographical footprints of photographic images, text (books, newspaper and magazine items), voice and music recordings, video/DVD and virtual reality scenes so that they become geographically searchable. Spatial data will also need higher granularity and more frequent update cycles and this will be dependent on developments in data collection technologies and the systems within which they are deployed (e.g. telegeoinformatics, Section 5.3.4).
- **Response times:** these will need to be speeded up. As will be discussed in Chapter 8, GIS queries have both high CPU costs and high Input/Output (I/O) costs, in other words, that are comparatively long transactions. Whilst some of the issues discussed under scalability and syntax are likely to speed up response times, more research will be required into how geographically naïve natural language requests are parsed into

database queries, indexing and fast retrieval of spatial data, efficient analytical algorithms and the structuring of phased responses through query sequencing and optimization.

- **Cognition:** this is central to how users of ubiquitous geo-technologies are going to be able to interact with them. There will need to be a better understanding of the reasoning behind the initiation of spatial queries (requests for information), spatial memory and learning, and working with naïve geographies and natural language. There needs to be more research into human-computer interaction with mobile devices as a means of accessing spatial information and into user preferences for information delivery (Section 9.5.2). This includes capturing and understanding user context(s) as an important means of adding utility to spatial information (Chapter 7).

The prize for the GIS community in embedding GI Engineered products into society is to become a mainstream technology, part of society's infrastructure. GIS would achieve its long-sought ubiquity (Li and Maguire, 2003; Sui, 2005) perhaps sharing the glamour of mobile phones but perhaps only with the humbleness of the smoke detector – there to reduce our anxiety. There are other implications. It has been argued that 'the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it' (Weiser, 1991 p. 94). What does it mean for GIS to 'disappear'? In a way SatNavs provide an example. It is not that maps or devices completely go away from view but rather the level of knowledge and skill needed to use such devices is well within the average person's grasp; what has disappeared then is any need to know the algorithms, database structures or location technologies (GPS) necessary for the internal working of a SatNav. When SatNavs are routinely installed in vehicles at the point of manufacture and no vehicle comes without one, a piece of GI Engineering will have then woven itself into the fabric.

Another example is provided by the use of postcodes or zip codes. These are small mnemonics that capture geographical location and some broad structure of geographic location (spatially hierarchical with in-code and out-code portions). These are now firmly embedded in the commercial and administrative mainstream such that they have now become an indispensable way of organizing information by geography and locating people and facilities within small

areas. This is also a good example of social learning and adaptation in the area of technological innovation (Williams *et al.*, 2005; Section 10.2). The postcode initially was designed as a device to assist the efficient delivery of mail. Whilst postcodes are still used for that purpose, society has hijacked them for other purposes not thought of at the outset. The mobile phone was originally conceived for business users only, but society has made them into must-have accessories even where, for some sectors, voice communication is secondary to texting. GI Engineers will have to be prepared for unexpected outcomes in the way society wants to use their products.

Finally, if GI Engineered products are to become part of the ‘fabric of everyday life’, then we can expect increasing legislative oversight with requirements to adhere to new legislation and develop new standards (Section 10.5). Not only must GI Engineered products be reliable and safe, but they also pose important social and ethical issues (Curry, 1995; Onsrud, 1995; Haque, 2003) around possible misuse of spatial data or even heightened vulnerability arising from the use of GI Engineered products (Section 10.6).



# Chapter 4

## Location-Based Services

### 4.1 Introduction

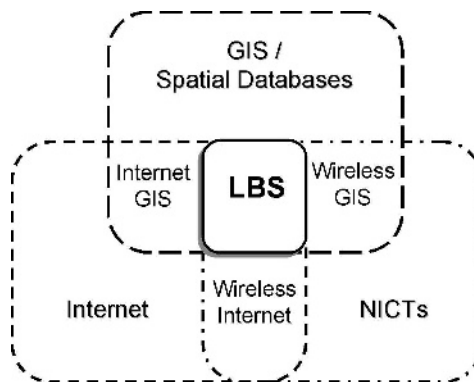
---

It is worth starting this chapter by re-stating our definition of LBS from Chapter 1:

*Location-based services (LBS) are the delivery of data and information services where the content of those services is tailored to the current or some projected location and context of a mobile user.*

And it is also worth repeating the Venn diagram from Chapter 2 as it positions the focus of this chapter in relation to its key facilitating technologies (Figure 4.1):

In this chapter, an overview of LBS is taken: what their initial drivers were, what their broad architecture is, what is their current



**Figure 4.1** The technological niche of location-based services (from Brimicombe, 2006).

state is, and what areas need the attention of GIS researchers. In doing so, it raises some important themes that will each form their own chapter for the remainder of this book.

### 4.2 Are Location-Based Services New?

---

As with a lot of things, once we start doing them in a digital environment, we think it's all new and we forget that very often we are supplanting something that was analogue. Certainly going digital affords us the opportunity of speeding up, expanding and diversifying to an extent that could only have been dreamed of in the analogue days, but still the antecedents were analogue. So what were analogue LBS? Well, not to stretch a point too far we could go back to Roman times. In building their network of roads across Europe, they introduced milestones to provide information (distance to or from a military base) tailored to the location of the soldier or citizen along the road. Though directional signs showing which way to go from one town to the next started to be introduced in medieval times, it was only with the popularization of internal combustion engine automobiles from the late 1880s onwards that whole classes of location-based services sprang up. In Britain two service providers established themselves: the Royal Automobile Club (RAC) in 1897 and the Automobile Association (AA) in 1906. Their services could be broadly categorized as:

- signage (warnings and signposting) and help in wayfinding;
- rescue from breakdown or running out of fuel;
- information for the avoidance of police speed traps.

The phase change from walking speed to break-neck speed meant that drivers needed forewarning of hazards or changes of direction to come, in time and right there on the side of the road where it was relevant. In case you needed more information or needed the breakdown service, then there were phone boxes specifically sited by the AA and the RAC (Figure 4.2). It was only in the 1930s that local authorities took over responsibility for road signage (though the AA continues to place temporary signs for special events and specific locations they deem insufficiently well signposted) and it was only the ubiquity of the mobile phone that eventually rendered most of the special phone boxes obsolete.





**Figure 4.2** A vestige of analogue LBS on a mountain pass with poor mobile phone reception. (photograph by the authors).

#### Prof. Brimicombe's journey to work:

I live in a village half way between London and Cambridge, it's Monday morning and yesterday I was writing the part of Chapter 4 you have just read. Today I've decided to count the number of directional road signs into work. I know the way but I'm intrigued by our analogue LBS infrastructure geared to wayfinding, the legacy if you like to what the RAC and the AA started over a century ago. However, I'm not counting all the Highway Code warnings and advisories; I'm only interested in all the signs, specifically located at the road side, that guide us where to go (a fine example is in Figure 4.3 that I stopped and photographed along the way). I've driven 2 km from my house through the village; I'm on my way to the next village and I've already passed 11 directional signs offering me choices to this place and that. By the time I've reached the motorway (in this case the M11), I've passed 26 directional signs in the 10 km since leaving home – an amazing density.

On the motorway there are fewer directional signs because there are fewer junctions, but because of the speed of the traffic, each turn off to a slip road seems to be indicated four or five times. To keep things safe at speed we obviously need plenty of reminding. Suddenly, as I approach the

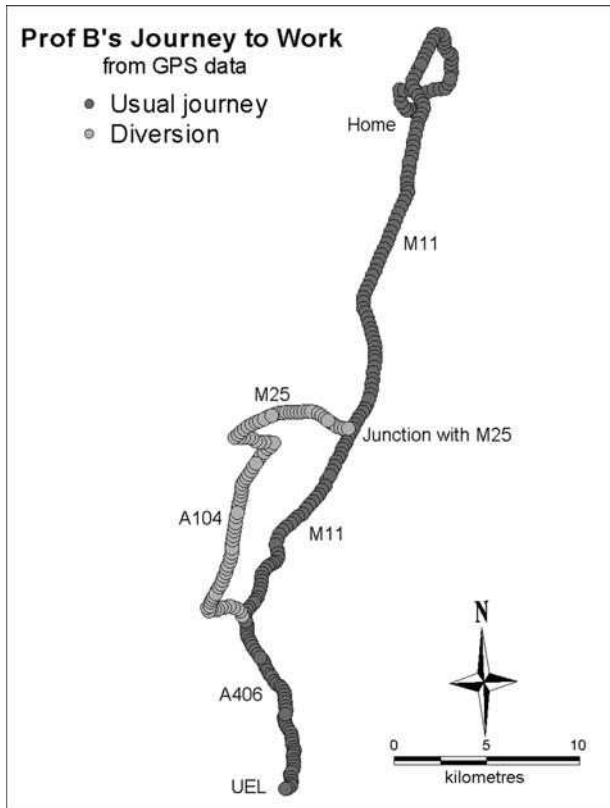


**Figure 4.3** Helping us find our way – directional road signs (photograph by the authors).

intersection with the M25 a computer-controlled dot matrix sign is warning me ‘M11 J6–J4 LONG DELAYS’ and as I look into the distance I can already see the brake lights of cars and lorries joining the queue. What to do? Sit it out? Turn off? Turn off onto the M25! Oh...Go north? Go south? Both ways will take me in the ‘wrong’ direction and the M25 is notorious for its traffic jams... a bad decision here could negate sitting it out on the M11. But I have no more information. I think I’ll go north because I remember it’s the shortest to another junction and I can get off onto smaller roads and perhaps find my way through to the university reasonably quickly. I haven’t got an in-car navigation system yet and many of the place names on the directional signs are only vaguely known to me – places I pass but never visit. So it’s largely guesswork from here informed by a schematic knowledge of the geography. Still, I believe I have a good sense of direction and the sun is in the south. Of course, lots of other drivers are doing the same thing and we start to clog up the minor roads. My eventual detour is shown in Figure 4.4 and it took only 20 minutes longer than my usual journey.

### Dr Li’s comment:

I just hate surprise diversions! I’m a careful driver and if I know I’m driving somewhere new I study the map in detail to familiarize myself with my intended route. But if I’m suddenly faced with having to make a detour into vaguely-known or unknown territory, I’d rather sit it out and be late than get lost. I’ve seen Prof. Brimicombe toughing it out and driving on anywhere rather than admitting he’s lost. But I don’t have his confidence in a ‘sense of direction’. I’m considering buying some in-car navigation – the prices have come down a lot recently!



**Figure 4.4** Map showing Prof. Brimicombe's usual drive to work and today's detour (from GPS data).

Prof. Brimicombe's journey to work (again):

It's Tuesday morning and I'm counting the directional road signs for the section of motorway I had to miss yesterday. For my usual drive to work of 54 km there are 67 directional road signs. Factor that up to the whole road network and it's a huge analogue infrastructure – more than I had realized. In addition, I pass some other interesting signage. There is one sign (easy to miss on a busy motorway) warning of forthcoming road works that will cause traffic disruption for several weeks. Then there is a warning that congestion charging for central London is in operation. There is one 'detour' sign (again, easy to miss and no doubt adding a sense of adventure to some unfortunate driver's day) and two signs put up by the AA to mark the way to an exhibition centre. There are also six radar speed traps. Although my journey is straightforward today, in other parts of London

the travel situation seems to be disintegrating – traffic and travel bulletins from a number of radio stations keep interrupting my enjoyment of a discussion I am listening to on the car radio, so much so that eventually I switch off the traffic announcements function.

Dr Li's comment:

I think you've made your point Prof. Brimicombe about there having been analogue antecedents to digital LBS – not just wayfinding of course. So much of what happens in society today is not just reliant on flows of information, but also on flows of goods and flows of people in unprecedented quantities. People want to know and do things based on where they are at that moment. And of course, Prof Brimicombe, you've highlighted a key problem – we don't want to be fed information that is either not relevant to us, or unnecessarily interrupts our activities. Most of all, I don't want location-based spam!

### 4.3 From Locating Services to Location-Based Services

---

Notwithstanding possible effects of any 'digital divides', most people of working age (and many youngsters!) in the West will be well familiar with using a PC to access the Internet and the World Wide Web. Let's not forget, though, that equal opportunity of access to a PC is by no means universal and in many developing countries the use of mobile phones is leap-frogging the penetration of the PC. Both, however, are means of locating services. In this section the distinction is going to be made between LBS and more general approaches to locating services which, whilst they may have the appearance of LBS – and may even be touted as LBS – do not fully conform to the definition of LBS given above. Because of the relatively recent appearance of aspirations towards LBS, true LBS can be hard to find. There are many at the conceptual stage, some initial offerings within the constraints of mobile technologies and, above all, plenty of research and development so that a host of offerings are just around the corner.

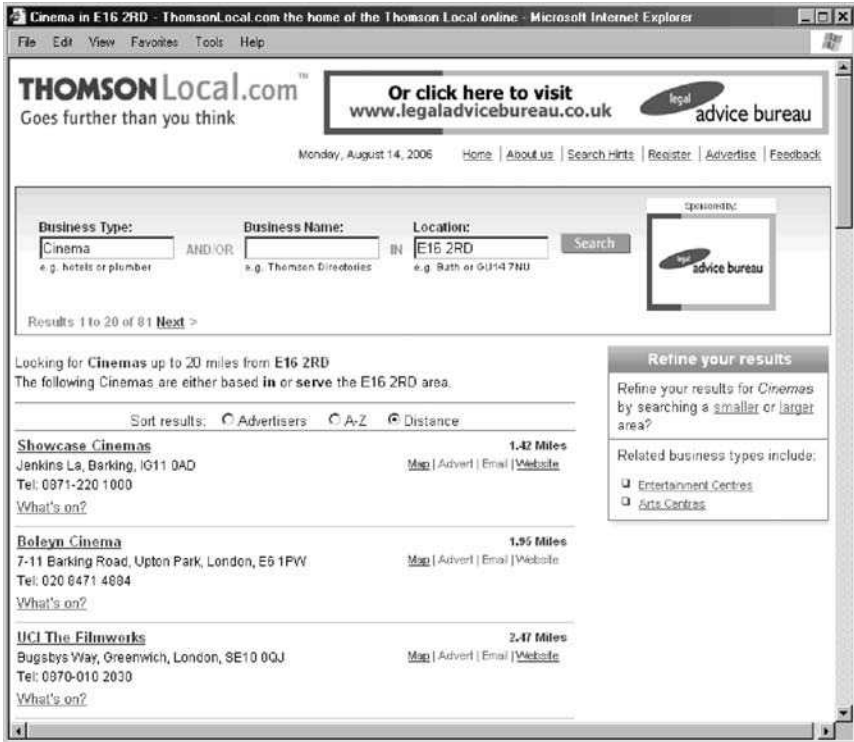
#### 4.3.1 Using the Web for Services

Most, if not all readers will be familiar with using an on-line search engine to find information and services. Popular examples are Google

(<http://www.google.com>), AltaVista (<http://www.altavista.com>) and Yahoo (<http://www.yahoo.com>). These are keyword-driven search engines that can be associated with logical operators (such as and, or, not). They will find you anything on the Web (and only on the Web) that matches the keyword search. Responses can be colossal (see Figure 1.3), useful and sometimes not so useful. The response to an identical query should be the same regardless of the location of the user, subject to in-country priority listings (i.e. those organizations that have paid for their pages to be referenced towards the top of the list and which may vary by country of access). Only if a location identifier is included in a query (e.g. London), or a query is tagged to be country restricted (such as a United Kingdom only search), should the response be location specific. But even then, the location identifier used in the query may have nothing to do with the user's current or projected location. So, useful as they are, the big name on-line search engines as currently implemented are not LBS.

Another class of Web-based search engine is location specific. These tend to be gazetteer and/or postcode driven, but can be queried based on a postal address. In Chapter 3 two Web sites for maps and route finding were looked at: MapQuest and ViaMichelin. Other examples of this class of search engine are on-line trade directories, such as the electronic Yellow Pages (<http://www.yell.com>), and on-line information about neighbourhoods, such as UpMyStreet (<http://www.upmystreet.co.uk>). Two broad approaches to answering queries can be adopted. One approach is primarily responding to a Geographical Base File (GBF), in which each gazetteer and postcode entry is listed against a seed co-ordinate in eastings and northings; the data to be presented are then clipped to the relevant geographical area based on a pre-set distance from the seed co-ordinate of interest. Thus in Figure 3.1 and Figure 3.2 the maps were produced in response to the postcode E16 2RD being entered and which is seeded at the postcode centroid as E543391 N180748 in the GBF. A number of on-line trade directories will calculate and sort their response by Euclidean (as the crow flies) distance from the user's specified postcode to the postcode of relevant services up to a maximum distance, again calculated on the fly through the GBF (e.g. <http://www.thompsonlocal.com>). Thus in Figure 4.5 a request for cinemas based on the postcode E16 2RD orders the response by distance.

The second approach is gazetteer-driven search engines, which operate through standard database *join* (Section 8.7.1) to identify pertinent information. Thus although the entry point into UpMyStreet



**Figure 4.5** Response from an on-line business directory sorted by distance from a seed postcode (from <http://www.thompsonlocal.com>; accessed 16 August 2006).

is a full postcode, much of the response relates to broader postal and administrative areas achieved through look-up tables to identify relevant information at neighbourhood, district and constituency levels. This type of Web service can be used in a number of ways apart from just finding out information about one's own neighbourhood. For example, the information presented would be useful to anyone moving to a new area of the country (say, to a new job) who wanted to evaluate which neighbourhoods would best suite their needs in seeking to purchase or rent accommodation. Do these qualify as LBS? Here we admit to being in a grey area – it depends. These search engines present services that are location specific, they even deliver maps. But 'location-specific' and 'location-based' have a nuanced difference in meaning. Finding a map of a location by gazetteer or postcode is just a digitally quick way of turning to the right page in a street atlas. Similarly, for finding pertinent entries in a trade directory.

As will be further discussed below, geographical content specific to a location is not the deciding prerequisite for LBS. ‘Location-based’ refers to the *user*, not necessarily to the content, though the content will have been tailored in a location- and/or context-specific way. Thus if Dr Li is at home and has a plumbing emergency, a query of an on-line trade directory for details of plumbers close by for a timely response would qualify as LBS within the definition – though if she did it all from her mobile phone whilst placing buckets to catch the offending leak and without even having to key in her postcode because the wireless network would automatically work it out, then Prof. Brimicombe would give his full approval to the example although LBS should not necessarily be so dramatic! If on the other hand Dr Li is wondering where to take her next holiday in Spain and is casually browsing what there is to see and do in Barcelona, Seville and Madrid then this would hardly qualify as LBS. Even if performed from her mobile phone, it would only qualify as ‘mobile Internet’ in Figure 4.1 above.

There are services on the Web that are considerably tailored to the individual user. A good example is the on-line bookshop Amazon. Prof. Brimicombe loves browsing bookshops – he can hardly walk past one without succumbing to the irresistible urge of wandering in – but he also buys on-line. When he goes to Amazon, a message comes up on the first page saying ‘Hello Prof. Brimicombe, we have some recommendations for you’. There then follows a suggestion of some books, mostly to his taste, that he might like to consider purchasing. He also receives e-mails from time to time of book offers that are tailored to his taste. How can this response from Amazon be tailored so that the Web pages delivered to Prof. Brimicombe are pretty much unique to him? It’s a combination of cookies and profiling from analysis of page requests and purchases made. A *cookie* is a text file that is either created and written to a user’s hard drive or read from a user’s hard drive by Java code that is embedded within the script of a Web page. So when the Web page is accessed it will check the hard drive to see if there is a pre-existing cookie and if so read the information it contains and if necessary update it; if not, it will create a new cookie containing the relevant information. And what is this information? There is the IP address of the machine being used, the log-in name of the individual, and time and date of last access. These allow the individual user to be recognized. The cookie can also be used to store other information, such as a log of the pages visited at that Web site. These logs can either be stored in the cookie client side or can be stored server side. Either



way, they can be analysed by the business concerned for the individual's preference for classes of products viewed and the actual purchases made and thus build up a profile of the individual customer. For an in-depth knowledge of modelling Web use in this way, the reader is referred to Baldi *et al.* (2003).

Being able to identify the individual repeat user and match it against a pre-existing profile allows businesses such as Amazon to tailor its Web pages on-the-fly. But are these LBS? Certainly in this example we are seeing a sophisticated and automated level of tailoring not seen in the previous examples. But even if one were mobile with, say, a wireless enabled laptop or PDA, the 'location' is the IP address of the machine and not its geographical location. In other words Amazon would not know if Prof. Brimicombe was using a mobile device from his office and thus tailor their response to his usual academic interests, or whether he was using the same mobile device from a holiday resort and might therefore be tempted by the latest best-selling novels or even some local history. Of course there is the potential with location-aware technologies (Chapters 6 and 7) to augment such services into true LBS.

### 4.3.2 Using Navigation Systems for Services

Let us now consider in-car navigation systems. Early in-car navigation systems were not much different to the on-line route finding or equivalent functionality of MapPoint as discussed and illustrated in Chapter 3. These are on the margins of LBS. In the last couple of years there have been increasing numbers of suppliers of 'SatNav' type in-car navigation systems, so much so that not only have the devices become smaller and simpler but prices have fallen dramatically. Whilst they are by no means inexpensive for what they are, prices have already fallen over the last few years in line with digital cameras, mobile phones and DVD players. In other words, they are on the way to becoming ubiquitously affordable.

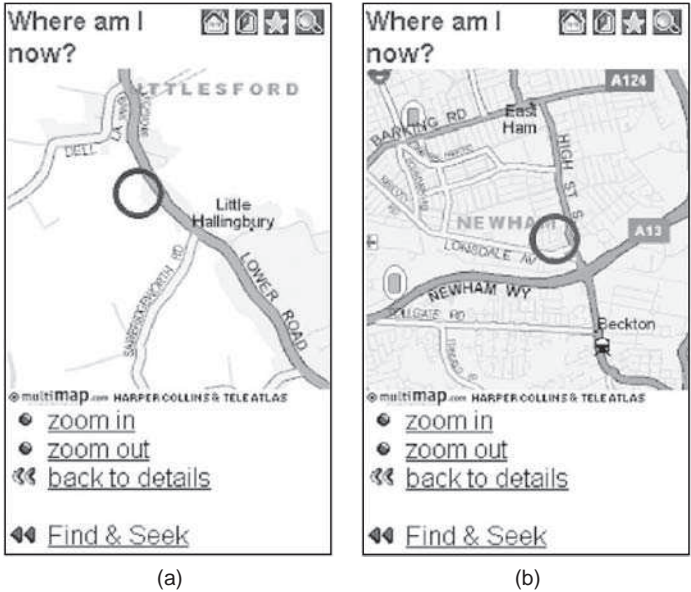
Basically, the first generation in-car navigation systems have become augmented with Global Positioning System (GPS) so that the user (whether in a car, on a bicycle or on foot for that matter) can be located within the map on the screen. For any given desired route, the in-car navigation system not only shows users where they are on the map display, they can indicate with arrows whether to go straight, turn left or right at forthcoming junctions,

and can even talk the user through so that, if driving, the user can concentrate on the road, not the map display. Because the information being given is continually being tailored to a user's geographical location then this would qualify as LBS. Such systems assume that users are driving and should for example follow one-way-systems and for that matter not wander aimlessly off the road network. They are thus designed within a restricted user context. Nevertheless, the basic in-car navigation system has no location-based dynamic capability beyond indicating the user's position and next step(s) along a predefined track. If there is a traffic jam ahead, such systems have no capability of giving advance warning and being pro-active in suggesting an alternative route. Systems with location-based dynamic capability are on the market (usually on the basis of an optional subscriber service) and thus take in-car navigation into the true world of LBS.

### 4.3.3 Using Mobile Phones for Services

Most mobile phone operators run a selection of services aimed at 3G or later handsets. These include mobile TV, news and weather, games, music and sports results and are clearly not LBS (that having been said, there are now emerging mobile games that are tailored to the locations of their users). One common class of service available through mobile phones centres on finding where one is (location) and seeking out the nearest facilities such as cash machines, restaurants and taxis. Many of these services offer to pinpoint your position from the mobile device itself.

Prof. Brimicombe uses Vodafone, which has a service called Find & Seek. He decided to give it a go, once where he lives in the country and once at the Docklands Campus of the University of East London. The results of finding his location via the wireless network are shown in Figure 4.6. The service first connects to a WAP page offering a number of options such as: 'locate me now', 'nearest cashpoint' and 'nearest taxi'. The nearest outlets and services are determined from an on-line trade directory and the user's current position determined automatically by means of the wireless network (Chapter 6). In Figure 4.6a, the service has located Prof. Brimicombe in the village of Little Hallingbury in the district of Uttlesford which is stated by the service as 'You are near Sawbridgeworth', which is indeed the nearest town (3.5 km away).



**Figure 4.6** Results of finding Prof. Brimicombe from the Vodafone Find & Seek service: (a) at home; (b) at work (map copyright: Multimap.com, Harper Collins and TeleAtlas; accessed 9 September 2006).

The circled location on the map appears to be the intersection point of the three nearest telecommunication masts (as best we can measure) but is still nearly a kilometre away from Prof. Brimicombe's house. Figure 4.6b gives the result of locating Prof. Brimicombe whilst at work. In this case, despite the higher density of masts expected in an urban area, the map indicates a location approximately 1.7 km from the campus. Clearly there is room for improvement here.

The accuracy of position fixing is further discussed in Chapter 6. If a user then wishes to determine, say the nearest cinema, the service requests the user to specify if this should be determined against their current location, another place name, a postcode, a London tube station or a district in London. Directions either for walking or driving can also be given. Of course, a specific charge is made for this type of service. This is typical of services offered through mobile phones at the time writing and indeed meets our definition of LBS in that the service can be tailored by the provider to the user's current location, or at least an approximation of it.

### 4.3.4 Location-Based Services

From the above examples, the reader should now be able to identify, in relation to the definition of LBS, those services that are definitely not LBS, those that definitely are and those that are in the grey margins and perhaps just about qualify. Our litmus test for LBS would be the response of a service to a query such as ‘what are the opening hours for Marks & Spencer?’<sup>1</sup> An LBS response would provide, as a minimum (as voice and/or text), today’s opening and closing times for the *nearest branch* of Marks & Spencer (with its address). Additional points would be scored by offering the options of:

- a map that shows both the location of the user and the said branch;
- the branch’s phone number;
- URL for mobile Internet connection;
- directions of how to reach the said branch by various modes of transport;
- any pertinent information such as special offers and sales;
- the offer to display opening times for other days of the week;
- the offer to locate and provide the above information on specific branch(es) of Marks & Spencer.

If, say, the request had been made at 5 p.m. and according to the database the nearest branch was just closing, the service could offer the option of providing information of the nearest branch with late opening hours. If, say, today was Sunday and all branches were closed, the service should have sufficient intelligence to either engage in an automated dialogue to clarify the information being sought or to offer the opening times for all days of the week for the *nearest branch*. In other words, even though the query may not appear at first sight to be geographical, the response is tailored to the location of the user, which, in most circumstances, *shouldn’t have to be asked* because it should be capable of being automatically determined as part of the service. Another example request might simply be ‘pizza’. An LBS response would include a selection of the nearest (perhaps classified as

---

<sup>1</sup>Marks & Spencer is a chain of high street stores selling clothes, shoes and food (<http://www.marksandspencer.com>).

‘take away’, ‘eat in’ and ‘take away & eat in’), the possibility of maps, directions, opening times, contact details, perhaps even sample menus and prices. If the time of the request was in the middle of the day, then lunch offers could be indicated, if it is late afternoon or evening then any special dinner menus and if really late at night, then only those still open. After making your choice, really good LBS would surely allow you to SMS, MMS or e-mail your friends the relevant information with any maps or directions automatically tailored to the current location of each friend.

Within the two examples just described, there can be discerned elements of LBS that are *pull* and elements that are *push*. ESRI (2000 p. 4), for example, provide both pull and push definitions for LBS. Thus, on the one hand, LBS can use the geographical location of a user, as determined through a location-aware device, to enable that user to pull down information that is pertinent both spatially and temporally. On the other hand, LBS can use the position of a location-aware device to determine that a user qualifies for, or would be willing to accept, relevant information pushed towards them either as part of the response to a query or independently of a query. The GSM Association (2003 p. 12) distinguishes a third element, that of *tracking*, where the locations of mobile devices are continually monitored for applications such as fleet management or mobile gaming. These distinctions are quite important ones as, for example, pull applications are always user initiated whereas push applications are likely to be controversial as they will raise issues of privacy, spam and in certain circumstances potentially dangerous distractions (such as whilst driving).

A common misconception around the definition of LBS is that somehow they must include maps. As can be seen from the above examples, maps can certainly be an efficient way of communicating spatial information, but not every response necessarily requires a map. Thus, if in the pizza query the user was only interested in knowing the telephone number of those take-away outlets for which the user was within their delivery catchments, then a map would be superfluous. As we have stressed before, the term ‘location’ in LBS refers more to the geographical position of the user as the key to tailoring information rather than the content of any response to a query. Of course, to tailor information and services by location, spatial technologies such as GIS and the use of spatial data must underpin LBS. But any idea that LBS equals maps delivered to a mobile phone is far too simplistic. Also, it should be pointed out that our definition includes

reference to a mobile user. Whilst the concept of LBS is focused on individuals on the move and can present a clear advantage for individuals when exploiting wireless networks, as long as the device being used can be location-aware, that is the user's location can be coupled with the device location, then LBS can work. Thus even where, for example, an individual's passive Radio Frequency Identification (RFID) or smart card containing contextual information (Chapter 6) were to trigger a tailored output to an information screen, say in an airport, it would be the known (fixed) location of the transceiver that would provide the location awareness rather than the RFID itself. Nevertheless, the main focus of LBS to date has been the delivery of information and services to a user's wireless-enabled mobile device (NICTs).

Finally in this section we need to briefly consider the word *context* in our definition of LBS. This refers to the current situation and activities of a user that may be pertinent to tailoring an information response in addition to knowing their location. As in the examples given above, an implicit context would derive from time of day, day of the week or month of the year (seasonality). Other contexts, as will be seen in Chapter 7, can be determined from the dynamics of the user, the environment and the technology involved. Thus, if the location of a user is being tracked by a service and that user is found to be travelling above a threshold average speed along the road network, then the service would conclude that alerts on adverse road conditions ahead would be welcome as might information of suitable stopping places every two to three hours of journey. Context may also be made explicit through setting profiles. This can be seen already in mobile phones where profiles can be set depending on whether one is in a meeting, at home or driving, so as to automatically change the settings of the mobile phone such as routing calls automatically to voicemail or having a louder ring tone. We see pre-set profiles of a similar nature being central to the way LBS will work as they further develop to ensure that content of services are contextually relevant and welcome. This aspect will be returned to in later chapters.

## **4.4 E911 and E112 Mandates**

---

For one London Borough (population: a quarter of a million) for an eighteen month period starting January 2004, the Command and

## **Location-Based Services and Geo-Information Engineering**

Despatch Centre that handles all emergency calls logged an average of 630 calls a day. Quite a number of these calls were from police, ambulance and fire service personnel themselves reporting information, requesting assistance and so on. Calls from members of the public for emergency assistance averaged 118 a day. Our analysis of the data showed that 42% of callers were from land lines, 58% were from mobile phones. Factor this up for the whole of London and an average of almost two thousand mobile callers a day are asking for emergency assistance. Factor this up further to the whole of the United Kingdom and it is tens of thousands of mobile phone calls every day. If they are mobile...where are they located in order to render them prompt assistance? Of course, a caller may have a very clear idea of where they are, but many do not. How many roads have you driven, cycled or walked down that you don't know the name of? If you were unexpectedly caught up in an incident either as victim or witness, would you be able to quickly provide a fix on precisely where you were? And then again, you might be somewhere that doesn't have an identifiable address, such as in a car park, along a country road or halfway up a mountain...what then? This is not the same problem for users of a land line where the telephone number can be quickly linked through a database to the address where the end of the line and hence the telephone is sited. The problem only relates to mobile phone users: needing to know where they are in an automated way in order to respond to emergency calls. In some countries, the solution to this problem has been one of the initial spurs for the growth of LBS.

### **4.4.1 E911**

In the United States, an important reason for owning a mobile phone is so that emergency calls to 911 can easily be made as and when required. It has further been estimated that 30% of all emergency calls in the United States are from mobile phones (FCC, 2005). In 1996 the Federal Communications Commission (FCC) mandated that wireless carriers (cellular network, mobile radio and broadband personal communication service providers) must be able to reliably identify the location of 911 calls from mobile devices. This is commonly known as the E911 Mandate where E is for Enhanced. This enhancement of 911 calls is based on adherence to a basic rule and two phases of roll out:



- **Basic E911 rule:** All wireless carriers must transmit to a public services answering point (PSAP) all 911 calls regardless of whether the caller is a subscriber or not.
- **Phase 1 E911:** Wireless carriers must be able to relay to a PSAP a mobile caller's telephone number (enabling call-back) and location of the base station or cell receiving the call.
- **Phase 2 E911:** Wireless carriers must be able to relay to a PSAP the precise location of a mobile caller in longitude and latitude and that these must meet accuracy standards depending on the type of technology used.

To meet these requirements, wireless carriers had to ensure that 95% of the handsets in use were E911 capable (i.e. were location-aware) by the end of 2005. For those carriers opting for a GPS solution, that is the embedding of a GPS chip within the handset, this required the upgrading of 95% of handsets in use to GPS-enabled models. For those carriers opting for network-based solutions (Chapter 6), they needed to upgrade their cellular infrastructure in order to meet accuracy requirements. These were set by the FCC at a radius of 125 m for 67% of callers (i.e. assuming a normal distribution,  $\pm$  one standard deviation = 125 m; by implication 95% of callers should be positioned within a 250 m radius). The roll out of E911 has been affected on the one hand by carrier's ability to meet deadlines for handset and infrastructure upgrades and a number have sought waivers to the original deadlines in order to have more time. On the other hand, PSAPs have also had to upgrade their equipment in order to receive and work with the new data. Although federal governments have introduced levies in the form of an E911 tax on all mobile callers, the funds have not always been used for the purposes of E911 upgrades but to plug other budget gaps (Williams, 2005). Despite missed deadlines and slower than expected upgrades of the system, E911 has had an important enabling effect for LBS in the United States in raising awareness of the positive benefits of locating mobile devices and mandated operationalization of the necessary infrastructure upon which LBS could be mounted.

### 4.4.2 E112

In 2003, some seven years after the FCC in the United States, the Commission of the European Communities (the European Union)

made recommendations for a set of location-enhanced 112 (E112) regulations. Whereas in the United States there was only the 911 emergency number for assistance, across European Union member states there was almost no commonality for emergency numbers in use. In the United Kingdom there is 999, in Sweden 90 000, Finland 10 022, Luxemburg 113, Spain 091 and France 17. It is clearly obvious that individuals travelling from one member state to another with a roaming mobile phone may not have known what number to call. E112 mandated for a common number to be available alongside the traditional numbers and that provision is made to respond in more than one language. Furthermore, the issuing of a 112 call had to be picked up by any available carrier regardless of whether or not the user was a subscriber. The carrier would also have to furnish the location of the user regardless of any prior permission by the user to do so or not.

The main difference between E911 and E112 is the flexibility taken towards the accuracy of the location coordinates that need to be provided. Whilst in the United States this is clearly specified, in Europe no accuracy obligations were put in place. This was not an omission but recognition of the possible high costs involved (on top of the very expensive 3G license fees that had been paid by the operators) and that through experience the accuracy needs would become better defined. Thus a mandated level of accuracy for location is not precluded in the future. In the meantime, as seen in Section 4.3.3, accuracy for mobile phones may be quite low and the precision variable. In the absence of any mandate, movement towards better accuracy is likely to be quite slow as it is an avoidable expense with no revenue. E112 is thus much less likely to be a catalyst for LBS than E911 is.

## 4.5 Keitai

---

In the Far East, particularly Japan, the spur towards LBS has been quite different. Whilst the United States is generally considered to be at the forefront of most IT-based innovations, it is in fact the Far East that is the leader in Web, e-mail and other Internet interactive services accessed by mobile phone. The dominant model for engagement in the United States and Europe has been the PC plus fixed Internet link. Home PCs never had such high penetration in Japan, China and Korea for reasons such as small living spaces (as in Japan), low penetration of domestic phone lines (as in China), economic recession in the late

1980s and early 1990s (Japan) or time lag in economic development (China). The Far East was faced with maturing Internet services at the same time as mobile telecommunications were being launched, thereby enabling these countries to leapfrog PC-based Internet access as the dominant mode. In Japan the word *keitai* is not just for a mobile phone but for a portable device that is dissolving the digital divide and is credited with helping Japan lift itself out of recession. Thus *keitai* has a status more akin to a cultural institution. The technology used in Japan is i-mode, which unlike other mobile phone access is always online, using General Packet Radio Service (GPRS) to maintain the always-on connection. The phones themselves have a specifically designed Web browser to display the content. The communication speed is relatively slow at 9600 bps but page size is typically small (about a kilobyte using compact HTML) and e-mails are limited to 500 bytes, so the overall effect is one of quick response.

Initial take-up of *keitai* was as a youth craze, particularly among young women due to very low subscription fees and, with downloads of icons, pictures and sounds, the ability to make messages cute (Matsuda, 2005a). It now has huge market penetration with three providers – DoCoMo, KDDI and Vodafone – dominating the market. Whilst the dominant form of communication is e-mail, a survey of usage in 2002 reported in Matsuda (2005b) showed that nearly 70% had used *keitai* for ring tone download, 23% for weather forecast, 16% for transport information and just under 7% for maps. Indeed, the authors witnessed a young lady in Shinjuku station, Tokyo, consulting her *keitai* as to which platform she needed to be at rather than look at the overhead display giving departure information (see also Figure 1.7).

With i-mode, LBS hold considerable potential though take up of transport information as a key LBS application appears slow from the figures given above. GPS and assisted-GPS *keitai* are now available on the market as a means of improving positioning over the current trend of using cell ID. However, as reported by Kohiyama (2005), there has been a pervading concern over issues of privacy and surveillance in relation to LBS applications where other people are able to determine an individual's location; this has led to a negative reaction. Buddy systems and communication preferences based on proximity (say, with nearby friends) generally get around privacy concerns by providing users with information based on relative position (how far they are away) rather than based on absolute coordinates. One application that has had good uptake is in-car navigation

where location relates to vehicles rather than individuals. Also popular is using *keitai* to find local services such as stores and restaurants within a neighbourhood based on a users current position.

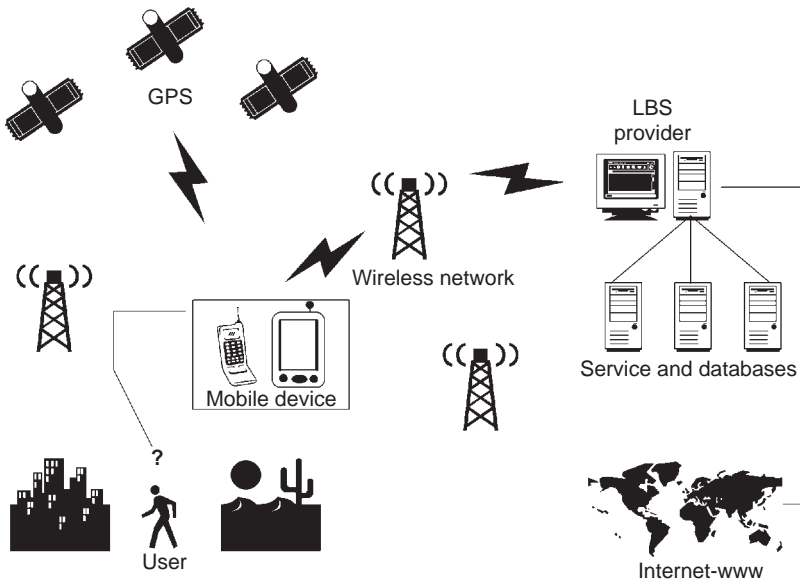
### 4.6 LBS Architecture

---

Having discussed in some detail the nature of LBS and some of the underlying drivers, it is appropriate at this stage to look at the architecture of LBS. Since LBS are a heterogeneous technology, they comprise of a number of sub-architectures to do with wireless communications, the Internet, GPS and so on. Aspects of these are discussed in other chapters as appropriate. In this section, the main components of the architecture are considered in two stages: firstly a broad overview of the main technological components followed by a more detailed view of how services can be brokered through the wireless communication network.

In Figure 4.7 there are five broad components in the system architecture that need to be brought together for effective LBS:

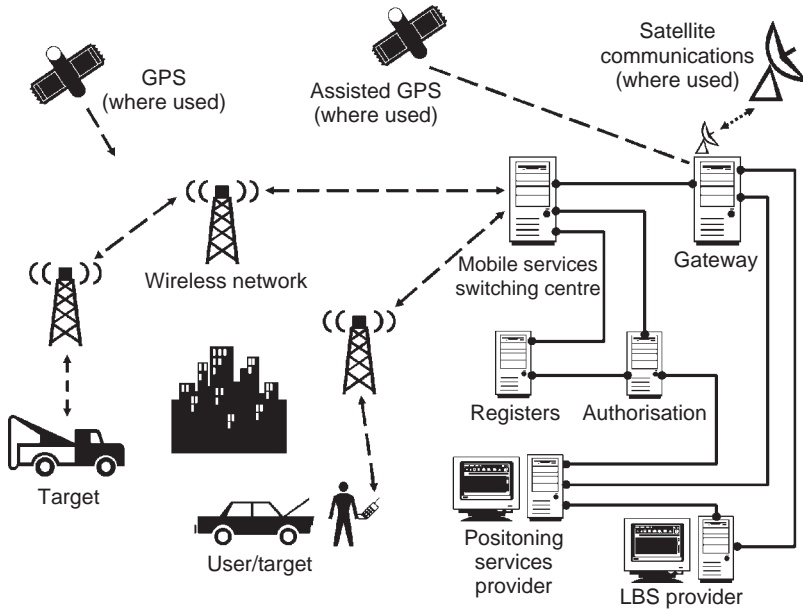
- Information about the real world in which the user is situated, accessible in databases and/or over the Internet and which can be tailored to the user's current or projected location.
- The choice and nature of mobile device currently being employed by the user to access LBS. Such a device would have to be either actively or passively location ware; where actively location-aware means that the NICT is able to determine and communicate its own position (e.g. by using internal GPS) and where passively location-aware means that the position of the NICT is able to be determined by other elements of the system to which it is linked (e.g. by the wireless network).
- The positioning systems that allow the geographical location of the user's mobile device to be known.
- The wireless network (whether this be Bluetooth, WiFi, mobile telecommunication network or indeed a combination of these) through which the user's query and LBS provider's response can be communicated between mobile device and server.
- The LBS provider, including the software (e.g. GIS) and other distributed services and components that are used to resolve the query and provide the tailored response to the user.



**Figure 4.7** Broad overview of LBS architecture.

Each element of the broad architecture just described has its own detailed architecture even right down to, say, the internal working of a mobile phone, but this level of technical detail is largely beyond the scope of this book, other than what has already described in Chapter 2. Nevertheless, it is necessary to expand a bit more on some of the intricacies that lie behind the delivery of LBS, say, across a mobile phone (cellular) network, as this has a number of implications for the topics addressed in subsequent chapters. Figure 4.8 therefore selectively expands the broad component architecture of Figure 4.7. It is based around a scenario of a driver requiring a breakdown service. The driver is the *user*, that is the entity requesting the service. In this scenario there is a single user requesting only one service (more normal scenarios would involve many users having either pull or push requests for services). The *targets* are the entities requiring positioning and thus drawn into the fulfilment of the service being requested. Thus the user is also a target, the other target(s) in this scenario being the truck(s) of the vehicle rescue service.

The request by the user to an LBS provider, from say a mobile phone, might take the form of a voice exchange (either to a real human or with a voice-synthesized service), texting using SMS or interactively



**Figure 4.8** Selectively more detailed architecture for LBS.

through a (WAP) Web site using GPRS. Either way, the user's request for help will be routed through a mobile phone provider's telecommunication network to a *mobile services switching centre* (MSSC). This is in effect a telephone exchange for mobile phones. Now it may well be that the user (the owner of the broken-down car) has roamed beyond his/her usual mobile service provider's range (as will happen when driving between countries in Europe). In this case, the MSSC will check with a *register* to find out which mobile phone provider is the *home provider* for billing purposes and will then check with that home provider the *authorization* for the relevant modes of communication (voice, SMS, GPRS) and services that can be made available through the user's mobile service contract. The user's request for help can then be routed to the relevant *LBS provider* intended to receive the request via a *gateway*, or gateway mobile switching centre (GMSC), which forms the interface between the telecommunication network (or portion thereof), the fixed line network(s) and other wireless networks (Section 2.3.1).

The LBS provider now needs to fix the user's position, in other words the user becomes a target. The request for positioning is

sent by the LBS provider to a *positioning services provider* (PSP) whose function is to contract and liaise with network and technology providers so as to perform the position fixing of target(s). The PSP checks that the user has lodged an authorization with the home provider to be a positioning target and then checks the register to determine whether the user's mobile phone is currently attached to the home provider's network and, if not, who is the current *visiting provider*. The PSP then evaluates the technology alternatives available for position fixing (Chapter 6) by interrogating the user's mobile phone back through the gateway and MSSC for the presence of integral GPS (and, if so, taking off the coordinates of the current position), and if no GPS then determining the available network method for position fixing and likely accuracy. The position fix and its accuracy are then relayed to the LBS provider. The LBS provider then needs to identify vehicle rescue services operating in that area (either from its own databases and/or over the Internet – see Figure 4.7) and one or more may be contacted. There may be a choice of vehicle rescue services on offer and each is likely to have its own fleet management service allowing the location of each truck (target) to be monitored, either in much the same way as described above under contract with a PSP or by using a combination of GPS and bespoke wireless communications. Thus each vehicle rescue service contacted by the LBS provider would, in an automated way, determine the nearness and availability of its trucks, evaluate a response time and provide its scale of charges. The user might then be sent by the LBS provider the best three offers of services; the user confirms one offer and the LBS provider confirms the order with the relevant vehicle recovery service providing the user's name, location and contact details from its client database.

The type of architecture described above is by no means fixed; what is outlined are the principles involved. An exciting and imaginative variant is shown in a video describing Hewlett-Packard's CoolTown technology concept (Hewlett-Packard, 2000) in which the vehicle itself, through its own diagnostics, determines that a breakdown is imminent and at the same time as informing the driver where to go in order to find the nearest service station through the in-car navigation system, notifies the service station of its impending arrival and arranges for a cab (taxi) to pick up the driver at that service station in order to complete the journey on time – all through vehicle-interrogated LBS! Is all this just around the corner? Well yes, it's all technologically feasible.



## 4.7 Application Areas

---

The range of services that it is expected to see emerging as LBS matures in the foreseeable future are summarized here. They have been classified by broad activity areas regardless of whether pull, push or tracking:

- **Navigation**, such as in-car navigation systems, which assist the driver by identifying appropriately optimized routes for a vehicle to be driven from some starting point to some pre-specified destination (or series of destinations). Some suppliers offer subscription services for real-time traffic updates.
- **Wayfinding**, which is the process by which individuals orientate themselves and discover routes, modes of transport and other spatially located objects, landmarks and points of interest.
- **Real-time tracking** of vehicle fleets, business associates, social contacts (so-called buddy systems to know when friends are nearby) or one's family members (such as tracking children home by parents who are still at work – popular in Japan).
- **Mobile commerce** (m-commerce) which would incorporate both transactions made by individuals on the move via a NICT and receiving (pushed) alerts (e.g. advertising, notification of discounts) to opportunities for transactions that are location-specific (such as at a nearby shopping mall).
- **User-solicited information** for all kinds of business and social purposes, such as weather forecasts, traffic conditions, delays to trains and flights, film showings and ticket availability, menus and local maps.
- **Location-based tariffs**, such as differentials in road pricing, pay-as-you-go car insurance and similar schemes.
- **Fulfilment** of field-based work orders including deliveries, maintenance inspections and data collection. Also included here would be the practice of *geofencing*, whereby field-based workers are tracked to ensure they remain within specified geographical limits during their working day.
- **Coordinating** emergency and maintenance responses to accidents, interruptions of essential services and disasters.
- **Artistic expression** in the community (e.g. Lane, 2003) that might include digital graffiti, location-based story lines and discovery trails.
- **Mobile gaming** where the players and actions are location-based.

## 4.8 Implications of LBS for GIScience

---

The range of applications areas listed above is quite broad amounting to what is predicted to be a very substantial and fast growing market sector. Building robust applications of LBS that can realize the commercial potential is a challenge for geo-information engineering and thus presents a research agenda for geo-information science. That agenda is summarized in this section and draws on the ideas given in Brimicombe (2008), Raper *et al.* (2007), Brimicombe and Li (2006) and Jiang and Yao (2006). Subsequent chapters focus on specific issues and in doing so expand on the agenda.

- **Data:** Whilst data availability is key to any information service, LBS have implications for data acquisition and integration and the granularity (resolution) of those data over and above what is currently available on typical Internet GIS applications (Tsou and Buttenfield, 2002; Grejner-Brzezinska *et al.*, 2004; Smith *et al.*, 2004). Responses to requests for services, particularly where paid for, will be expected to have a commensurate level of utility linked to the level of detail and currency of the information provided. The information sought is likely to be very localized, will have to be accurate and up-to-date. This increase in detail will bulk out current databases by several orders of magnitude, with implications for data handling and timely response to queries. Also, given the dynamics of many urban areas and the rate at which change occurs, there are issues in maintaining the currency of the data. Finally, the types of data that might be harnessed in providing a response are unlikely to be restricted to maps but might include imagery, sound clips, textual data and virtual reality clips – geographically referenced multimedia – which will challenge the way geographical databases are currently constructed and queried.
- **Locating the user:** Knowing the location of the user requesting a service is fundamental to LBS – otherwise the response cannot be tailored to location. The location may be the user's current location or some (future) projected location. The latter may be expressly stated (e.g. 'in Edinburgh tomorrow') or

determined in the shorter term on the basis of the user's current location and trajectory (speed and direction). A user's location can be expressed in a number of different ways: latitude and longitude, an address, a local neighbourhood or a district. These are progressively easier to determine. There are a number of technologies available for position fixing, each with its advantages and disadvantages and level of accuracy, as will be discussed in Chapter 6. As seen in the example above (Figure 4.6), the challenge will be in consistently fixing the user's location with sufficient accuracy so as to maximize the credibility and utility of information provided by LBS. Key here is the term 'sufficient accuracy' to denote the accuracy required to give a response of adequate utility which will vary depending on the nature of the request. Thus for an update on weather conditions the user need only be fixed within the nearest few kilometres, whereas to find a broken-down vehicle the user needs to be fixed within 10 metres. Very few requests would require sub-metre accuracy in positioning the user.

- **Contextualizing the user, environment and technology:** Dey (2001, p. 5) defines *context* as 'any information that can be used to characterize the situation of an entity'. The entity in question is the user and part of the characterization is the user's geographical location. The environment might usefully be further contextualized according to whether, for example, it's an urban or rural situation. The time of day and prevailing weather conditions can also add context as to what might be an appropriate LBS response to a query. Thus being directed across a wooded park as the shortest route to the nearest cash machine may be considered a good choice during the day, more so on a dry, sunny day. It would be considered rather a poor response on a rainy day and particularly so if it was after dark. Attitudes to personal safety and what constitutes a risk may depend on the user's gender. This is situational context related to user characteristics and behaviour. Certain types of contextual information might be derived from data mining and spatio-temporal analyses of the user's activity history, particularly from tracking data. Thus commuters who daily follow the same route (either by car or by public transport) might automatically be notified of any disruptions to their usual journey that are likely to affect them that day (or even in

the near future). The issue of context is discussed further in Chapter 7.

- **The spatial query:** In general we have a low threshold to impatience when waiting for responses from IT. It is unfortunate then that geographically-based queries take longer to transact than more traditional type queries of well-structured attributes in a database (Shekhar and Chawla, 2003). Adding to the length of time necessary for a response will be the time taken to transmit the query to the LBS provider, time taken to fix the user's position and the time taken to format the result of the query and transmit it back to the user. So there are challenges in optimizing query processing times so that users perceive a prompt reply. Geographically-based queries often rely on topological relations between objects (as discussed in Chapter 3, and again in Chapter 8). These relations currently tend to be for 2-D spaces and for static objects. For LBS these will need to be expanded to include dynamic and mobile objects. Another consideration is that most users of GIS technologies have to date been trained users. If LBS are to have popular appeal, then they must be geared to the 'naïve user', that is a user that has no formal training or expertise in database or GIS. The challenge here is in allowing the user to use natural language queries and have the system 'translate' them into standard query languages such as SQL but incorporating suitable levels of fuzziness to represent the inexactness or ambiguities that can arise from natural language queries.
- **Communicating the response:** There are a number of challenges here, not least being the delivery of responses to small screens such that they remain intelligible and meaningful to the naïve user. Responses to queries will need to take account not only of context but also user preferences for different modes of communication. Some users may feel comfortable with maps, others may prefer text or voice or even 3-D images. Scale and level of detail are also important and may well depend on context. Thus finding a particular address nearby would call for detailed information, requesting an update on the weather might be satisfied with a less detailed response. Technical challenges for GIS concern on-the-fly generalization and customization of maps to take account of scale, content and appropriate symbology. Also, LBS responses do not just concern maps but may include relevant voice,

video clips and virtual reality which will need to be appropriately integrated and made accessible. Again, these issues are further discussed in Chapter 9.

- **Interoperability:** LBS are heterogeneous technologies, that is they involve many different technologies in the delivery of services. Despite the general technological convergence that is happening (Chapter 2), interoperability remains a significant issue, not least for GIS (Chapter 3). Interoperability must work at several levels: technically, in that information can physically move across the system; semantically, in that the information being gathered and passed from one service to another retains its meaning; and commercially, in that business models operating within the value chain allow the acquisition of services cost-effectively for the user.
- **Legal and social issues:** Whilst LBS have enormous potential for enhancing our working and social lives, there are also perceived dangers of abuse, particularly related to privacy. Notions of privacy are culturally constructed. Thus, in the United Kingdom where the CCTV coverage is probably the densest in the world, to be recorded by numerous cameras on any trip outside the home is not seen as an infringement of privacy but an issue of safety. Other Europeans, such as the Swiss, would find this level of capturing one's every movement intolerable (citing personal communication). Nevertheless, in delivering LBS where users and targets need to be positioned, there needs to be a system of safeguards to protect privacy where desired and to prevent the abuse of tracking information and profiles constructed around individual users.
- **Business models:** for LBS to be successful, they have to be cost-attractive to users and make a profit for providers. Business models are therefore just as important as the technical aspects of achieving LBS (Chapter 10). Business models do not feature strongly in GIScience research as traditional models of mapping authorities as providers with data users as purchasers of their products (or, in some countries, provided at cost of dissemination) has been a time-honoured approach. Given the heterogeneity of LBS, other business models are likely to compete with and even undermine the traditional model, just as has been happening with the music industry first with Napster and then with YouTube. In the LBS market there is everything to be played for.

# Chapter 5

## Data for Location-Based Services

### 5.1 Introduction

---

The information society as part of the overarching context for location-based services (LBS) was discussed in Chapter 1. When we talk about ‘information’ we are increasingly referring to ‘digital information’. Even if we read things in print, they inevitably got into print by digital means (e.g. word processing). The term ‘information’ is also being used increasingly loosely. In a recent, short article (22 column centimetres) on the world’s digital output, the terms information, data and content are used interchangeably. A *datum* is a fact, often in the form of a measurement (see data types in Section 3.4.1) or as an indication that something is perceived to exist or has occurred. *Data* are thus collections of facts – the building blocks of information, evidence and knowledge. *Information* is something about which we are informed (have communicated to us – or become known to ourselves) usually as a result of analysing or interpreting data. Information is often (should be) the basis for rational decision making. Both data and information can exist independently (as in databases or on Web pages) whereas both evidence and knowledge are linked to the thoughts and activities of people. *Evidence* is the bringing together of data and/or information from different sources in a consistent and valid way to provide new knowledge, or refine existing knowledge, about phenomena (things that happen). *Knowledge* is a blend of skills, assimilated (digested) evidence and the imparted wisdom of others that

form our individual basis for knowing things with a level of certainty (and hence predictability). Finally, the term *content*, in an informational sense, can be used as an umbrella term for data, information, expressed evidence and statements of knowledge that are stored or communicated in an accessible format; thus we can refer to Web content as being all the myriad of things that can be accessed on the Web.

Our world is full of digital content. Exactly how much content has been the subject of estimates and debate. Studies for 2006 (reported by Wray, 2007) conclude that the digital output for the year was 161 exabytes. An exabyte is a staggering billion gigabytes, so 161 exabytes would fill 161 billion iPod Shuffles. E-mail traffic accounted for 6 exabytes (3.7%). Much of the rest was accounted for by digital telephony (e.g. mobile phones), use of digital cameras (stills and video) and digital TV. London's traffic surveillance cameras generate about 8 million gigabytes daily. All of this can be compared with the entirety of our pre-digital outputs since the dawn of time (including the spoken word), which, if stored digitally, would not exceed the e-mail traffic for 2006! But how much of the digital content generated in 2006 was new and original? It seems that only about a quarter falls into this category, the other three-quarters being forwarded e-mails, downloaded copies and pirated material. 40 exabytes of original content a year, though perhaps not all of it useful as data and information, nevertheless presents a daunting task for society in how to harness this quantity of material.

This plenitude of digital content has only been a recent phenomenon. For the spatial sciences, for example, the 1990s were a period of transition from data poverty to data richness. In recent years, digital spatial data sets have grown rapidly in scope, coverage and volume (Miller and Han, 2001). This change has been facilitated by:

- improved technology and wider use of Global Positioning System (GPS), remote sensing and digital photogrammetry for primary data collection;
- the introduction of new technologies, such as light detection and ranging (LiDAR), that allow improved data capture in the third dimension (height);
- the operation of Moore's Law resulting in increased computing power to process raw data coupled with the falling cost of data storage;
- the advent of data warehousing technologies;



- the growing availability of digital secondary data sources such as census, commercial and administrative data;
- increasingly efficient ways of accessing and delivering data on-line.

These technical advances in hardware, software and data have been so profound that their effect on the range of problems studied and the methodologies used to solve them have been fundamental (Macmillan, 1998) and have indeed paved the way for LBS. The issue today is not so much the feasibility of collecting digital data *per se* (Section 5.3) but as to whether they have the right coverage (the area of interest), can be of sufficient currency (up-to-date) and have sufficient granularity (detail) to provide answers to queries that meet threshold levels of utility for customers. Data are key to the successful implementation of LBS; therefore, how can we ensure that we have the right data to provide information that is useful so that customers will sign up for services?

## 5.2 The Size and Granularity of the Problem

---

If data collection techniques have been getting more sophisticated (Section 5.3) and data have become much more plentiful, what is the problem concerning data for LBS? Let us consider an example. Suppose Dr Li wanted to drive from her home just outside London to, say, York in the north of England – a journey of some 320 km. Finding her way from home to the nearest motorway to take her north is no problem since this is the area in which she lives and she has accessed the motorway network many times before. As she goes north she follows a succession of cities (easily remembered: Cambridge, Peterborough, Doncaster) that are progressively signposted along the motorway network until York is indicated. In other words, to reach a broadly defined destination only macro information is required; and it's the same irrespective of whether the journey is by car, coach, train or plane. Dr Li's problem (and nearly everybody else's) is finding the precise address at the journey's end in an unfamiliar area. This is commonly referred to as the problem of 'the last mile', which on city scales of complexity could perhaps be more usefully stated as 'the last ten kilometres' or 'LTK information'. To satisfy this, one needs micro level data that is highly current. Dr Li wants to be routed from the nearest arterial road to the exact address without being directed into a major traffic snarl up due to, say, road works. Because a large customer base may

want LTK information for anywhere in the country on any day, the data problem facing an LBS provider is one of maintaining granularity, both spatially and temporally.

The transition from data poverty to data richness has been most noticeable for macro- and meso-scale data sets with an annual update cycle. Achieving this for micro-scale data with daily or weekly update cycles (as relevant) and particularly the attributes of features is more difficult and requires data collection infrastructures that are expensive to organize and implement. The data requirements for LBS exceed those that are currently available for typical Internet Geographical Information Systems (GIS) applications, corporate GIS and public sector GIS (Tsou and Battenfield, 2002; Grejner-Brzezinska *et al.*, 2004; Smith *et al.*, 2004).

Overall, demand for LBS will inevitably depend on the utility that can be gained from accessing the service. The level of utility will largely depend on the currency, granularity and fitness-for-purpose of the information provided, alongside response times and comprehensibility of information formats (text, voice, maps and so on). Fitness-for-purpose is further discussed in Section 5.5, whereas issues of response times will be discussed in Chapter 8. Data for LBS will need to be fine-grained and up-to-date. For example, one commercial database on petrol station locations has 106 attributes per site, including opening times, shop type and availability of car wash as well as up to 28 photographic images of the facilities (<http://www.catalist.com>). It is likely then that the sizes of data sets necessary to support LBS at sustainable levels of utility are set to grow dramatically as the spatial granularity is refined and as additional attributes are added. A further major consideration in urban areas, given their dynamics and complexity, will be the problem of maintaining currency of data. The size of some typical United Kingdom data sets on the market, expressed as numbers of records, is given in Table 5.1.

It is clear, however, from the example just cited of the petrol station locations, that the size of the database also depends on the number of attributes being stored against each entry – the more attributes there are (to add utility to the data) the more effort is involved in keeping the data current. To gauge the effect on the sizes of data sets of moving from macro- and meso-scales to micro-scales, Table 5.2, in which we give the number of attribute fields, the number of records and file sizes in MID/MIF format, has been constructed for Greater London. The MID/MIF format is a data exchange format from MapInfo that contains data on projection type, coordinate

**Table 5.1** Indicative size of some data sets useful for LBS (UK coverage, 2003).

Data type	Indicative size (number of records)
Addresses	24 million
Postcodes	1.7 million
Street gazetteer	764 000 entries (Navteq)
Financial and legal services	190 000 banks, building societies, estate agents, etc.
Hotel, restaurants, pubs	162 000 locations
Local services	230 000 local service and retail companies
Recreation and culture	87 000 locations

**Table 5.2** Comparison of some GIS data sets for Greater London, 2005/2006.

Greater London				
Dataset	Features	Fields	Records	MID/MIF format
Navteq Standard Streets	Highways	20	5 687	1 MB
	Main Roads	20	34 368	6 MB
	Minor Roads	20	148 217	25 MB
	All	20	188 272	32 MB
OS Integrated Transport Network	All	19	230 653	71 MB
OS CodePoint	All	15	281 014	45 MB
OS AddressPoint	All	25	3 642 200	930 MB
OS Landline	Buildings	4	5 614 401	839 MB
PointX	All	19	356 844	89 MB

OS = Ordnance Survey

system and feature coordinates in a MIF file and the attributes of each feature in a MID file. This exchange format is readily imported into most GIS and is a useful way in the current context of comparing file sizes for different data sets. Table 5.2 illustrates a number of issues relevant to data for LBS:

- Navteq Standard Streets is a road centreline database used in a number of Web-based GIS applications, in-car navigation systems and for fleet management (<http://www.navteq.co.uk>). Illustrated here is how the number of records and file sizes substantially increase with a densification of the road network from highways down to minor roads;
- The Integrated Transport Network (ITN) is a component of Ordnance Survey's MasterMap data. It, too, is a road centreline database but has a higher level of spatial granularity

compared with the Navteq data. Not only are more minor features included (such as alleyways, which are perhaps not normally accessible by vehicle) adding to the number of records, but more importantly in terms of file size is the level of detail in the digitization. The Ordnance Survey uses more digitized points to represent curved features (such as roundabouts) to give them a smoother appearance – but this results in much large file sizes. Thus the 22% increase in the number of records coupled with the finer detail of digitization results in a 122% increase in file size compared with the Navteq data set.

- Turning now to addressable features, CodePoint is a data set of postcodes of which there are over a quarter of a million in London. Each postcode is represented as a point at the address-weighted centroid (i.e. is the average location of all the addresses within the postcode). To increase the granularity to these individual addresses there is the AddressPoint product, which provides a geographical point for each addressable property (though a block of flats might be represented by a single point). File size has increased 20-fold. If a point on a property is insufficient and we wish to see building outline then the Landline product has a smaller file size but far fewer attributes. The number of records though has increased by nearly two million or 54%! This illustrates an important problem in LBS: not all buildings for which people might want directions have a separate address or indeed an address at all. An example of the latter might be a multi-storey car park, which is never assigned an address because letters are never delivered to it.
- Finally in our table of examples is the PointX product. This is a data set of points-of-interest (<http://www.pointx.co.uk>) from accountants, through cash points and feng shui consultants, to zoos. This is a bit like a trade directory with a geographical point for each outlet. File size is larger than the road network.

The data sets listed in Table 5.2 represent the absolute minimum that would be required to offer customers viable LBS. The total range of data types required to support full LBS implementation is unclear at this stage of development, but is likely to include:

- **Base mapping:** vector and raster data sets over a range of scales including road networks, railways, building outlines, elevation data and administrative boundaries.

- **Points-of-interest:** various categories of prominent places and landmarks (with attributes) that are both addressable and nonaddressable.
- **Services-of-interest:** the equivalent of fully geocoded electronic yellow pages cum business directories with detailed attributes of each service, including URL link to Web pages.
- **Events-of-interest:** from weather forecasts, traffic news on accidents and/or levels of congestion, roadworks and diversions, updates on public transport services, events such as pop concerts, marathons and flower shows.
- **Navigation data:** public transport routes and intersections, street-level routing data including one-way systems, restrictions, key signage and other road furniture such as traffic lights and bus stops.
- **Imagery:** satellite, aerial and terrestrial imagery for visualization.
- **Sound clips:** such as commentaries on features of interest and voice-synthesized or pre-recorded instructions.
- **Moving image clips:** from Web-cams and CCTV; also film trailers, advertisements and the like. Typically, file sizes for video clips are about 1 MB a minute.
- **Virtual reality visualizations:** for full 3-D 'wander through' visualizations and explorations of locations of interest.

Within the above list, it is useful to distinguish, from a technical perspective, three broad sets of database objects: *static* (short to medium term invariant, such as road and rail networks), *dynamic* (short term periodically updated, such as weather forecasts) and *mobile* (spatially on the move either intermittently or continuously and requiring almost continuous updating). Furthermore, as the list clearly suggests, data types accessible through LBS would need to be broadened beyond those traditionally the focus of GIS applications. Geographical coverage of data would need to be national, even international. Furthermore, many attributes of interest may not necessarily be resident in databases but may, for example, need to be harvested from Web pages via a search engine or by using software agents. This makes the integration of all data for LBS into a single database repository very unlikely as a business model. Indeed, the trajectory of the underlying technologies is towards interoperable, open systems software (e.g. Buehler and McKee, 1998), the use of distributed services and software components as facilitated by CORBA, Java and .NET (Peng and Tsou, 2003) and the

use of distributed data objects (Tsou and Battenfield, 2002) as an evolution of distributed databases in a networked environment. So whilst some core data sets for LBS, such as road networks, may be kept centrally on a coordinating provider's server(s), all other data will be distributed over intranets and the Internet, for which there will be protocols and agreements covering access, use and charging. In such a distributed environment with large and diverse candidate data sets to be considered, metadata will play a key role (Flewelling and Egenhoffer, 1999; see Section 5.3 below).

Metadata descriptions of the content and quality of data sets are necessary because of the impracticality of browsing through large and distributed data sets in order to ascertain content relevance. Whilst metadata content standards exist (e.g. FGDC, 1997) and form the basis for on-line data clearing houses (Section 5.5), such content advises on data reliability at the point of purchase whilst providing only limited indication of their information potential, when combined with other data sets, to resolve a query with sufficient utility (Brimicombe, 2003). LBS providers will need to compile (and continually update) metadata databases that reflect user perspectives of information utility for a range of user contexts (e.g. daytime vs nighttime) so as to automate rapid data selection and combination.

### 5.3 Data Collection Technologies

---

Activities of surveying and mapping features of the Earth's surface go back at least some 4000 years to the Ancient Egyptians. Systems for the consistent measurement of length, area, weight and time are fundamental to any organized society. As with most areas of science and technology, the microchip has revolutionized the way measurements of the Earth's surface and of the built environment are collected, stored and processed to form useable inputs to GIS. Whilst it is still possible to discern particular approaches to measurement, such as land surveying and remote sensing, the various technologies are being increasingly integrated into digital mapping systems. These systems are increasingly aimed towards automated data collection for the construction and visualization of 3-D models, particularly of urban areas (e.g. Brenner and Haala, 2001). This has ramifications for LBS as 3-D visualizations have the potential for more realistic and intuitive presentation of information to a user (Rakkolainen and Vainio, 2001),

but see Chapter 9 for further discussion. In this section some of the key components that go to make up state-of-the-art systems are described, but the reader must understand that this is a rapidly evolving arena.

### 5.3.1 GPS and Inertial Navigation Systems

Central to nearly all forms of measurement and mapping is the Global Positioning System (GPS). Initiated as a programme in 1973 by the United States with the first satellites launched in 1978 and becoming fully operational in 1993, GPS has become indispensable for geographical positioning and navigation. So much so, that sole reliance on a system owned by a single state and with no guarantees of service is now considered too risky. The Russians of course have developed their own system (GLONASS), the European Union is rushing ahead with its own (Galileo) and other states such as China are also either implementing new or supplementing existing systems to safeguard their own national interests. For simplicity all satellite-based positioning systems will be referred to as GPS. Chapter 6 provides a detailed explanation of GPS in its role of locating the user for LBS. This sub-section will limit itself to a brief introduction.

GPS is based on a constellation of 24 satellites that orbit the Earth at an altitude of approximately 20 200 km. Radio signals are emitted by the satellites over a number of frequencies that can be picked up by receivers regardless of weather conditions, time of day or position on the Earth. The only criteria, and this is an important one for LBS, is that receivers must have an unimpeded ‘view’ of the sky. In other words, GPS receivers operate poorly or not at all under vegetation cover, inside trains, in buildings, tunnels and so on. Generally, the signals from three satellites are required for a dependable 2-D ( $x, y$ ) position fix and from four satellites for a 3-D ( $x, y, z$ ) position fix. Handheld receivers typically have an operational accuracy of  $\pm 3\text{m}$  to  $\pm 15\text{m}$  depending on the configuration of the satellites in view, though most manufacturers provide a range of equipment and accuracies (<http://www.trimble.com>; <http://www.garmin.com>). As with all technologies based on the microchip, GPS receivers have undergone considerable miniaturization such that they can now be integrated into mobile phones.

Inertial navigation systems (INS) are quite different to GPS although they still allow the user to track position from some known point. INS use three sets of gyroscopes and accelerometers carefully



calibrated inside a vehicle (car, helicopter, aircraft) and aligned to the orthogonal axes of the 3-D coordinate system in use (northing, easting, elevation). The vehicle can then travel in 3-D space and have its position tracked by measuring the forces applied in acceleration and changing its position. Coupling INS with GPS allows the INS to have on-going calibration and for both the position and orientation (pitch, roll, yaw) of the vehicle to be tracked with accuracies down to centimetres for position and tens of arc-seconds for orientation. This has implications for remote sensing (discussed below) in that the position and orientation of an imaging or measurement sensor (that is being flown in a helicopter or aircraft) can be known at all times and can thus provide for an automated means to rectify and transform the images into the desired ground coordinate system.

### 5.3.2 Remote Sensing

Remote sensing can be defined as the acquisition of data about objects using a sensing device that does not require direct contact with the objects themselves. Thus the use of a camera to obtain data about objects (as opposed to taking souvenir snaps) would constitute remote sensing. Use of cameras in this way from balloons and aircraft goes back at least a century and had certainly become routine by World War II. Aerial photographs have very high resolution, down to the manhole in the street, and can either be interpreted for the features contained within the images (aerial photographic interpretation (API)), such as the nature of the geology or vegetation, or used for measurement (photogrammetry), such as in the derivation of topographic contours or the mapping of buildings, land parcels and roads (see <http://www.getmapping.co.uk>). For both these uses it is usual to use partially overlapping images which when viewed together permit a 3-D stereoscopic visualization of the landscape. As with nearly all technologies, aerial photography and photogrammetry have moved into the digital age with digitally acquired or scanned photographs being rectified and measured in a semi-automated fashion by software for the fast production of maps.

Satellite imaging for civilian purposes started in the 1960s with meteorological satellites but was quickly followed in the early 1970s by satellites with imaging systems designed to observe the Earth's surface rather than its atmosphere. Whilst the Americans and the Russians

used some traditional film-based cameras from space, the new generation of un-manned satellite imaging devices were wholly digital, so the data could be transmitted back to Earth. The imaging systems generally work by scanning a strip or swathe orthogonal to the direction of orbit with successive scans used to construct the images. They are also designed to be multi-spectral, that is each swathe being split into different bands of the electromagnetic spectrum, most commonly blue, green, red and infrared, thus extending the imaging beyond the visible spectrum. Just as digital cameras are sold today labelled according to the number of megapixels with which an image is captured, satellite imaging is most usefully classified according to its resolution, that is the ground area covered by one pixel of the image. Of most use to LBS are the high resolution imaging systems of satellites such as Ikonos (<http://www.satimagingcorp.com>), which has a pixel size of one square metre (for panchromatic) and 4 m (for multi-spectral), sufficient then to discern large vehicles. Swathe width is 11 km. In other words, to provide coverage for greater London (Figure 3.9b) would require six passes of the Ikonos satellite given that there is lateral overlap of the swathes taken on successive orbits. The imaging system can be tilted  $\pm 30^\circ$  in any direction allowing the acquisition of stereoscopic imagery, which can then be used for 3-D visualization and photogrammetric purposes.

Described thus far has been *passive* remote sensing, in other words, imagery that passively records reflected light from the Earth's surface or off objects. Such systems are limited to daytime operation and, if the Earth's surface is to be imaged, the weather must be cloud and haze free. *Active* systems are those that provide their own source of energy and then record the strength of the reflected signal. Two such technologies of interest for LBS are radar and lasers. Aircraft- and satellite-borne radar has been in use since the 1960s with higher resolutions being deployed in the 1990s onwards. These have a resolution of down to 10 m, and whilst this is a lower resolution than the passive systems described above, it nevertheless has two important attributes: firstly, they can operate day and night (since they generate their own energy source) and, secondly, they can operate in all weathers (since radar wavelengths can penetrate clouds). This provides an invaluable capability, for example to detect and map flooding and other hazards/disasters that occur during poor weather or poor light conditions. LiDAR (light detection and ranging) is a system for laser-based remote sensing. Usually mounted on an aircraft or helicopter with GPS and INS for positioning, LiDAR emits vertically downward pulses of light

and measures the properties of the return signal to determine very accurate measurements of height. Because the pulses are spaced every few centimetres, a very dense data set is collected; this can then be filtered and used to visualize the height and shape of buildings, vegetation and all manner of street furniture as well as the slightest change in topography (see <http://www.earthdata.com>).

### 5.3.3 Ground Survey

Land surveying is the art and science of measuring distance (horizontal and vertical) between objects, measuring the direction of line between objects and the angles between lines. It has been the time-honoured approach to mapping features and boundaries, to calculating areas and sub-dividing areas and to drawing up cross-sections for land management, construction and a host of other applications. It used to be that all maps were compiled using land surveying techniques, but for the last 50 years it has been supplemented first by aerial photography and then by satellite remote sensing to increase the speed and cost-effectiveness with which maps and now digital coverages can be compiled. Nevertheless, land surveying remains the most accurate. Again because of the microchip, together with GPS and the ability to generate and measure the return signal from beams of infrared and laser light, land surveying has been revolutionized almost beyond recognition. The equipment has become considerably automated whilst software is used to carry out the calculations. Land surveying has also become digitally integrated with GPS and remote sensing to form automated systems for data integration and the production of 3-D city models for LBS and other applications (Grejner-Brzezinska *et al.*, 2004).

Whilst land surveying, GPS and remote sensing provide geometric data, field surveys are carried out to sample check (ground truth) automated mapping methods, to collect more detailed attribute data and as a means of monitoring changes to attribute data. Whilst some attributes can be collected during a land survey or from remote sensing, many attributes tend to be collected separately and from a range of sources (see also Section 5.3.4). Key to field surveys these days are mobile GIS deployed using PDA or tablet PCs. This has come about due to the increasing power of PDA, their wireless connectivity, the availability of add-on GPS and the increasing sophistication of the GIS software that can be installed. The position of the field operative can be displayed in relation to base mapping, thematic overlays and

**Table 5.3** SWOT analysis of mobile computing methods for field surveying (adapted from Wagtendonk and de Jeu, 2007 p. 653).

<b>Strengths</b>	<b>Opportunities</b>
Efficiency in data collection	Wireless information retrieval
More uniform data collection	Real-time streaming of field data
Minimized post-processing time	Monitoring and virtual supervision
Higher location and attribute accuracy	Dynamic changes to data collection strategy
Real-time error control and validation	Seamless integration of field and lab/office work
Searchable access to field manuals etc.	Faster data distribution to field operatives
Wireless integration with other devices	
<b>Weaknesses</b>	<b>Threats</b>
More difficult system development requiring specific skills	Insufficiently well-trained field operatives
Development, implementation and maintenance costs	New risks in data loss during transfer and integration
Vulnerability of equipment	Lack of resources for regular updating of systems
	Growing technical dependency

remote sensing imagery; attributes can be entered through a series of customized data entry forms that do preliminary on-site checking of consistency so that gross errors (or blunders) can be rectified before moving on. A SWOT analysis of mobile computing methods for field surveys has been carried out by Wagtendonk and de Jeu (2007) (Table 5.3) because, although these methods would intuitively seem more effective, there is an absence of any convincing proof of either their success in the long term or any estimates of their added value. The study found that whilst there were clear efficiency improvements during the field data collection phase, the largest benefits came in the post-processing phase when traditionally analogue records would have had to have been transcribed to a digital format. Thus a digital approach allows new data to be checked and integrated into the main database more quickly. Another key benefit is the repeatability of data collection where, in a sample checking or monitoring phase, the system brings successive field operatives back to the exact same location (through the GPS and map coverages).

With the growing popularity of in-vehicle navigation systems (SatNav) and the need to maintain the accuracy of the road network data stored within them, companies that furnish such data have come to rely on drive-by surveys. This is a field survey technique in which a vehicle is equipped with high accuracy differential GPS, laptop(s), pen

tablets (for digital note taking), voice recording (also for annotation) and wireless communication. The vehicle is then driven along the road network whilst an operative records changes to the existing database. This is a technique extensively used, for example by Navteq whose European database covers over five million kilometres of road. 'The company has over 300 European field researchers . . . constantly driving the roads updating important details from new housing estates to petrol stations – plus all the latest information on traffic flow changes . . .' (Little, 2006 p. 56). The importance placed on driving the roads, despite the millions of kilometres that need to be covered, is that the information being recorded can be observed and understood from the perspective of a vehicle user.

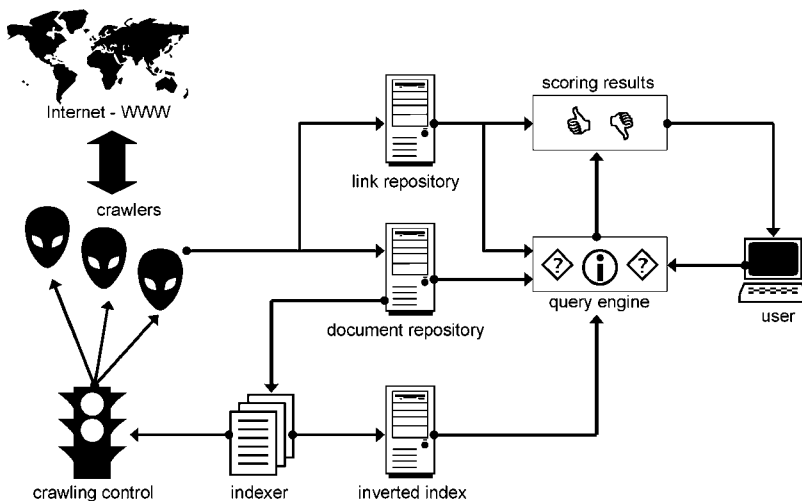
Maintaining a specially equipped vehicle and a crew of two for every 35 000 km of road network in order to keep a database current is expensive and will add considerably to the purchase price of the data. A consortium of data providers and vehicle manufacturers has been exploring ways in which their customers, through a 'cooperative map updating framework', can assist them in maintaining the currency of the information provided (Visintainer and Darin, 2006). The framework currently being explored includes deviation reports (from the known road network) detected through the SatNav's GPS and Web interfaces through which customers can notify data providers of changes to attributes. Research being carried out at the University of East London is also focusing on deploying the SatNav user as a passive data collector (Ekpenyong *et al.*, 2009). Once a deviation from the road network is detected, a user's SatNav switches into data collection mode and records the vehicle's trajectory until it regains the stored road network. The data are then processed by a snap-drift neural network that has been coded onto the SatNav microchip, and classified by road type (e.g. motorway, minor road) and immediately added to the database 'on probation'. When the SatNav is next connected on-line to the provider for database updates, the 'on probation' data is up-loaded to the provider who then matches the data with that from other users, remote sensing imagery and so on in order to confirm the correct road geometry and road type to add to their main database. Getting users to update databases in this way should significantly reduce the cost of updating.

### 5.3.4 Non-Traditional Approaches to Data Collection

The ever growing need for automated and cost-effective methods of spatial and attribute data collection with improved granularity is

fostering the development of non-traditional techniques. CCTV, although technically a form of remote sensing, has for sometime been used as a means of developing traffic bulletins, though it requires an individual to view the images and generate the bulletin. However, a range of roadside and in-road devices using radar, acoustic sensors and infrared detectors are being increasingly deployed not only to count traffic and pedestrians round-the-clock but also to classify the traffic into vehicle types and to determine levels of congestion so as to provide near-instant warnings of events that are happening on our roads. Laurini *et al.* (2001) have classified this type of spatial data collection, where remote sensors telemeter data across fixed-line or wireless links to an operations centre that is monitoring some spatial phenomenon (e.g. weather, traffic, transport of hazardous materials), as TeleGeoInformation.

Another important yet non-traditional approach to data collection is the use of Web crawlers (also known as spiders and Web robots). It is likely that every reader of this book has used an on-line search engine such as Google or Yahoo, if not on a daily basis! Web crawlers are software agents that autonomously navigate the Internet checking known Web documents and links (to make sure they are still active) and to download new documents. This forms the basis of how an on-line search engine works, the architecture of which is illustrated in Figure 5.1. The documents retrieved by the crawlers are stored in the document repository and the links in the link repository. The



**Figure 5.1** Architecture of an on-line search engine (based on Baldi *et al.*, 2003).

documents are then indexed for fast searching whilst the link structure is modelled as a key element in scoring results and determining which documents end up high on the list of responses. Whilst this architecture is a general one for an on-line search engine, it can also be modified so that crawlers can specifically function to search out attributes of, for example, businesses and points of interest.

Finally, wireless networks and the use of mobile devices themselves as a source of data for LBS should be considered. In Chapters 6 and 7 the ways in which users can be located and contextualized are discussed. Where movement is being tracked, even if it is only to hand over communication with a specific mobile device from one network cell to another, updates can take place almost instantaneously. Often, such system data are only of instantaneous interest and not recorded. For LBS, however, data derived from tracking the location of mobile devices can be usefully mined, though its 'shelf life' or currency will be short to medium term. For the medium term, a set of daily tracks, say over a number of months, may be analysed to establish the pattern of an individual's commuter journeys. These stored patterns can then be used to trigger useful alerts should delays be expected along that individual's commuting route. For the short term tracking, imagine a scenario where a number of users of a cellular network are driving along a motorway. Their progress at speeds of between 80 and 130 kph will be clearly evident from the hand over rate from one network cell to another as they progress. If suddenly none of them are progressing to the next cell, then they must be stuck in a traffic jam and an alert issued. Another example would be to use local changes in signal attenuation across a network to inform customers of where it is raining. These are examples of data generated by the operation of the network. On the other hand, many radio stations offering travel bulletins (road, rail, ferry, air) encourage their listeners (who themselves may be stuck in some travel snarl-up) to phone in and report the situation so that other listeners can be more generally warned. Also being seen within the news media is increasing use of mobile phone cameras by the public to capture, on stills and video in real-time, events of public interest. In this way mobile devices are increasingly a means of capturing content.

### 5.3.5 Update Frequencies

Given the dynamic world we live in, most data have a finite shelf life, after which they become too out-of-date for use – unless, that is, they



become historically interesting. Most if not all LBS queries are seeking up-to-date, even up-to-the-moment information. Clearly, there are cost implications of keeping data current and the more often the data need to be checked and updated the more expensive it is going to be. Already mentioned in Section 5.3.3 is Navteq's approach to driving around the roads of Europe in specially equipped cars in order to verify and update its databases. Whilst logistically it would be possible for every road to be revisited once a year, in practice some areas are more likely to experience rapid change than others. In city centres, for example there is a greater likelihood of changes to signage, restrictions, bus lanes and one-way systems affecting drivers than there would be for rural roads. Update cycles take account of this. There is also a subtle link between update frequencies (and the cost of maintenance) and nominal scale of representation. Large scale mapping, such as in Figure 3.11, requires more detail to be surveyed at a higher precision. Small changes can become quite noticeable thus requiring a higher update frequency. On the other hand, small scale mapping, such as in Figure 3.9b, can be highly generalized and only large changes (such as a new motorway) become evident and thus require a lower update frequency to maintain their usefulness.

As a small case study we can look at the attributes of the Ordnance Survey Integrated Transport Network (ITN) data for Greater London (featured in Table 5.2) where changes to the data are recorded. An analysis is presented in Table 5.4, which summarises length of road (carriageway) in metres by class of road. The ITN layer was created in 2002 to supersede a product called OSCAR and, as part of creating a stable and consistent product, an update cycle was initiated. Table 5.4a shows modifications to road segments whilst Table 5.4b shows new segments of roads. The tables seem to suggest that the update cycle covered the bulk of modifications to existing roads in the first three years before inserting new roads. The cycle, however, is more subtle than this. The Ordnance Survey has quality standards stating how soon certain classes of features should have been surveyed and entered into the national mapping database (Ordnance Survey, 2007). Thus 'prestige sites' that require public unveiling or are likely to feature in the media are surveyed and added to the national database before their official opening. 'Category A' features such as changes to or new metalled carriageways are subject to continuous revision and should be entered into the national database within six months of completion. The nominal scale of the mapping is 1 : 1250 and is therefore very detailed, so even slight changes to road geometry

**Table 5.4** Analysis of updates in the attribute table of the Ordnance Survey ITN layer for Greater London; (a) modified segments by length in metres; (b) new segments by length in metres (base data Crown Copyright).

(a) Modified

Year	Total length	Motorway	A road	B road	Local street	Minor road	Private	Alley	Pedestrian
2002	5 629 177	15 621	488 880	149 569	4 197 455	617 247	156 574	1 132	2 698
2003	4 725 899	35 824	547 163	145 604	2 626 117	490 958	445 215	432 659	2 358
2004	3 253 441	42 312	690 698	111 424	1 595 083	228 156	443 307	138 882	3 579
2005	2 850 342	25 301	429 482	78 133	1 354 759	318 050	522 807	118 160	3 650
2006	1 611 184	16 345	158 070	40 406	908 052	167 481	238 141	80 324	2 366
TOTAL	18 070 043	135 403	2 314 293	525 136	10 681 466	1 821 892	1 806 043	771 157	14 651 (a)

(b) New

Year	Total length	Motorway	A road	B road	Local street	Minor road	Private	Alley	Pedestrian
2002									
2003									
2004	20		20						
2005	125 759	417	17 536	6 076	55 470	13 038	21 707	11 389	127
2006	168 240		11 866	2 973	55 964	9 039	59 220	29 160	17
TOTAL	294 019	417	29 421	9 050	111 435	22 076	80 927	40 548	144 (b)

or the introduction of traffic calming features (e.g. sleeping policemen) have to be incorporated. So what Table 5.4b is indicating is that new features are nearly all revisited within three years, subject to modification and therefore appear in Table 5.4a. It is noticeable, however, that the amount of revision required as the product is bedded-in has declined year on year from over five thousand kilometres in 2002 to less than two thousand kilometres in 2006. Another subtle pattern is how the revision of public roads peaks earlier and for private roads peaks later in the cycle. What is also clear is the enormity of the task in maintaining spatial databases for LBS, and what is represented in Table 5.4 is only the roads (let alone the buildings, points of interest and so on) and only for London!

## 5.4 Data Quality Issues

---

In Chapter 3 (Section 3.5) it was noted that uncertainty in information products is a central issue in GIScience. It is very unlikely that any database is going to be 100% correct. Furthermore, the older a database is, the more likely its content will be out of date and therefore in error with respect to the current situation. How much damage can this do? Well, in general, it only takes one glaring error in some information we are trying to use for us to start distrusting it. If we are paying for that information as part of a service (such as LBS) then we might start to have doubts about purchasing any further information in the future. Then again, less obvious errors can permeate information culled from a number of different databases or resulting from the analysis of a number of GIS data layers. For example, suppose two sources of data (such as GIS layers) are both 90% correct, then if we were to combine them with a Boolean AND statement ( $data_A \text{ AND } data_B$ ) then the result may well be only 81% correct ( $0.9 \times 0.9 = 0.81$ ). As the number of data sources or data layers is increased so the likely accuracy of the combined data set will decline. The actual resulting accuracy will depend on the exact nature of the query being resolved, but the example just given raises the importance of data quality for LBS and the need to model the level of uncertainty inherent in information products. In this section sources of data uncertainty and the need for quality statements (metadata) to be welded to data sets are looked at. The reader is also referred to Goodchild and Gopal (1989), Medyckyj-Scott *et al.* (1991), Brimicombe (2003, Chapters 8–10) and van Oort (2006) for a fuller treatment of the subject.

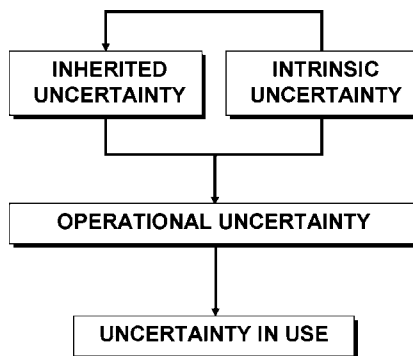
### 5.4.1 Understanding Error and Uncertainty

Firstly, it is necessary to state some definitions so that terms such as error, accuracy and uncertainty can be clearly understood:

- **Error:** is the deviation of data from the truth, or what is perceived to be or accepted as the truth. This assumes that there is an objective truth that can be known and measured, but as we will see below, absolute objectivity can be hard to guarantee. Errors can be further classified as:
  - **gross:** blunders such as typographical errors or accidentally putting the decimal point in the wrong place;
  - **systematic:** a uniform shift or bias in the data such as might arise from using poorly calibrated instruments or rounding up measurements, say, to the nearest metre;
  - **random:** variations arising from repeated measurement of objects and phenomena either due to equipment precision or natural variation; random errors are expected to be normally distributed (bell-shaped curve) about the mean (average).
  - **residual:** the difference between some modelled data (such as by formula or simulation) and the real-world observation of the data in the form of: [*observed* – *expected*].
- **Accuracy:** is the logical dual of error and is the degree of conformance of data with the truth, often expressed as unity minus total error from all sources.
- **Precision:** is the level of consistency with which measurements can be made and is often synonymous with the number of decimal places that can be reliably used. High precision refers to the fact that a set of repeated measurements are close together, whereas low precision would mean that the repeated measurements are widely scattered. Since precision refers to the amount of detail that can be accurately discerned, then high precision is synonymous with high resolution.
- **Resolution:** is the level of detail discernible within the data, which, as discussed in Section 3.4.1, is also linked to the scale at which data are represented.
- **Vagueness:** arises from poor or loose definition, can be linked to low precision and low resolution, and arises particularly where objects and phenomena are *fuzzy* in terms of what is being measured and how to measure them.

- **Ambiguity:** arises where definitions and/or data disagree or conflict, sometimes due to vagueness and data error, but often due to differing expert opinion.
- **Reliability:** reflects the level of trust or confidence one can place in a data set to meet one's needs at the point of access or purchase. The level of reliability is strongly influenced by the metadata.
- **Metadata:** is data about data; statements and tables that accompany data sets that describe the content: how they were collected, by whom, their consistency, completeness and reported accuracy (Section 5.4.3).
- **Fitness-for-use:** whereas the term reliability refers to individual data sets, fitness-for-use refers to the overall assessed quality of information and evidence derived from analysing single or multiple data sets. As in the simple example given above whereby overall accuracy may decline as two or more data layers are combined in some analysis, so depending on the application at hand, the final result may be judged as fit-for-use (such as in decision making) or else the information may be deemed inadequate to the task despite the initial reliability placed on the individual data layers used.
- **Uncertainty:** is an umbrella term referring to the inevitable inaccuracies, inexactness or inadequacies that exist in most data sets and the propagation of these through analyses to raise concern, doubt or scepticism in the mind of the user as to the nature, validity or usefulness of the results.

Four broad categories or sources of uncertainty can be recognized in relation to data services such as LBS, as summarized in Figure 5.2:



**Figure 5.2** Sources of uncertainty (from Brimicombe, 2003).

- **Intrinsic:** this arises during primary data collection, that is when data are being collected in the field and then compiled into data sets. There may be operator bias, poor definition of which features and events are being captured, inappropriate sampling schemes or use of poorly calibrated or faulty equipment. Although a range of quality control measures may be in place and adhered to, as previously stated, it is unlikely that data sets achieve 100% accuracy.
- **Inherited:** this arises from the use of secondary data, that is data sets compiled and made available by third parties. Such data sets will include the intrinsic uncertainty which arose when they were first compiled with additional uncertainties of age (length of time since compiled or last updated) and the way the data are presented (e.g. tabulations, aggregations, suppression of low numbers or personal details to safeguard individual privacy) or the particular definitions of terms used to describe class features (ontology) including variable names that may lead to semantic confusion.
- **Operational:** this arises from the computer handling and manipulation of data, from the way floating points are handled to the algorithmic logic and numerical computation of software. Data may also be corrupted when transferred across networks or from one software to another.
- **Use:** this is by the end user who, on the one hand, may be naïve of the above sources of uncertainty or too trusting of computer outputs or who, on the other hand, despite a healthy level of criticality, may be misled into making sub-optimal or wrong decisions due to the accumulated uncertainties (low fitness-for-use) in the informational products.

If operationally-induced uncertainties can be kept to a minimum through, for example, the use of benchmarked software, then apart from blunders in software use (such as using a wholly inappropriate query to extract data from a database without any form of checking the integrity of the query), the major sources of uncertainty will tend to be intrinsic and inherited. Referring back to Section 5.3.5 on update frequencies, it will be obvious that the age of a data set in relation to the frequency of change in the objects or phenomena being recorded in that data set will have an impact on accuracy and reliability. Most data sets have a ‘shelf life’ or ‘sell by date’ beyond which they are of little interest to LBS, which is mostly concerned with the here and now.

Of course from a scientific perspective such data sets should not be destroyed but archived, as eventually they will have historical interest. Ultimately, in an LBS context, it is up to providers to adequately handle the expected accuracy of their information products and for users to pay more attention to accuracy statements (Section 5.4.2) and the implications of how information is presented (Chapter 9).

### 5.4.2 Assessing Data Accuracy

It should be a duty of data providers to supply accuracy statements about their data sets (Section 5.4.3) and briefly considered here is how such accuracy statements can be derived. There are two important aspects to be considered: one is the accuracy with which an object might be positioned and portrayed graphically (geometric accuracy) and the other is the correctness of the attributes associated with those objects (attribute accuracy). Of course, a data provider may rely on its quality assurance (QA) procedures during compilation (e.g. it may be an ISO 9000 company) and thereby assume that the data sets produced are of consistently high quality. More traditionally (and should also be part of QA procedures) accuracy is established through sample surveys, usually of a higher order, by which is meant that the sample survey is carried out by more accurate means than the main data collection exercise. Thus if the main data collection has been from aerial photography, the accuracy assessment would most likely be carried out by ground survey (often referred to as ‘ground truth’).

For geometric accuracy, sample surveys tend to focus on well-defined points having no attribute ambiguity (Bureau of Budget, 1947; American Society of Civil Engineers, 1983; American Society of Photogrammetry and Remote Sensing, 1985). Results are usually reported as the root mean square error (RMSE), which is a measure, plus or minus, within which the average error will occur (Formula 5.1). Of course, if only unambiguous features have been checked, then this is the accuracy of unambiguous features and more fuzzy features might have a lower accuracy. Published accuracy levels for Ordnance Survey maps are given in Table 5.5. From this table it is clear that smaller scale maps (i.e. with smaller representative fractions) can be expected to have lower geometric accuracy. This is due to the need for higher levels of generalization in order to meaningfully represent features (Section 3.4.3).



**Table 5.5** Geometric accuracy of Ordnance Survey topographic data (reproduced from Brimicombe, 2003).

Nominal scale	RMSE	95% confidence	99% confidence
1 : 1 250	< 0.5 m	< 0.8 m	< 1.0 m
1 : 2 500	< 1.1 m	< 1.9 m	< 2.4 m
1 : 10 000	< 4.1 m	< 7.1 m	< 8.8 m

$$RMSE = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n}}$$

(5.1)

where  $e$  = the residual errors (*observed*–*expected*),  $n$  = the number of observations.

Attributes are usually tested separately to the geometry, though again by sample survey and often with the aid of mobile computing and GPS (Section 5.3.3). At each sample location the attributes are checked and any changes noted. To assess the accuracy, a confusion matrix is constructed (Table 5.6); this plots the mapped

**Table 5.6** Numerical example of a confusion matrix from which values for proportion correctly classified (PCC) can be derived (after Story and Congalton, 1986).

		Reference data			
		Residential	Commercial	Industrial	<i>Total</i>
Classified data	Residential	28	14	15	57
	Commercial	1	15	5	21
	Industrial	1	1	20	22
<i>Total</i>		30	30	40	100
Producer's accuracy		User's accuracy		Overall accuracy	
Residential 28/30 = 93%		Residential 28/57 = 49%		(28 + 15 + 20)/100 = 63%	
Commercial 15/30 = 50%		Commercial 15/21 = 71%			
Industrial 20/40 = 50%		Industrial 20/22 = 91%			

attributes (classified data) against that found in the sample survey (reference data). From this matrix the percentage correctly classified (PCC) can be calculated. The only problem is that the matrix can be read a number of ways. The data provider would look at each attribute class by the columns to derive a producer's accuracy. Thus if 30 residential locations were sample checked and only two were wrong then the PCC for 'residential' is 93%. On the other hand, a user would note that only 28 residential locations out of 57 on the map were correctly classified and therefore the user's accuracy for 'residential' is down to 49%. An overall measure of PCC takes the sum of the diagonal divided by the total number tested.

There are, of course, a number of other ways of calculating the accuracy of geometric and attribute data, but given here are the most frequently encountered.

### 5.4.3 Metadata

The compilation of data sets, particularly spatial data sets, is time consuming and expensive. Re-using data in projects or applications beyond that for which they were originally collected is therefore attractive. However, how does one find and then quickly establish the pertinence and reliability of a data set to one's needs. Finding data is partly to do with tapping into infrastructures (Section 5.5) but decision making about data use largely revolves around the use of metadata. *Metadata* is literally data about data (Lillywhite, 1991), that is higher level information about a data set. Good metadata should be an authoritative guide to data provenance, ownership (or stewardship), content, compilation and accuracy or reliability. There are standards that guide the production of metadata. In Europe, for example there is *ISO 19115 Geographic Information – Metadata*, which each member state then implements through its own local standards, so that in the United Kingdom there is the *GEMINI Discovery Metadata Standard* and the *e-Government Metadata Standard*. There is even a downloadable tool that will assist data producers/guardians in the creation of metadata: MetaGenie from <http://www.gigateway.org.uk/metadata/metagenie.html>. In the United States there is the *Content Standard for Digital Geospatial Metadata* developed in 1994 by the US Federal Geographic Data Committee; this is mandatory for all federal agencies. Metadata file(s) should be automatically provided with data sets at the point of search and/or access.

The best way to understand the role and value of metadata is actually to study a metadata statement that accompanies a data set. The example chosen is the *London Household Survey 2002*, which can be accessed through the UK Data Archive ([www.esds.ac.uk](http://www.esds.ac.uk)). This is a large data set (despite it being only a 0.27% sample of households) such that cursory browsing of the data is both difficult and not very instructive. It actually contains two data sets, one for household responses and another for the individuals interviewed (2–3 persons per household). So one file has 773 variables for 8153 households giving 6.3 million data entries; the other has 685 variables for 20 910 individuals giving 14.3 million data entries. Not the sort of data sets one can casually browse and gain an insightful overview! There is a 573 page user guide, which is a daunting prospect if you want to discover whether the data would be useful for some application or not, but clearly indispensable if the data are to be accessed and analysed in detail. Extracts from the metadata statements accompanying the data are given in Tables 5.7, 5.8 and 5.9.

**Table 5.7** Extract from the metadata for data set 5149 at the UK Data Archive (copyright UK Data Archive).

---

5149 Greater London Authority (GLA) Household Survey, 2002 (London Household Survey, 2002)

Depositor:

University of Manchester. Cathie Marsh Centre for Census and Survey Research

Principal Investigator:

Greater London Authority. Data Management and Analysis Group

Data Collector:

[an individual's name given]<sup>a</sup>

Sponsors:

Housing Corporation

Greater London Authority

Other Acknowledgements:

[another individual]<sup>a</sup> at the Cathie Marsh Centre for Census and Survey Research (CCSR), University of Manchester, carried out some work on the survey to prepare it for archiving.

Abstract:

This survey was commissioned by the GLA and undertaken by [an individual's name given]<sup>a</sup>. Over 8150 interviews were achieved, sufficient for detailed data analysis at London city level and less detailed analysis at 'groups of boroughs' level.

No specific report of findings has been written, because the survey is intended primarily to support policy development, including the linkages between multiple aspects of needs. The data have so far been used by a wide range of GLA policy teams, government research analysts and academics. The survey has also formed the basis for follow-up interview surveys (based on the 75% + of respondents who gave permission at the end of the main interview), on topics such as access to e-governance.

### Main Topics:

The survey covered a wide range of policy areas in moderate detail – household and personal characteristics (including income and savings), employment, transport, crime, health, disability, housing needs, moving intentions and history, use of the Internet and access to standard lifestyle commodities.

### Coverage:

Dates of Fieldwork: February 2002–August 2002

Country: England

Geography: Greater London

Spatial Units: Electoral Wards/Divisions (England); Postcode Sector; Local Authority Districts

Observation Units: Individuals

Kind of Data: numeric data; Individual (micro) level

Universe Sampled:

Location of Units of Observation: subnational

Population: residents of London aged 18 and over during 2002

### Methodology:

Time Dimensions: cross-sectional (one-time) study

Sampling Procedures: multi-stage stratified random sample

Number of Units: Target: 8000. Obtained: 8158.

Method of Data Collection: face-to-face interview

Weighting: weighting used. See documentation for details [5149userguide.pdf]

Language(s) of Written Materials:

Study Description: English

Study Documentation: English

### Access:

Access Conditions: the depositor has specified that registration is required and standard conditions of use apply. The depositor may be informed about usage. See terms and conditions for further information.

Availability: ESDS Access and Preservation, UK Data Archive

Contact: Help desk: [help@esds.ac.uk](mailto:help@esds.ac.uk)

Date of First Release:

7 April 2005

Copyright:

Greater London Authority

---

<sup>a</sup>Individuals' names have been suppressed from the metadata statement for publication in this book.

**Table 5.8** Extract from the data processing notes for data set 5149 at the UK Data Archive (copyright UK Data Archive).

UK DATA ARCHIVE: IMPORTANT STUDY INFORMATION
Study Number 5149 – Greater London Authority (GLA) Household Survey, 2002
DATA PROCESSING NOTES
Data Archive Processing Standards
The data were processed to the UK Data Archive’s B standard. A substantial series of checks was carried out to ensure the quality of the data and documentation. Firstly, checks were made that the number of cases and variables matched the depositor’s records. Secondly, logical checks were performed on a sample of 30 + 10% of the remaining nominal (categorical) variables to ensure they had values within the range defined (either by value labels or in the depositor’s documentation). Thirdly, any data or documentation that breached confidentiality rules were altered or suppressed to preserve anonymity.
All notable and/or outstanding problems discovered are detailed under the ‘Data and documentation problems’ heading below.
Data and documentation problems
None encountered.

**Table 5.9** Extract from the variable definitions for data set 5149 at the UK Data Archive (copyright UK Data Archive).

Pos. = 10	Variable = hhldtype	Variable label = NEW household type variable
This variable is numeric, the SPSS measurement level is ordinal.		
Value label information for hhldtype		
Value = 1	Label = 1 pensioner	
Value = 2	Label = 1 non-pensioner	
Value = 3	Label = All pensioner couples	
Value = 4	Label = Married and no ch	
Value = 5	Label = Married and 1 dep ch	
Value = 6	Label = Married and 2+ dep ch	
Value = 7	Label = Married and all nondep ch	
Value = 8	Label = Cohab and no ch	
Value = 9	Label = Cohab and 1 dep ch	
Value = 10	Label = Cohab and 2+ dep ch	
Value = 11	Label = Cohab and all nondep ch	
Value = 12	Label = Lone male and 1 dep ch	
Value = 13	Label = Lone male and 2+ ch	
Value = 14	Label = Lone male and all nondep ch	
Value = 15	Label = Lone female and 1 dep ch	
Value = 16	Label = Lone female and 2+ ch	
Value = 17	Label = Lone female and all nondep ch	
Value = 18	Label = Multi hhld and 1 dep ch	
Value = 19	Label = Multi hhld and 2+ dep ch	

Value = 20	Label = All students
Value = 21	Label = All pensioners (not cples)
Value = 22	Label = Other
Value = 23	Label = Unresolved

Pos. = 11    Variable = pension    Variable label = At least 1 pensioner in  
household

This variable is numeric, the SPSS measurement level is ordinal.

Value label information for pension	
Value = 1	Label = Yes

Pos. = 12    Variable = depchild    Variable label = At least 1 Dependant child in  
household

This variable is numeric, the SPSS measurement level is ordinal.

Value label information for depchild	
Value = 1	Label = Yes

The first part of the metadata statements (Table 5.7) gives details of where the data is deposited, who carried out the survey and which organization(s) paid for it. There then follows an abstract giving an overview of the survey, the geographical level at which it should be used and a summary of the types of applications for which the data have been used. Then comes a statement on the topics addressed ‘in moderate detail’ through the interviews. Next there is some important information about when and how the survey was carried out and the methodology by which the data were obtained. Finally in this section are the conditions for access and who the data copyright holders are. The second extract (Table 5.8) illustrates how metadata can tell a potential user something of the accuracy or, in this case, the reliability of the data. Some independent checks have been carried on data quality with emphasis on logical checks to ensure consistency of the data. The third extract (Table 5.9) demonstrates the need for a data dictionary that allows the user to identify what each variable is, how it has been coded and where it is placed within the data set. Thus, whereas variables 11 and 12 on the presence of pensioners and dependent children are a simple yes/no answer, variable 10 on household type can have one of 23 different answers. Notice that this latter variable is labelled ‘NEW’, alerting the user, who might be comparing this survey with an earlier one, that this variable has been inserted or changed and is therefore not directly comparable with any household type variable in earlier surveys, so helping users avoid what might be a

simple mistake but which could be seriously detrimental to the fitness-for-use of any analytical products.

To summarize, the main elements that need to be addressed through metadata are:

- **ownership/guardianship:** who owns the data (copyright holder), who might have taken over the guardianship of the data and where the data are deposited;
- **lineage:** the provenance of the data, including the methods employed and which organizations were involved;
- **coverage and resolution:** which geographical areas are relevant and what are the spatial and temporal resolutions applicable to the data;
- **quality:** whether or not geometric and/or attribute content has been tested for accuracy as well as the degree of completeness and the logical consistency of both the data and accompanying documentation;
- **purpose, usage and constraints:** the original purpose for collecting the data, how they have been used by the owner and, importantly, any known or advised constraints on the use of the data in any new applications.

## 5.5 Organizing and Accessing Data

---

In the early 1990s there started to be a growing recognition that ad hoc access to and integration of spatial data sets from across a range of providers was less likely to adequately serve the needs of society than if organizational arrangements were put in place to facilitate standardized, consistent services for access. This, in turn, would stimulate the production of good quality spatial data and a virtuous cycle would emerge that had overall benefits for society. Such a strategy, it was believed, would help promote economic development, provide for better governance and support moves towards better environmental sustainability. Thus was born the concept of a Spatial Data Infrastructure (SDI). This is intended to be an environment in which quality assured data sets can be accessed and retrieved by users in an easy and secure manner (Coleman and McLaughlin, 1998). SDI has clear implications for LBS because any service provider (or service broker) in satisfying a query is likely to have to extract the relevant



elements from a number of distributed data sets (Smith *et al.*, 2004). For SDI to be achieved, four things need to be in place:

- **communication networks through which data are accessed:** this would encompass both the Internet and wireless communication networks, an important consideration for LBS being the speed and scalability of service;
- **policies regarding data availability, granularity and use:** this would include agreement on which data sets are made available, for which purposes they can be used as well as issues regarding guardianship and responsibility for updating and safeguarding privacy;
- **data quality, interoperability and transfer standards:** to ensure correctness, completeness and consistency of data, the degree of integration that can be achieved and the format (e.g. XML) in which data can be accessed;
- **contracts and protocols regarding access, copyright and pricing:** governing the acquisition and mechanisms of payment (where necessary) by approved users.

SDI can be regarded as organizationally hierarchical from individual companies and agencies, which tend to be the data producers or brokers of individual data sets, through local government up to national and then global levels (Chan and Williamson, 1999). Evidence for the emergence of SDI lies in the plethora of Web gateways or portals that now exist through which spatial data can be accessed. Where these portals are exclusively for serving spatial data they are termed *geoportals* (Maguire and Longley, 2005; Tait, 2005). Two such geoportals can be found at <http://www.geo-one-stop.gov> and [www.gigateway.org.uk](http://www.gigateway.org.uk). The function of geoportals is to organize content, provide directories and search tools, allow data browsing and visualization; some provide basic data integration tools. Underlying most geoportals is a metadata catalogue through which data suitability (e.g. coverage, scale, theme) can be queried and candidate data sets listed. The detailed metadata records can then be accessed and studied. Key to any geoportal and the wider aspirations to SDI is the quality of the metadata content of which there may be thousands of published entries in a geoportal's metadata catalogue. Administration and quality assurance of these metadata statements are critical. Currently, the main means of access to geoportals is by the individual user. This we believe will change with increasing use of agent-based software (Section 3.3.2) acting on behalf of LBS or other service providers.

## **Location-Based Services and Geo-Information Engineering**

These would be automatically launched on receipt of a query in order to find and retrieve through geo- and other portals the relevant information required to provide a response. This then will require new forms of protocols and contractual arrangements to be agreed between portal administrators and service providers.

# Chapter 6

## Locating the User

### 6.1 Introduction

---

One essential aspect in LBS is fixing the location of users through their mobile device. Location of users is regarded as the spatial context in LBS applications. Such location data are often used as a key in real applications, from which the overall context informs mobile applications. In this chapter the various technologies currently available to provide location data about mobile devices used for LBS are discussed; these include satellite positioning systems (the most commonly used being the United States operated Global Positioning System (GPS)), positioning methods based on mobile telecommunications networks (network-based), technologies used for short range positioning (such as Bluetooth and tags), and hybrid positioning solutions. The principle and characteristics of each technology are described as well as their strengths and weaknesses.

There are a number of issues that need to be borne in mind when considering different technologies deployed for locating the user in LBS applications. An important consideration is both the accuracy and the consistency of the location data obtained. The level of positioning accuracy can vary depending on the type of technology used; and so does the consistency of the location data achieved. For example, GPS can provide sufficiently accurate positioning data where there are clear lines of sight to the necessary number of satellites; however, the accuracy can drop quite dramatically when in an urban environment with high

rise buildings (urban canyon effect). On the other hand, using a network-based technology such as Cell-ID, the accuracy of positioning data can also vary from one place to another depending on the spacing of masts and hence the size of cell serviced (Section 6.4.1). Stability of different positioning methods, which concerns both accuracy and consistency, will have a direct influence on the level of services that can be provided through LBS. Another issue concerns latency. *Latency* refers to the time period used to locate the position of a device and is an inherent characteristic of positioning technologies, which can be important for certain types of LBS applications.

Yet another issue relates to overheads and consumption of power. Overheads are both the volume of messages exchanged between a mobile device and the system (signalling overhead) and the amount of processing required at the mobile device and in the system for a position fix (computational overhead). The overheads affect the consumption of power in a mobile device while obtaining its position. High power consumption uses more of the device's battery resources and leads to a shorter usage time. It is important for mobile devices to conserve power given current battery technology. Thus there needs to be a pragmatic balance between power consumption, positioning accuracy and coverage. Otherwise, it can detract from the use of LBS. Furthermore, different levels of cost and requirements on devices, networks and infrastructures need to be considered when using different types of positioning technologies. These issues will be discussed in relation to each positioning technology in the sections that follow.

The purpose and type of LBS application determines what level of accuracy and consistency of position fixing is acceptable. User location data can be classed into different levels of granularity, which is closely related to different types of applications. Location data can be point-oriented, line-oriented and area-oriented. Point-oriented location data with high accuracy can be at a single point with coordinates, or for a very small area with an indicator of accuracy (e.g.  $x, y, \sigma$ ; where  $\sigma$  might be the RMSE – see Section 5.4.2). Position fixing to a point can be used for applications such as emergency services and other services requiring an accurate fix of user location. Line-oriented location data can be used for LBS applications with line-based geographic features (i.e. roads, rails and rivers), such as vehicle/fleet management services. Area-oriented location data can have different levels of accuracy (medium to low) depending on the area covered. Such location data can be sufficient for those LBS

applications that only need to know whether users are within a certain service area, such as providing local visitor information, weather reports, tariff services and congestion alert services.

Thus, depending on the service requested by a user, the positioning technology might be switched, changing the overheads and latency along with the accuracy and consistency. However, it needs to be pointed out that the accuracy of position fixing is not the only factor needed to ensure that applications deliver more useful information with less error. For example, there are recorded instances of vehicles being driven up the wrong side of motorways and other dual carriageways because drivers misread the in-car navigation system (SatNav) despite the accurate positioning of the car on the screen. This results from a confusion or misinterpretation of what is being indicated to the driver despite being accurately positioned (see Section 10.5.3 for further discussion). Thus other aspects of user context play an important role alongside location data (Chapter 7) as well as the way in which information and services are communicated to the user (Chapter 9) in order to assure LBS applications.

## **6.2 Positioning Technologies**

---

There are a number of position fixing technologies as well as combined hybrid solutions to locating users. Positioning methods can be generally categorized into network-based, device-based and hybrid methods. They can also be differentiated as to whether they are integrated or stand-alone positioning infrastructures, and again depending on whether they are satellite-based, network-based (cellular) or indoor infrastructures. Device-based positioning is usually based on GPS and hence both a satellite-based and stand-alone infrastructure. Network-based positioning can be regarded as integrated, as the network is also used for communication and data transmission with gateways to other networks and systems. Various positioning technologies are discussed based on the categories of device-based, network-based and hybrid integrated. In addition, a range of technologies which can be used to locate users within a comparatively short range are described.

For device-based positioning (also known as terminal-based or receiver-based positioning), the mobile device determines its position using signals it receives. In other words, signal measurements and the

computation to determine a position are performed by the receiver located within a mobile device. No network connection is required. There is more privacy in using device-based methods. However, there is high power consumption for mobile devices to carry out positioning and positioning stops once signals can no longer be received (such as indoors or in tunnels). Global Navigation Satellite Systems (GNSS) are considered as a class of device-based positioning. The GPS is the most well known of these. Other similar systems are the Russian GLONASS system, the European Galileo system that is being designed for civilian purposes and scheduled to start operation after 2009, and the Chinese Beidou (Compass) Navigation Satellite System that is currently being trialled and will soon be operational. The basic principles of GPS technology, which is similar to all other GNSS, are described in Section 6.3.

Network-based positioning methods use the transmitter base stations of a mobile telecommunications network to locate a mobile device by measuring the signal travelling to and from a set of base stations. Through signal measurements, the direction and/or length of an individual radio path can be computed and the position of a mobile device can be derived using computational geometry. Connection to server-side services is needed to position mobile devices when using network-based methods. There is low power consumption and fewer requirements made of the mobile devices. Usually, a range of network-based positioning methods (Section 6.4) can be operated using mobile telecommunication networks. The networks are used both for positioning and communication and must share the bandwidth. Such network-based positioning can work in indoor environments as long as there is sufficient signal strength. If position data with high accuracy are required, there can be high costs in terms of signalling overhead and thus reduced capacity for voice/data transmission. One of the main drivers for the development of network-based positioning methods is the E911 mandate (Section 4.4.1) that requires network operators in the United States to provide mobile user location data to specified accuracies.

A range of other technologies can also be used for positioning, particularly for comparative short range positioning. Such technologies include Bluetooth technology, Radio Frequency Identification (RFID) and Wireless Local Area Networks (WLANs), all of which are described in Section 6.5. Various hybrid positioning approaches are discussed in Section 6.6. In some hybrid positioning solutions, measurements are carried out both in the device and on the network.

Such positioning is not entirely device-based and should more properly be regarded as device-assisted.

## 6.3 Global Positioning System

---

The Global Positioning System (GPS) is a satellite-based global radio navigation system. It was primarily developed as a defence system by the US Department of Defense. The first satellite was launched in 1978, and the system started to be fully operational in 1993. GPS consists of 21 satellites plus three spare ones orbiting the Earth every 12 hours. The GPS satellites are in orbit at high altitude (about 20 200 km). The technology used by GPS enables the system to fix any position in the world at any time of the day when there is a clear view of the necessary number of satellites. The GPS configuration assures the visibility of five to eight satellites at any point on the Earth at the same time. With the ‘differential’ mode (Differential GPS (DGPS), Section 6.3.3), high accuracies at sub-centimetre levels can be achieved. Prior to May 2000, GPS signals from the satellites were selectively degraded to reduce the accuracies available for civilian use, and although this policy has been abandoned it could still be re-introduced. Currently, the GPS provides both military and civilian positioning services. Over the years, GPS receivers have become small enough and cheap enough to be used by the general public as handheld gadgets or incorporated into mobile phones. The GPS can thus be used in a wide range of applications, such as precise positioning, navigation, surveying, mapping, engineering and of course for LBS. The general design principles of the GPS can be summarized as follows:

- be suitable for all classes of platforms whether they be aircraft, ships or land-based vehicles as well as other satellites;
- provide positioning in real-time along with the ability to determine the time and velocity;
- reference all positioning to a single global geodetic datum;
- be capable of preferentially providing higher accuracies to certain classes of users (such as the military);
- be scalable to unlimited numbers of users worldwide;
- have low unit cost (it is in fact free to use) with low power overhead.

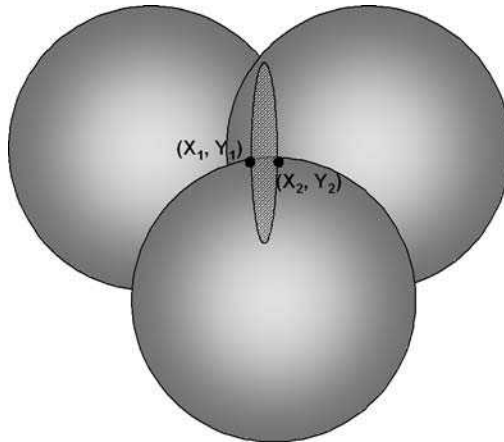


Many LBS applications (such as in-car navigation) are supported by GPS positioning. Locating users through GPS-enabled devices is a device-based positioning method. A GPS-enabled mobile device means that a satellite navigation receiver is built into the device. Devices with built-in GPS can offer a higher degree of location accuracy that can be within  $\pm 3$  to  $\pm 15$  metres. However, such accuracy may not be consistent in all environments. The accuracy may vary from rural areas to high-rise urban areas due to its need for a clear view of the sky and signals from three or four GPS satellites. There are still other disadvantages for such devices used in mobile situations, such as the higher power consumption and an increased cost in the manufacture of the device.

### 6.3.1 Basic Principles of GPS

The basic principle of positioning using GPS is trilateration based on distance measurements (range measurements), using satellites as reference points, although there have been advances in the data processing over the years since the system became operational as well as enhancements to the satellite constellation, such as the Wide Area Augmentation System (WAAS) which increases accuracy. A position on the Earth is determined by the distance measured between the GPS receiver at a particular position and a range of satellites. Mathematically four such measurements are needed to calculate an exact position. Three simultaneous measurements from three different satellites can narrow down the position to two points on the intersection of three spheres (Figure 6.1). Usually one of these two points is an impossible (incorrect) point which can be distinguished and rejected by the algorithms coded into GPS receivers. Therefore, in theory, three satellites ranges are sufficient to determine a position. In practice, the fourth satellite is used as an additional measurement to eliminate what is known as receiver clock offset. In situations where the altitude is already known (e.g. at sea level for mariners), the position can be determined using one less satellite measurement.

The distance from a position on the Earth to a satellite is determined by the time taken for the radio signal to travel between that satellite and the GPS receiver: *distance = speed of light* (approximately 300 000 km/hr)  $\times$  *travel time*. The radio signal generated by GPS includes carrier signals, Pseudo-Random Noise (PRN) codes and navigation messages. GPS signals are described in Section 6.3.2.



**Figure 6.1** Two position points determined by three satellite measurements.

In order to measure the radio signal travel time between a GPS receiver and a satellite, it is essential to establish the exact time the signal is sent from the satellite. For this purpose, both the satellites and the GPS receiver are synchronized by generating the same code, referred to as Pseudo-Random Noise (PRN) code, at exactly the same time. When the PRN signal from the satellite arrives to the receiver, this PRN sequence is compared with the receiver's PRN sequence. These two PRN sequences will not be synchronized due to the delay in the satellite PRN signal arriving because it has had to travel about 20 200 km. The receiver then delays its own PRN signal until these two PRN sequences are synchronized. This delay time of the receiver's signal is equal to the travel time of the satellite's signal. This is, in principle, how the radio signal travel time between a satellite and a receiver is measured by employing PRN code. The radio signal travel time is then converted to a distance measurement by multiplying it by the speed of light. Although the PRN code looks almost like a long string of random pulses, it is actually generated as 'pseudo-random' sequences repeated every millisecond. The term *pseudorange* refers to the distance (range) measurement just discussed. Another mathematical model for distance measurement is *carrier phase* ranging which provides a higher level of accuracy in positioning. These measurements are described in more detail in Section 6.3.2.

Accurate timing is essential to achieving accurate distance measurements between a receiver and satellites, and consequently in

achieving an accurate position using GPS. The time used by satellites is generally very accurate due to the atomic clocks installed on board. However, the clocks installed in GPS receivers have less accuracy and may result in errors in measuring the signal travel time. This is known as *receiver clock offset*. By introducing a fourth satellite signal, trigonometry can be used to estimate and eliminate the receiver clock offset. There are four unknown parameters including three for coordinates in a GPS-based position fix and one for the receiver clock offset. The processors in the GPS receivers apply algebra to the problem, and the clock offset is readily computed. If three perfect measurements locate a GPS receiver in a 3-D space, then four less perfect measurements can eliminate any timing offset as long as the offset is consistent. Thus, four measurements are required. When there are more than four satellites visible, the optimal solution is derived by using a least squares solution. The design of GPS receivers can be affected for these four measurements. For a single-channel receiver, it has to sequence through four separate measurements for four different satellites before it can calculate an answer. It can take between 2 and 30 seconds to fix a position and is not suitable, say, for monitoring velocity.

The positions of satellites in space are also required in order to calculate the position of a receiver in addition of knowing the distance between satellites and the receiver. The satellites in GPS are at the altitude of about 20 200 km (11 000 miles), and thus their orbits can be predictable. The position of each satellite in its orbit can be obtained at any time by GPS receivers on the ground with an *almanac* stored in the memory of the processor. The almanac is a data file that GPS receivers can use to obtain the satellite position. It includes orbit information of all satellites in GPS, clock correction and atmospheric delay parameters. Satellites have a tendency to drift from their predicted orbital positions mainly due to *ephemeris* errors. These errors are caused by the gravitational pulls on the satellite from the Sun and the Moon, and also by the pressure of solar radiation. Such variations in satellite orbit are usually very minor. These minor variations in satellite altitude, position and speed are constantly monitored and measured by the US Department of Defense. Such information, including all minor corrections, is sent back to the respective satellite and is in turn broadcast to GPS receivers. Satellites in the GPS constellation pass over specially sited ground monitoring stations twice a day. In this way the accuracy of the predicted orbit of all GPS satellites is kept typically to within a few metres (Grejner-Brzezinska, 2004).

The positioning accuracy of GPS can be influenced by a number of other factors. One type of error is caused by the delay in satellite signals when travelling through the Earth's ionosphere. The ionosphere is a layer of electrically charged particles (ions) about 30–300 miles above the Earth. Electromagnetic signal transmissions can be interfered with and distorted by these particles in the form of a nonlinear dispersion of the transmissions affecting their speed, frequency and direction. Therefore, errors can occur in measuring signal travel time. These errors can be eliminated or mitigated to a certain degree with mathematical modelling approaches. By looking at the relative speeds of two different signals, the variation in the signal speed can be measured. The basic idea is this: when light travels through the ionosphere, it slows down at a rate inversely proportional to its frequency squared. Thus the lower the frequency of the signal, the more it gets slowed. By comparing the arrival time of two different parts of the satellite's signal with different frequencies, it is possible to deduce what kind of slowing effect they must have experienced. This type of error correction is sophisticated and can only be carried out on the most advanced *dual-frequency* (dual-channel) GPS receivers, and not on single-channel receivers.

Errors can also be caused by the delay in GPS satellite signals travelling through the Earth's troposphere, where signals can be affected by the water vapour in the atmosphere. The troposphere is the portion of the atmosphere containing the weather systems and hence water vapour and clouds. The errors caused by the troposphere are similar in size to those of the ionosphere. However, the delay is not frequency dependent. In another words, the amount of delay in GPS signals with different frequencies can be the same. Thus, the dual-frequency correction approach cannot be applied here. To eliminate the tropospheric effect, empirical models which consider temperature, pressure and relative humidity are often used. Such models can reduce the major part (90–95%) of the tropospheric effect.

Another error source in GPS is *multipath propagation*. This is where satellite signals interact with objects (such as buildings, water bodies and even the ground itself) to produce multiple reflections and diffractions such that the same signal arrives by both direct and indirect paths. Multipath propagation can cause these signals to interfere with each other and distort the amplitude and phase of the signals. Therefore, errors occur in the distance measurements of signals, and the positioning accuracy is reduced. The extent of the multipath effect is likely to be random and unpredictable, depending on

satellite geometry and the type of reflective surfaces in the receiver surroundings. A properly designed choke ring antenna can reduce the multipath effect for the waves from surfaces and for signals reflected from the ground.

There are several other sources of error that can reduce the positioning accuracy of GPS. The accuracy can be affected by the relative angles in the sky between the satellites used for positioning, which is often referred to as Geometric Dilution of Precision (GDOP). Usually, the wider the angle between the satellites relative to the GPS receiver, the higher accuracy in position fixing that can be achieved. If for example the satellites being used by a receiver all line up in the same direction or are too clustered then the accuracy will tend to be poor. It is not that one satellite is better than another. The geometry can magnify or lessen all the uncertainties. A GDOP value of six and less indicates good geometry. There are also other dilution factors, such as Position DOP (PDOP), Vertical DOP (VDOP) or Horizontal DOP (HDOP). All these dilutions of precision are normally computed by GPS receivers and provided in real-time as a quality assessment.

Finally, an historical source of error in the measured range to the satellite was Selective Availability (SA). As mentioned above, SA resulted from the United States Department of Defense policy of denying to non-military GPS users the full accuracy of the system. The effects of SA could be mitigated through differential techniques (Section 6.3.3). However, SA levels have been set to zero since 2 May 2000 at 04:00 UT (Universal Time) with immediate noticeable improvement in civilian use of GPS. Other error sources in GPS can be error potential in satellite clocks and GPS receiver noise. For more information on GPS error sources, readers are referred to Lachapelle (1990) and Hofman-Wellenhof *et al.* (2001).

### 6.3.2 A More Detailed Look at GPS

#### 6.3.2.1 System Components

GPS consists of three essential segments: the satellite segment, the user segment and the control segment (Rizos, 2002). The satellite segment is the satellite constellation of 24 satellites (including three spare ones) orbiting the Earth every 12 hours. To provide positioning with the coverage of the whole Earth, the minimum number of satellites required is 21. Three spare ones have been put in orbit in case one or

more of the constellation fails or needs servicing. These 24 satellites circulate the Earth on six orbits with four satellites on each orbit. Each orbit has a 55 degree inclination angle and there is a 60 degree separation between each orbit.

The user segment can be viewed as all the GPS receivers employed in positioning applications. The types of applications for which GPS are used have an impact on the receiver configuration. As GPS is used for positioning in a wide range of applications, requirements on the user segment (receivers) can vary. For the precise positioning used in surveying, mapping and engineering, positioning accuracy will be a primary requirement for GPS receiver configuration, whilst for other applications requiring, say, a fast response, the latency time for first position fix (Time To First Fix – TTFF) might be the primary requirement. For LBS applications, accuracy requirement for position fixing can vary widely as LBS cover such a broad range of applications (Section 4.7). Thus, various requirements for receivers such as TTFF, battery consumption and device size have to be considered along with positioning accuracy. Therefore, the purpose and type of LBS application can have an influence on requirements of the user segment.

The control segment of GPS comprises a number of monitoring stations (including one master control station) which maintain the system operability including monitoring and determining the satellite orbital positions. There are currently five monitoring stations located across the world to cover each GPS satellite most of the time. The five stations are at Hawaii, Colorado Springs, Ascension Island, Diego Garcia and Kwajalein (Küpper, 2005). As discussed in Section 6.3.1, the orbital positions of the GPS satellites can have variation primarily due to the ‘ephemeris’ errors. Therefore, the satellite positions are constantly monitored by the control segment to determine their exact position in space. Control messages are transmitted back to each GPS satellite by the control segment.

### **6.3.2.2 Signal Structure**

The primary measurement used for positioning in GPS is the radio signal travel time between satellites and receivers, from which can be derived a distance measurement as discussed in Section 6.3.1. The radio signal generated by a GPS satellite oscillator includes three basic components: carrier waves (pure sinusoidal waves) L1 and L2, Pseudo-Random Noise (PRN) code and navigation message. All

signals transmitted by the GPS satellites are coherently derived from a fundamental frequency ( $f_0$ ) of 10.23 MHz. The L1 carrier has the frequency of  $154 \times 10.23$  MHz ( $154 \times f_0$ ), whilst the L2 carrier the frequency of  $120 \times 10.23$  MHz ( $120 \times f_0$ ). The carrier modulation enables the measurement of the radio signal travel time between a satellite and a receiver.

PRN code has two types: precise P(Y)-code and coarse-acquisition C/A-code. P(Y)-code (P-code and Y-code) is modulated on the L1 and L2 carriers, which is a dual frequency positioning solution standard for high precision geodetic applications. C/A-code is superimposed on the L1 carrier only, which is a single frequency type used mainly for civilian GPS receivers with a moderate accuracy level. Access to the C/A-code is free to all users and is known as the Standard Positioning Service (SPS). SPS has a guaranteed horizontal positioning accuracy of  $\pm 100$  m or less (95% of the time) and vertical positioning accuracy of  $\pm 156$  m or less (95% of the time). The Precise Positioning Service (PPS), which uses both C/A and P/Y codes and is available exclusively to the military, uses an Anti-Spoofing (AS) policy in which an additional W-code is implemented to encrypt the P-code to the Y-code. PPS has a guaranteed horizontal positioning accuracy of  $\pm 22$  m (95% of the time) and vertical position accuracy of  $\pm 27.7$  m.

The navigation message in GPS is the data message needed for a real-time positioning in a GPS receiver. The navigation message includes the satellite ephemeris, clock correction data, almanac and other information about the state of satellites and their signals. The navigation message is contained in the GPS signals and transmitted from the satellites to receivers.

### **6.3.2.3 Signal Measurement**

Two fundamental types of range (distance) measurement methods used in GPS are: pseudorange and carrier phase. In general, pseudorange measurement is used as a standard method whilst carrier phase ranging is applied to achieve very high positioning accuracy. Pseudorange is a geometric range between the satellite transmitter and a GPS receiver. As discussed above, such measurement becomes distorted by the atmosphere and other effects and by the lack of synchronization between the satellite clocks and the GPS receiver clock. The distortions are recovered from the measurement of differences between the time of signal transmission from the satellite and the



time of its reception by the GPS receiver. This time measurement is carried out using the PRN code (Section 6.3.1). There are two types of pseudorange: P-code pseudorange and C/A-code pseudorange. In general, more precise measurements of pseudorange can be derived from shorter wavelengths of PRN code. The C/A-code range measurement precision (noise) is much lower than that of the P-code by a factor of ten. The pseudorange method of position fixing is a function of the unknown GPS receiver coordinates, the clock errors between satellite and GPS receiver, and the signal propagation errors. The pseudorange measurement is used in GPS as a standard measurement method.

The carrier phase ranging method uses the difference between the phase of the reference signal generated by the GPS receiver carrier and that of the incoming carrier signal from a satellite. The carrier phase has two parts: fractional phase and integer ambiguity. Integer ambiguity refers to the number of full phase cycles between the GPS receiver and the satellite. It remains constant throughout a period of continuous signal tracking. At the start of the signal acquisition, the GPS receiver can measure only a fractional phase and an initial unknown integer ambiguity. Special ambiguity resolution algorithms are then used to resolve the integer ambiguity. After the initial period of signal acquisition, the receiver can count the number of integer cycles that are being tracked. The carrier phase measurement can, in this way, be expressed as the integer number of cycles counted since the start of signal tracking and the sum of the fractional part measured to millimetre accuracy. This is why the longer the period of tracking, the more accurate the position can be. These measurements can then be converted to a range by multiplying the measured phase with the corresponding wavelength. The carrier phase ranging method can achieve high accuracy positioning in GPS, however, there is a need to use more sophisticated and higher cost GPS receivers.

Another GPS measurement used primarily in kinematics applications for velocity estimation is instantaneous Doppler frequency. This is achieved by determining changes in the phase range due to a change in the GPS receiver's position.

#### **6.3.2.4 GPS Positioning Accuracy**

In Section 6.3.1, a range of error sources that influence the positioning accuracy of GPS were discussed in general. The positioning accuracy can also rely on the number of satellites used for positioning, the type of measurement methods employed (pseudorange or carrier phase

## Location-Based Services and Geo-Information Engineering

**Table 6.1** GPS positioning accuracy (based on Rizos, 2002).

Point positioning		Relative positioning (see Section 6.3.3)	
Standard Positioning Service (pseudorange)	SA off	3–10 m	Kinematic survey (carrier phase) $\geq 5$ mm
	SA on	10–100 m	DGPS services (pseudorange) $\geq 50$ cm
Precise Positioning Service (pseudorange)		1–5 m	Static survey (carrier phase) $\geq 2$ mm plus 1 ppm (up to $< 0.1$ ppm)

Note: 95% confidence level.

ranging, Section 6.3.2.3) and the data reduction algorithm used. Furthermore, the accuracy varies depending on whether users are stationary or moving (static vs kinematic mode) when positioning takes place, and whether the positioning is performed in real time or after post-processing. More details can be found in Grejner-Brzezinska (2004) and Küpper (2005).

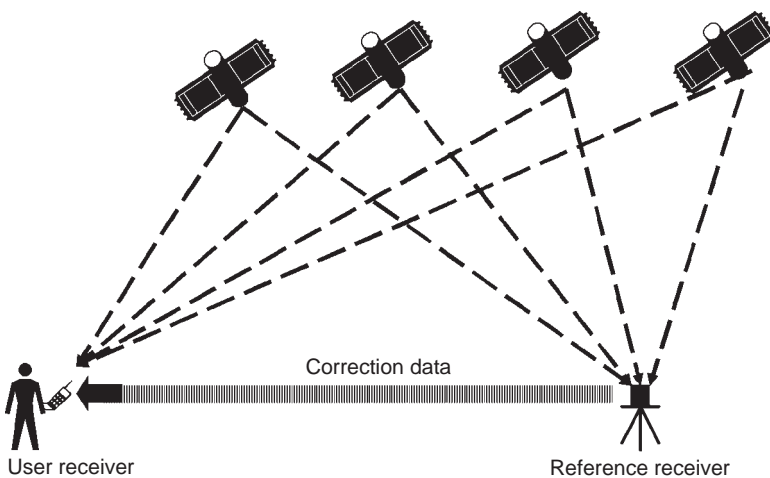
GPS provides two basic types of positioning services as discussed in Section 6.3.2.2: the Standard Positioning Service (SPS) and the Precise Positioning Service (PPS). SPS is generally for civilian usage, whilst PPS is mostly for military and other authorized users. SPS provides positioning and timing services based on the C/A-code on the L1 carrier, and PPS provides positioning, velocity and timing services based on the C/A-code and P-code on both L1 and L2 carriers. The positioning accuracy of GPS is shown in Table 6.1 with different positioning services and measurement methods.

### 6.3.3 Principle of Differential GPS

To achieve more accurate position measurement than standard GPS, a relative (differential) positioning technique has been developed known as Differential GPS (DGPS). In point (absolute) positioning, the position is determined through a single receiver and multiple satellites. The principle was explained in Section 6.3.1. The configuration of the satellites and any remaining errors will directly affect the accuracy of the point positioning. The basic principle of DGPS is to remove these errors by determining the difference between the measured distance at an unknown point whose position is to be fixed and an approximation of the true distance as calculated by a receiver at a known position.

Thus DGPS uses at least two receivers for positioning. One GPS receiver acts as a reference receiver (also known as a reference station), which is placed at a known stationary position with accurately surveyed coordinates  $(x, y, z)$ . The other GPS receiver, known as the user receiver, is at an unknown position that can be mobile. Both the reference and the user receiver track the same satellites simultaneously. The difference between the two simultaneous measurements by both receivers is calculated. Thus, the systematic error sources, which are common to the reference receiver and the user receiver, can be determined. DGPS positioning is an effective way of increasing positioning accuracy by removing the systematic errors. The basic concept of DGPS is illustrated in Figure 6.2. It is important that both the reference and user receivers track the same satellites to generate their measurements, otherwise severe errors may result.

The GPS error sources are spatially and temporally correlated between a reference and user receivers for short to medium separation distances. Therefore, the systematic errors in the measurements can be estimated at specified time intervals. These systematic errors can then be made available as differential corrections and sent to GPS receivers being used nearby by wireless communication. These corrections, thereafter, can be used by these receivers to remove the systematic errors from the measurements collected at all unknown positions. The correction can also be applied to all other receivers in the same locale,



**Figure 6.2** The basic principle of differential GPS.

to eliminate error in their measurements. Over distances of 60–70 km the systematic errors in GPS measurements due to the troposphere, satellite clock and so on are likely to be very similar and therefore differential corrections to significantly reduce systematic errors are relatively safe over such distances.

The correction data used for removing the systematic errors, transmitted from a reference receiver to user receivers, can be implemented in two ways. One is using the block shift method by which the coordinate solution is transmitted from the reference receiver to user receivers. The process is as follows: obtain the known position of a reference receiver; compare the known position with the position determined via GPS for the reference receiver; determine the corrections on  $x$ ,  $y$ ,  $z$ ; transmit the corrections to user receivers; apply the corrections to 'raw' field coordinates, or save to a file for later correction through post-processing. Although the block shift method is easy to implement, it is less flexible. Another method is range correction. It works by correcting the range instead of correcting computed coordinates. The process is as follows (many parts are similar to the block shift method): obtain the known position of a reference receiver; determine the 'true' range using the known position from the reference receiver; compare the 'true' value with the 'observed' range value obtained from the GPS reference receiver; generate the correction data for all pseudorange data; transmit the correction data on ranges to user receivers for positioning. This method is more flexible than the block shift method. The range correction is based on the pseudorange; therefore, the user GPS receivers can determine a position using any combination of corrected ranges.

Positioning with DGPS is implemented by DGPS services. DGPS services can enable users to obtain higher positioning accuracy. They are usually provided by governments, industries and professional organizations. However, in order to use such services, additional hardware is required to be able to receive and process the differential corrections. DGPS services normally involve some type of wireless transmission system. In general, VHF or UHF systems are used for short range transmission whilst low-frequency transmitters are used for medium range transmission (beacons). For the coverage of entire continents, L-band or C-band geostationary satellites are used known as Wide Area DGPS (WADGPS). WADGPS involves multiple GPS reference base stations that generate corrections and send them to a master control station. The master control station then uploads the corrections to a geostationary communication satellite that in turn

transmits the corrections to the user receivers. The positioning accuracy of WADGPS, such as OmniSTAR, is at the sub-meter level. Some examples of DGPS services are:

- Wide Area Augmentation System (WAAS), which is operated in United States and supported by the Federal Aviation Administration (FAA). Its major objective is to support aviation navigation and precision approaches. The accuracy is expected at about  $\pm 7.6$  m at 95% of the time. It is a public service and any appropriate GPS receiver may have access to it.
- LADGPS, which is a ground-based Local Area DGPS and supports real-time positioning, usually over distances of up to a few hundred kilometres.
- LAAS, which is a FAA-supported Local Area Augmentation System with the main objective of supporting aviation navigation and precision approaches.
- European Geostationary Navigation Overlay Service (EGNOS). Available in Europe, a geostationary satellite has to be visible to use this service. This requirement is often not met because the satellite is not high enough in the southern sky in many urban environments in Northern Europe. Although the EGNOS correction data can be transferred through an Internet link, there is then a need for a continuous reliable Internet connection to receivers.
- The 'National GPS Network', which provides differential GPS information in the United Kingdom from fixed sites for post-processing (see <http://www.gps.gov.uk> for a range of information on this).
- Another approach is establishing a local network of Continuously Operating Reference Station (CORS), especially those requiring the highest accuracy. The US National Geodetic Survey, US Department of Transportation and International GPS Service (IGS) deploy and operate these networks. Normally, all users have a free access to the archived data.

### 6.3.4 Indoor GPS

The emphasis of indoor GPS is on location-sensing systems for indoor environments by applying the advantages of GPS. As discussed in

Section 6.3.1, positions can only be determined through GPS when there are signals received from three or four GPS satellites – and this usually requires an unimpeded view of the sky. GPS cannot work inside buildings because the signal strength is not high enough to penetrate a building. One of the indoor GPS methods is to use pseudolites (pseudo-satellites). These are transmitters that generate a GPS-like signal. The signal is designed to be very similar to GPS satellite signals so that a minimal modification to existing GPS receivers is necessary to allow them to operate as pseudolite-compatible receivers. (Giaglis *et al.*, 2002; Zeimpekis *et al.*, 2003). As with true GPS, at least four pseudolites are needed to be visible for positioning. Indoor GPS can be used in positioning mobile devices with low power consumption. Another solution for using GPS positioning indoors is to increase the sensitivity of the GPS receivers. This can be achieved by installing a large number of parallel working correlators to the GPS receivers. Conventional receivers usually have two to four correlators and are not sensitive enough to receive the GPS signal in indoor environments. Increasing the number of correlators improves the receiver's ability to detect signals.

### 6.3.5 Other Satellite Systems

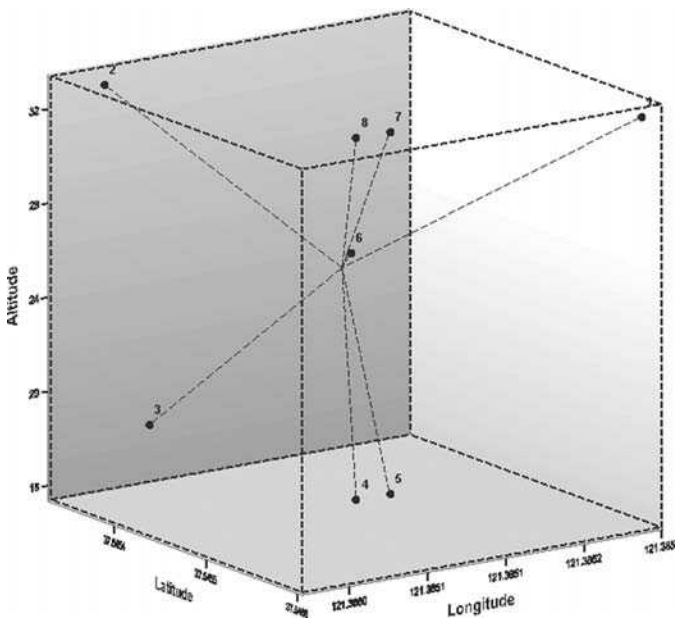
Apart from GPS, there are a number of other GNSS (Global Navigation Satellite Systems) based on similar principles and with similar basic features. One of them is the Russian GLONASS system, which was originally developed for the Russian military and became operational in 1996. However, its long term stability has been questioned given the current situation. Another system is the European Galileo system (civilian), which is managed and financed by the European Commission and the European Space Agency. Galileo is expected to become fully operational sometime after 2009. There is also a Chinese system Beidou, also known as Compass Navigation Satellite System. Three trial Beidou navigation satellites have been launched since 2000, with the objective of becoming a fully operational GNSS.

### 6.3.6 An Example of GPS Use

GPS technology can be used in a wide range of applications, including aircraft, ship navigation, survey, mapping and engineering, natural

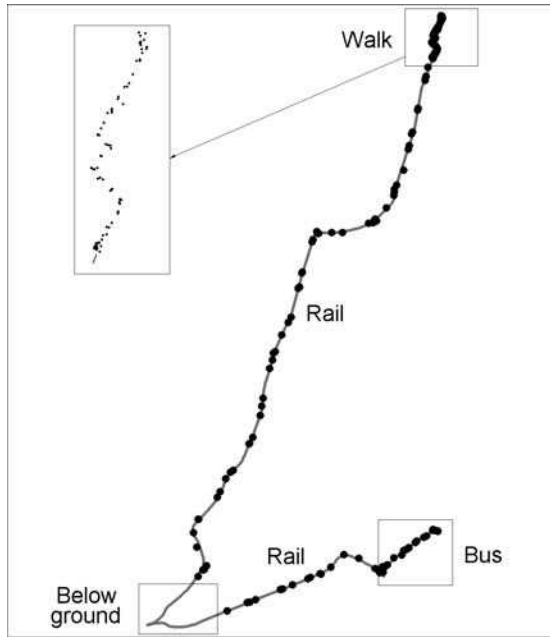
resource management, emergency response, transport and fleet management, in-car navigation and routing. For LBS applications, GPS positioning has the advantages of a global coverage and high accuracy compared to current mobile network-based positioning (see Section 6.4). However, the high TTFF and the high battery consumption of receivers are obvious disadvantages for its use in LBS. Another drawback is that positioning can't be carried out amongst dense high buildings and in indoor environments. On the other hand, GPS positioning can be integrated with mobile network technology to provide hybrid positioning approaches and enhance the positioning accuracy and consistency of GPS-enabled mobile devices (Section 6.6). Before moving on to network-based positioning technologies, it will be useful to examine some live GPS data so as to illustrate some of the issues just discussed.

Whilst TTFF is one factor in GPS performance, a GPS receiver needs to locate all the satellites within view and then to compute which ones provide the best fix. During this start-up period the GPS fixes can 'wander' considerably and can be highly inaccurate. This is illustrated in Figure 6.3 for a stationary GPS receiver recording once



**Figure 6.3** Wandering of a GPS position fix for a stationary receiver during the start-up period.





**Figure 6.4** A GPS track for a multi-modal transport scenario.

a minute from start-up. The first eight readings are shown in relation to the average point: the 2-D position appears to settle down from the fourth reading onwards, but the 3-D position only settles down from the sixth reading. The second example (Figure 6.4) is a track of an individual who begins by walking to a station, then takes a train and finally completes the journey by bus. The inset shows the walked section. The GPS data do not portray a neat line of fixes following the roads walked but have a scatter as successive readings have different levels of accuracy. This appears contrary to the impression of high accuracy given by SatNavs, where the location of the user remains nicely on the road network. The reality is that the fixes from a SatNav (or GPS-enabled mobile phone) do have a scatter as in Figure 6.4 but that the internal algorithm snaps the fix to the nearest road. The position fixes during the train ride are more spaced apart due to the increased speed of travel. However, this section of the journey also shows an erratic track with gaps in the fixes due to insufficient view of the satellite constellation from time to time – most noticeably in a tunnelled section below ground where the user also changed train line. Whilst position fixing using GPS can be very accurate (to within

a few metres) compared with the other techniques discussed below, the tracks can be quite messy at a detailed level with some missing data.

## 6.4 Network-Based Positioning Technologies

---

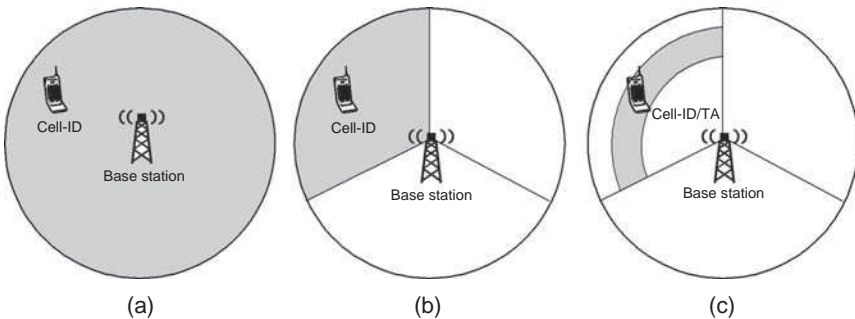
The development of network-based positioning technologies was initially driven by the regulatory Enhanced 911 (E911) mandatory from the FCC (Telecommunication Act 1996), and also later in Europe by E112 (Section 4.4). In principle, network-based positioning technologies use a network of base stations to locate a mobile device by measuring the signal travelling between the mobile device and a set of base stations. The direction and/or length of individual radio paths can be computed through signal measurements, and the position of a mobile device can then be determined from geometric relationships. In this section, the principles of some commonly used network-based positioning technologies are described. These technologies include Network Cell Identification (Cell-ID) technology (Section 6.4.1), Angle of Arrival (AOA) technology (Section 6.4.2) and time delay methods, including Time of Arrival (TOA), Time Difference of Arrival (TDOA) and Enhanced-Observed Time Difference method (E-OTD) (Section 6.4.3). One of the main advantages in locating the position of a user's mobile device for LBS through the use of these network-based positioning technologies is that they are based on the telecommunication networks servicing mobile devices with voice and data and do not require any extra software or hardware to be installed in mobile devices. The term mobile device generally refers here to a handheld, mobile wireless-enabled device such as a mobile phone or PDA that is wireless enabled. A mobile device is also known in some texts as a terminal or a mobile station.

### 6.4.1 Network Cell Identification

The Network Cell-ID method (Cell-ID) is also known as Cell of Origin (COO) or Cell Global Identity (CGI). It is a network-based proximity positioning method. The basics of the cell were discussed in Section 2.3.1. The Cell-ID method is a basic technique to locate mobile devices, often used in GSM mobile networks. The principle of the

Cell-ID method is to use the centre of a cell's coverage and its cell size to determine the position of a device within that cell coverage. The method identifies the approximate position of a mobile device through locating which cell base station the device is using at the given time. The position determined by the network is not necessarily the co-ordinates of the geographical centre of a cell, but the mast location and the size of a cell. A base station may have just one omni-directional antenna located in the centre of a cell (Figure 6.5a). More commonly in urban built-up areas, a base station has three directional antennas. Therefore, three cells are covered by one base station (Figure 6.5b). The three cell areas served by a base station are referred to as cell sectors. The size and shape of a cell can vary considerably from a rural to an urban area according to the density of mobile users that need to be served.

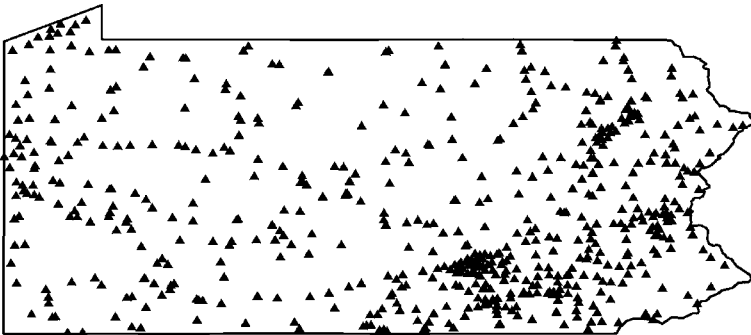
The location accuracy of the Cell-ID method varies subject to the size of the cell and the density of mobile device users. This density is used to determine the statistically most likely position of a user. It can also depend on the shape of a cell, which can vary considerably due to the geographic features in the cell, such as hills and large tall buildings. In urban areas, the accuracy can be in the range of 50–500 m (typically 200 m) for micro-cells, and 500 m to 5 km (typical 2 km) for macro-cells. But in rural areas, the accuracy can be in the range of 1–35 km due to larger cell sizes. When using indoor pico-cells, the accuracy can be in the range of 10–50 m. For some special extended cells, such as for near-shore and coastal users, the maximum radius can be as large as 120 kilometres. Even though the positioning accuracy using Cell-ID



**Figure 6.5** The basic principle of Cell-ID and Enhanced Cell-ID methods: (a) single cell from an omni-directional antenna; (b) cell sectors from three directional antennas; (c) Enhanced Cell-ID with timing advance (TA).

is relatively low and varies depending on cell size, there is no specific hardware and software support required for devices using the Cell-ID method. There is also no specific hardware requirement for networks, but additional software support is needed for networks to determine the centre of cell coverage and cell size. Cell-ID is the most commonly available technology among mobile network operators. Data on masts can be found at <http://www.sitefinder.ofcom.org.uk/> for the United Kingdom and at <http://wireless.fcc.gov/uls/index.htm?job=transaction&page=weekly> for the United States. Figure 6.6 shows mast configuration for Pennsylvania, United States. Note how there are different densities, the clusters being urban areas. The near linear arrangement of masts in some areas reflects major highways and the towns along them.

The basic Cell-ID method can be combined with the timing advance value for a mobile device to improve its positioning accuracy. This method is called Enhanced Cell-ID, also referred to as Cell-ID/TA. Mobile devices and base stations transmit based on a strict timing schedule that allows multiple mobile devices (mobile phones) to use a single frequency. To ensure that the transmission from a mobile device arrives at the base station exactly in its allocated slot regardless of its distance from the base station (i.e. to compensate for the time it takes for the transmission to occur), the base station instructs each mobile device to advance its timing by an amount that is a function of the distance from the base station. Based on this principle, the timing advanced value in Cell-ID/TA can be represented as a doughnut-shaped area that is the approximate location assuming the base station is at the centre of cell. If the base station uses directional



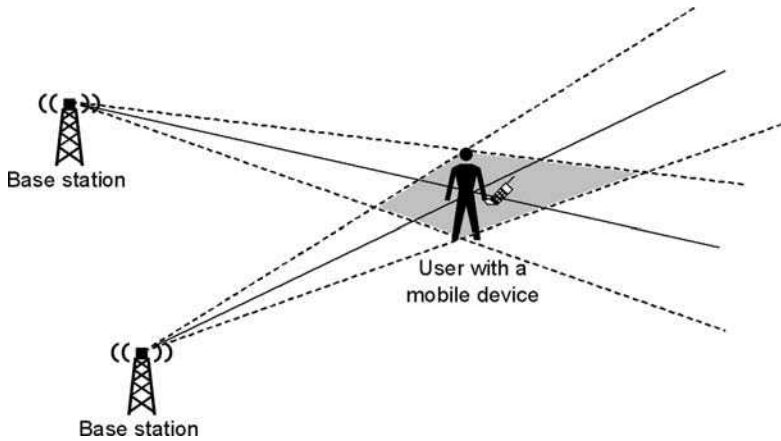
**Figure 6.6** Configuration of cells and cell sizes for Pennsylvania, USA (from FCC data).

antennas, the timing advanced value can be represented as a section of a doughnut-shaped ring which is the approximate device location. The principle is illustrated in Figure 6.6c. The time advance value could be one of 64 values, with each of them being a 550 metre wide band. Therefore, the use of timing advance value with Cell-ID method enhances location accuracy compared to using the basic Cell-ID method. However, the degree of enhanced accuracy using Cell-ID/TA method still varies. In urban areas, where signals are more likely to be bouncing off buildings, multi-path propagation often occurs resulting in the signal path being longer than the direct path. Therefore, the location of a mobile device indicated through timing advance may be much further than its real location. In these cases, the accuracy might not be improved by using Cell-ID/TA. For rural areas, where multi-path propagation is less likely to occur, the location accuracy can be considerably improved by using timing advance. This method also requires no additional hardware support for either network or mobile devices in order to be used.

Another enhanced Cell-ID method estimates the distance between a mobile device and its serving base station by measuring the strength of the signal from the serving base station compared with other base stations for neighbour cells. The measured signal strength level at the device is used to estimate the distance by applying propagation models and network planning tools. It can provide more accurate location than the basic Cell-ID method. The signal strength enhanced Cell-ID is likely to have improved positioning accuracy in rural areas than in urban areas.

### 6.4.2 Angle of Arrival

Angle of Arrival (AOA) is another technology used for network-based positioning. The location of a mobile device is determined by measuring the angle of the signal received from two base stations. The basic principle is illustrated in Figure 6.7. Base stations are equipped with directional antenna arrays (Figure 2.5b) that can measure the broad directionality of signals from mobile devices according to signal strength. Thus the incoming signal from a mobile device is measured at a base station to establish a straight-line locus from the base station to the device; the location of the mobile device is along this line within the cell. Another AOA measurement at another nearby base station from the same mobile device will determine a second straight-line



**Figure 6.7** Basic principle of the AOA positioning method.

locus. The intersection of these two lines gives the position fix for the device in two dimensions. The observed angles will be an approximation of actual angle values because of a sub-optimal resolution of antenna arrays. In theory, two base stations are required to position a device, in practice three or more angle measurements are likely to improve the accuracy.

In general, positioning with AOA is able to reach accuracies of within 300 m. However, the accuracy can be reduced significantly in rural areas. This is because the measurement is an angular locus and so the accuracy of location fix decreases the greater the distance a mobile device is from a base station (as would be expected in rural areas). However, multi-path propagation can affect AOA efficiency in urban areas when there is no straight line of sight between base stations and a device. If three or four base stations are used in AOA measurement, positioning accuracy can be improved. To facilitate the positioning using AOA method, considerable hardware support is required in the configuration of networks. Base stations need to be equipped with an array of antennas, rather than one antenna per cell, in order to measure the angles required by the AOA method.

### 6.4.3 Time Delay Methods

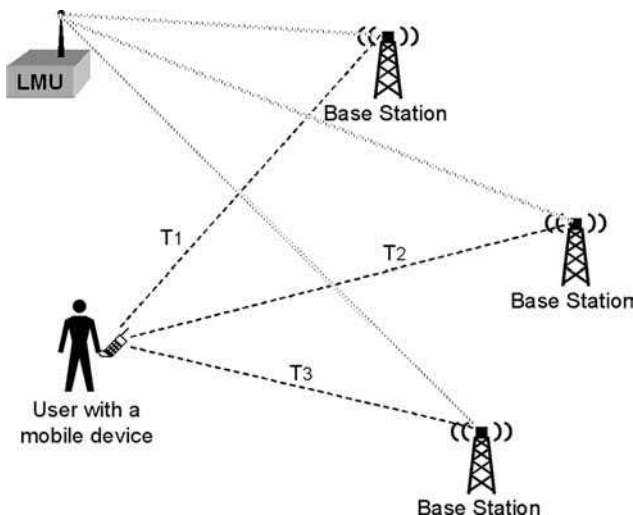
Time delay methods are network-based positioning technologies which-use the time delay principle of a signal transmitted between a

transmitter and a receiver. Because the speed of electromagnetic waves is constant in ‘free space’, distances can be measured by calculating the time delay of a radio wave transmitted between the two points. Time of Arrival (TOA), Time Difference Of Arrival (TDOA) and Enhanced-Observed Time Difference (E-OTD) can all be regarded as time delay methods.

### 6.4.3.1 Time of Arrival – TOA

TOA, also known as Absolute Time of Arrival, is the method by which the position of a mobile device is determined by the time delay (travel time) for signals between a number of base stations in a network and a user’s mobile device. The range between a transmitter and a mobile device can then be calculated by the travel time of signals between them as illustrated in Figure 6.8. A high degree of synchronization within a network of base stations is required for the TOA method. Furthermore, the position of each base station in a network also needs to be known to a sub-metre accuracy. The TOA method can provide positioning accuracy within a range of 125–200 m, which can be higher within areas of low multi-path effects.

An alternative approach to TOA involves the measurement of the round-trip of a signal transmitted from a base station and then



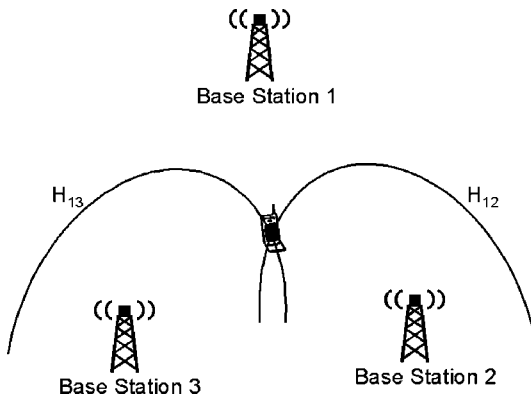
**Figure 6.8** Basic principle of TOA positioning method (LMU is required for GSM networks – see text for explanation).



echoed back by the mobile device, giving a result that is twice the one-way measurement. This approach does not require time synchronization between the base station and the mobile device and is the most common means of measuring TOA. Within a GSM network (a 2G network; Section 2.3.1), base stations are not synchronized with each other. To overcome this problem, additional network elements are required to transmit a beacon signal from a fixed known location. These beacon transmitters are called Location Measurement Units (LMUs) and effectively allow base stations to be synchronized with one another (also shown in Figure 6.8). Thus to position mobile devices within GSM networks using the TOA method, LMUs need to be incorporated into the network. This requires considerable network investment, although existing mobile devices can be used with little or no modification to software. For 3G networks, LMUs are not required because the base stations are already synchronized.

#### 6.4.3.2 Time Difference of Arrival – TDOA

The TDOA method calculates the time difference of a transmitted signal from a mobile device arriving at three nearby base stations in a network. The basic principle of the TDOA method is illustrated in Figure 6.9. The time difference of signal arrival allows the relative distance of a mobile device from each base station to be determined. To position a mobile device at least three measurements are required. Each of the measurements defines a hyperbolic locus on which the



**Figure 6.9** Basic principle of TDOA positioning method ( $H_{12}$  is the hyperbolic locus from base stations 1 and 2,  $H_{13}$  is the hyperbolic locus from base stations 1 and 3).

mobile device must lie (Zeimpekis *et al.*, 2003) and it is the intersection of such loci that allows a position fix on the mobile device. For the TDOA method, transmitters and receivers need to be time synchronized. Because the difficulty of having both the clocks at transmitters and receivers precisely synchronized, the TDOA method synchronizes several transmitters to a common time base and then measures the time difference of arrival at the receiver. A mobile device normally transmits signals to more than one base station. Although a mobile device will be registered with a base station and assigned a Cell-ID, it will still continue to transmit signals to other nearby base stations in order for a hand-over from one cell to another to work efficiently. Thus as long as a mobile device is within range of and maintains contact with three base stations then TDOA can be used.

In theory, the positioning accuracy of the TDOA method is between 50 and 200 m, the usual reported accuracy being about 125 m. There is no specific hardware and software required for mobile devices in using the TDOA method; however, for GSM networks there is a need for high synchronization within the network of base stations to support TDOA. Therefore, there is significant network investment required, such as additional hardware, for this method to be used in a GSM network. Similar to the TOA method, LMUs are required for network base stations to use the TDOA method within a GSM network. In 3G networks the base stations are synchronized thus removing the need for LMUs. Furthermore, because the cells in a 3G network are smaller, there is an increased likelihood of a mobile device being within range of three base stations. This can increase the achievable accuracy to around 20 metres.

### **6.4.3.3 Enhanced-Observed Time Difference – E-OTD**

E-OTD technology is a network-based positioning method. It is a modification of the TOA and TDOA methods described above in that the position is calculated by the mobile device rather than by the network software. There are thus two configurations for E-OTD: one using time of arrival for signals sent to base stations, the other using time difference of arrival. Both types carry out timing measurement at the mobile device, the position is then reported back to the network. The E-OTD method requires the installation of specific software within the mobile device. However, E-OTD requires less dependency on network synchronization of base stations as the signals to the three base stations are all initiated by the mobile device. The positioning accuracy using

E-OTD is generally better than 150 m, and can be as high as 50 metres. E-OTD can also work with GPS to provide a hybrid positioning solution as an augmentation to satellite-derived positioning. For more details of E-OTD positioning see, for example, Küpper (2005).

#### **6.4.4 Advanced Forward Link Trilateration and Enhanced Forward Link Trilateration**

Both Advanced Forward Link Trilateration (A-FLT) and Enhanced Forward Link Trilateration (E-FLT) are network-based positioning methods, unique to CDMA-based mobile networks, because they are inherently synchronous in their operation. The A-FLT method uses measurements of the phase delay between signals sent to two different base stations; this is then compared to the same data taken from another pair of base stations. A-FLT works in a similar way to the TDOA method by which the position of a mobile device is derived from several base stations. Software changes are required for mobile devices to use the A-FLT method, and there is the need to have LMUs for base stations in some 2G mobile networks. For 3G networks (i.e. Cdma2000), LMUs are not necessary because the base stations in these networks operate synchronously. The A-FLT method has accuracy ranges of 50–200 m. E-FLT positioning method is also mainly based on TDOA technology using forward-link signals received by mobile devices. E-FLT can be used for legacy CDMA mobile phones with no changes needed to handsets. The E-FLT method has a positioning accuracy of about 250–350 m.

### **6.5 Short Range Positioning Technologies**

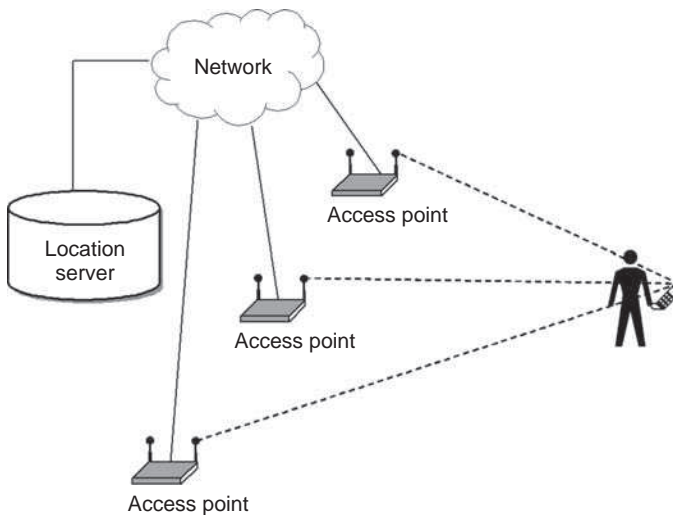
---

The positioning technologies discussed above can be used for large area coverage: GPS positioning has global coverage and network-based positioning methods cover the operational area of whole mobile telecommunications networks. There is, however, a range of technologies that can be used to position devices over comparatively short ranges. They are frequently employed for positioning in indoor environments. These technologies include Wireless LANs (WiFi), Bluetooth, Radio Frequency Identification (RFID), Ultra Wide Band positioning, as well as ultrasonic, infrared, camera-assisted and

sensor-assisted technologies. Here the focus is only on a selection of these positioning technologies, including WiFi, Bluetooth and RFID. They tend to be mainly used to identify a proximity location where mobile devices connect to a network.

### 6.5.1 WiFi-Based Positioning

WiFi is a wireless networking standard defined in the IEEE 802.11 series. WiFi are also known as Wireless Local Area Networks (WLANs). WiFi-enabled devices (often equipped with a WiFi adapter) can be connected to a wireless network via access points on the network with a speed of up to 54 Mbps. These local networks are often connected to a larger network such as a corporate network. The details of WiFi can be found in Section 2.3.2. The basic principle of using WiFi technology to position a mobile device (referred to as WiFi-based positioning) is illustrated in Figure 6.10. The location of a device is determined by measuring the strength of signals received at two or more access points. The signals communicated between mobile devices and access points are often referred to as beacons, which contain packets of information. There are two ways of transmitting these signals for positioning: uplink and downlink. For uplink, a WiFi-enabled mobile device generates beacons. These beacons are received



**Figure 6.10** Illustration of WiFi-based positioning.

by those access points that are in range, that are able to receive the signal, and the network establishes the position. For downlink, access points transmit beacons containing their ID along with other information. The device continuously receives beacons from those access points in the nearby vicinity, detects the best quality signal for transmission and identifies the access point for positioning.

The measurement of beacons for WiFi-based positioning can be carried out in three ways. The first is to determine the position of a device from the position of access points according to the quality of the signals received. This is the simplest method for implementing WiFi-based positioning. The second is to compute the position of a device from the distance between the device and a number of access points. Such distances can be calculated by the path loss of a signal during the transmission between the device and access points. The accuracy can be improved using this method of WiFi-based positioning; however, this can depend on the configuration of access points relative to the mobile device. The third approach uses ‘fingerprinting’, whereby the position of a device is derived by matching the patterns of signal strength from access points against a table of pre-collected patterns at various positions across the area of network coverage. This latter approach can achieve much higher accuracy than the other more conventional approaches. Measurements of beacons can be sent to a dedicated location server for processing. This server can sit anywhere on the network and reduces power consumption of mobile devices in achieving a position fix. In general, using measurements of signal timing is not suitable for WiFi-based positioning because it is very difficult to measure the time differences of signals travelled over such a short range.

WiFi-based positioning only works in areas where there is network coverage from access points. This method can work in indoor environments. It is not suitable for a very large scale implementation; therefore, the coverage is relatively small although it may be larger than other indoor positioning technologies such as Bluetooth (see Section 6.5.2). Currently, WiFi technology is widely used in offices, homes, airports, cafés, hotels and public places. WiFi-based positioning can also be used for outdoor environments, usually in dense urban areas. The accuracy of this method is generally not very high, and also depends on the density of access points and the strength of signals transmitted. In addition, beacon signals emitted from access points can contain some specific location information about the access point (e.g. room number in a building). In these

cases, the level of positioning accuracy can be increased. WiFi-based positioning can have a relatively higher accuracy compared to other techniques when used for tracking in a closed and defined indoor environment. 'Fingerprinting' data are generally needed for this type of application. In general, WiFi-based positioning provides a proximity location rather than a more specific  $x,y$  position. For example, it may return which terminal or part of a terminal at an airport a mobile device is located. Using WiFi-based positioning, there is no requirement for additional infrastructure except for the existing WiFi network (WLANs). However, devices have to be WiFi enabled; and information on WiFi access points (a WiFi database) is often required. For those LBS applications for indoor environments or some open urban spaces with WiFi access, such positioning technologies can be readily used to locate users.

### 6.5.2 Bluetooth Technology Used for Positioning

Bluetooth technology provides an ad hoc approach for low power, short range wireless connections for voice and data transmission between various devices within a nominal 10 m range. It operates in the globally available unlicensed 2.4 GHz ISM radio frequency band. Bluetooth-enabled devices can be linked up for ad hoc networking with other Bluetooth-enabled devices when they come within range (Section 2.3.2). Most mobile devices are currently fitted with Bluetooth-enabled transmitter chips.

Although there is no Bluetooth specification designed to support positioning services, Bluetooth technology can be used to locate user mobile devices. Various solutions have been developed based on Bluetooth technology for short range and indoor positioning. Bluetooth-enabled devices can transmit signals containing information such as device identity and profile. Such signals can be picked up by a host device and used to identify the presence of other devices when they are within the 10 m communication range. Signals can also be processed and computed to determine the position of a device. Position information should be able to be exchanged between Bluetooth devices locally or with the location server in a network. In addition, the tag and reader positioning approach discussed in Section 6.5.3 can also be used in Bluetooth-based positioning. The time required for scanning signals between Bluetooth-enabled devices is a minimum of 1.28 seconds, and maximum time for retrieving a position is 15.4 seconds. The presence of

other devices nearby could increase the time needed for positioning the intended device.

The received signal strength from a device decreases logarithmically with distance both in indoor and outdoor environments. Based on this relationship, Bluetooth signal strength information can be used to position and track the users/devices enabled with Bluetooth. One positioning approach is to determine the position of a device by triangulating a set of signal strength readings gathered through antenna located at different positions (Thapa and Case, 2003). The approximate position of a device will be the intersection of three circles generated by known signal levels from three different antennas. The accuracy of using this approach will be higher than just detecting whether a device is in range. However, there is a need for at least three access points within range. The accuracy of Bluetooth-based positioning is relatively limited, and determined by the maximum range covered. The typical range of accuracy can be from 10 to 100 m depending on the power class of device. However, since the most commonly used nominal range is 10 m, this is likely to be the nominal accuracy.

The main advantage of Bluetooth-based positioning is that it can be deployed rapidly, with easy maintenance and low cost. Positioning with Bluetooth technology can be used by those LBS applications where short range coverage with approximate area accuracy is sufficient. For example, this might be short range 'push' information services and alerts. Another advantage of Bluetooth-based positioning is that it is designed to consume low levels of power and is suitable for small mobile devices. Moreover, the internal authentication and encryption provided by Bluetooth technology provides for better security than other wireless technologies, which rely on external means of security.

### 6.5.3 Radio Frequency Identification

Radio Frequency Identification (RFID) technology is often used over a very short range (typically 1–3 m) with low power and low cost. The radio frequency used by RFID systems is at different ranges from 100 to 500 KHz (low frequency), 10–15 MHz (intermediate frequency), through to 850–950 MHz and 2.4–5 GHz (high frequency). RFID reader and RFID tag are two basic components in a RFID system. Both the reader and the tag each have an antenna and a transceiver for



two-way communication between them (i.e. exchanging radio signals). The RFID reader is powered, and has a processor and an interface. It can be connected to a server. The RFID tag, known also as RFID transponder, usually has a microchip attached to its antenna. The RFID tag can be either passive or active. Passive tags have no power supply themselves, and only obtain their power through the radio signals transmitted by RFID readers. Passive tags normally have a small size memory that can only store limited information such as an ID. Such passive tags also have a very short communication range – generally within a few metres only. Active tags have their own power (i.e. equipped with a battery), with more memory for storing extra data and capable of processing some calculations. The communication range of an active tag can reach to tens of metres. RFID technologies have been widely used in applications such as identifying objects, controlling stock and managing access.

RFID technology can be used to provide proximity position data of users. RFID tags can be identified by RFID readers when they are within communication range. The tags then respond to the readers and transmit desired information stored in their memory (in a microchip). Thus, the proximate position of RFID tags to a reader can be identified. There are several methods of identification. One is to store a serial number that identifies a user/device or an object on a microchip within RFID tags. RFID readers are connected to a location server via a network. In this method, RFID readers are generally fixed at specific physical locations such as building entrances and exits, desirable points of a public space or on any walls of a building. RFID tags are equipped within user devices or objects. When the user/device with a RFID tag moves into the communication range of a RFID reader, the reader prompts the tag and the antenna in the tag enables its chip to transmit the identification information to the reader. The reader then sends the information to the location server. In this way, the proximate position of a user/device can be obtained. Another method is to locate RFID readers within mobile devices. When a RFID-enabled device moves into the communication range of a RFID tag located at a fix position, the position data of the tag can be transferred into the reader. Thus the proximate location of an RFID-enabled device can be determined. Using RFID technology for positioning doesn't require line of sight. RFID tags can be read by a RFID reader as long as they are within the communication range. The accuracy of this positioning approach depends on the communication range of RFID radio signals.

### 6.5.4 Other Non-Radio Signal Technologies Used for Short Range Positioning

Positioning methods can also use a range of non-radio signal technologies. A few examples are described here.

Infrared is one widely used technology and can be developed for positioning a device over a short range in indoor environments. The infrared technology is based on a direct line of sight, narrow angle communication between an emitter and a receiver working within the infrared frequency spectrum. It is often referred to as Infrared Data Association (IrDA). The signals have a short working range of a few metres (usually 1–2 metres). Infrared signals cannot penetrate walls, but support bi-directional communication. The data transfer rate is in the range of 9600 bps to 4 Mbps. The infrared signals will be reflected and scattered slightly when they reach objects. There are two standards for IrDA: IrDA control and IrDA data. IrDA control is usually used for low speed applications such as controlling keyboards, mouse and joysticks. IrDA data standard is applied for high speed short range applications.

An infrared-based positioning method generally locates a user/device by proximity. Signals used for positioning can contain information such as an identifier, location and time. Receivers are usually located at certain known positions and can pick up signals from a transmitter attached to a user/device when they are in working range. The proximity of the user/device can then be determined. The method can also work the other way round. Using infrared technology for positioning was applied to location-aware applications as early as the 1980s. The Active Badge system is a well known example (Want *et al.*, 1992). In this study, the system determined the location of a user in an office environment and forwarded relevant incoming phone calls to a location near to the user. Each user wore an Active Badge, which was an infrared tag emitting signals periodically containing personal ID. Infrared receivers installed in various locations within the building picked up the signals. The proximate position of the user was then determined and submitted to the main system by the receivers. In general, the infrared-based positioning method has the benefit of low cost and low power consumption. It can also be used in conjunction with other positioning methods. As with Bluetooth and RFID technologies, infrared is suited to particular types of LBS applications where short range proximate location is sufficient.

Ultrasound is another technology that can be used for positioning. The basic principle is to determine position via the time measurement of signals and lateration (multiple distance-based) calculation. Because ultrasound signals have very low propagation velocity compared with radio signals, synchronization is generally not needed for lateration calculation. A user's mobile device needs to be equipped with an ultrasound device that can either be a transmitter or a receiver depending on whether the ultrasound-based positioning method is network-based or device-based. In simple terms, the network-based solution of ultrasound-based positioning uses the ultrasound device as a transmitter. The transmitter emits signals that usually contain unique identity information. Numbers of ultrasound receivers are installed in fixed locations and are able to detect and receive the signal from the user mobile devices when in range. These receivers are usually connected to a server via a network. The signal information can be processed and the device position is then determined. For device-based ultrasound positioning, a mobile user device is equipped with an ultrasound receiver instead of a transmitter. The coverage areas are installed with ultrasound transmitters that emit signals. The receiver inside a user device can receive signals from transmitters within range. The position of the user device is then determined either by computing the distance between the transmitter and the receiver or by applying lateration calculation using signals from at least three transmitters. The ActiveBat system is one example of using ultrasound technology to locate users (Ward *et al.*, 1997). In order to reach a higher accuracy, to within the centimetre range instead of a proximity position, a network of close-meshed ultrasound devices is usually required for the ultrasound-based positioning method. The ultrasound devices in such networks can either be transmitters or be receivers. Ultrasound does not require a line of sight between transmitters and receivers but it has a limited transmission range, which makes the ultrasound-based positioning a method only suitable over short ranges.

Another positioning solution is using Ultra Wide Band (UWB) technology. UWB signals are very short pulses of only a few nanoseconds that are generated simultaneously across all frequencies. This can best be described as a blast of electromagnetic noise. The technology enables fine time measurement due to its high time resolution. It can be employed for TOA/TDOA measurements and lateration to achieve high accuracy. UWB positioning is a practical approach for indoor environments. This method does not require line of sight.

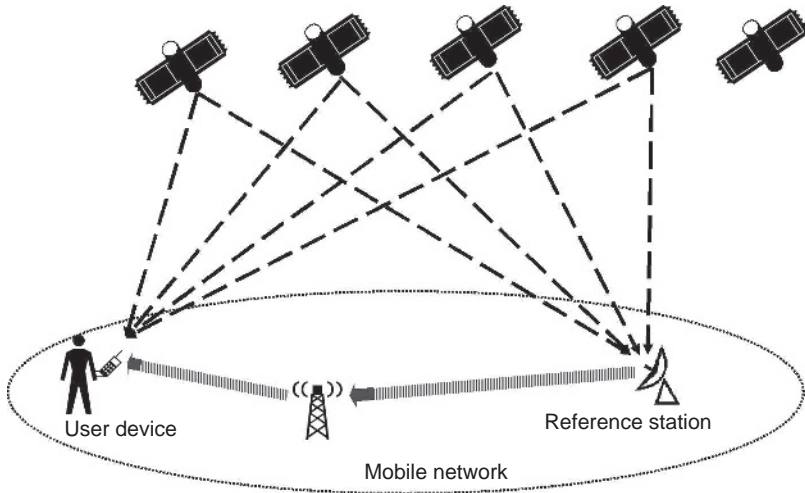
Finally, the ‘dead reckoning’ technique offers yet another positioning solution. It makes use of the initial position of a mobile device and tracks the position from measurements of acceleration, speed and the direction of travel (see inertial navigation systems in Section 5.3.1). For small mobile devices this approach is only approximate, though larger in-vehicle systems can be very accurate. The technique is most useful where GPS signals are weak to supplement the position fix.

## 6.6 Hybrid Positioning Approaches

---

Hybrid positioning approaches aim to take advantage of different types of positioning methods by integrating them. One main hybrid positioning approach is to incorporate device-based and network-based positioning technologies. Device-based positioning, such as GPS positioning, offers high accuracy. However, standard GPS positioning has a relatively long position fix time with high battery consumption and poor signal reception indoors and in dense urban environments. On the other hand, network-based positioning can work indoors and in dense urban areas across a broad range of devices. In general, however, network-based positioning has lower accuracy. The hybrid positioning approach of incorporating these two technologies is to provide increased accuracy and availability. It is based on the integration of GPS technology with mobile network-based positioning technologies. This is commonly referred to as Assisted-GPS (A-GPS). A-GPS positioning can work with most mobile telecommunication networks, such as GSM, UMTS and CDMA2000. With A-GPS, mobile devices (receivers) will have higher sensitivity and less power consumption. The positioning fix time can also be decreased. This section focuses upon A-GPS as an example of a hybrid approach.

The A-GPS positioning method uses both mobile wireless network and GPS technologies to provide improved positioning capability. The main principle is that a mobile network is used to support mobile devices equipped with a GPS receiver (GPS-enabled) to obtain their position with additional information and assistance. The GPS-enabled devices have data from both the GPS satellites and the mobile network. In A-GPS, a number of reference stations can be made available within the mobile network infrastructure to carry out tasks such as computing correction data and compiling GPS satellite signal



**Figure 6.11** An illustration of the principle of A-GPS positioning.

data to assist in position fixing. Reference stations can be used to provide DGPS corrections as a wide-area DGPS network. Reference stations are generally linked to a mobile location centre in the network. As illustrated in Figure 6.11, signals are transmitted between the mobile location centre and reference stations, and between the mobile location centre and the receiver in a mobile device (Küpper, 2005). There will be extra requirements on the network in supporting the signal transmission between reference stations and the mobile location centre.

There are a number of ways in which A-GPS operates. One is that the mobile telecommunications network incorporates GPS receivers to obtain the signals from the GPS satellites and transmits the unprocessed data stream to mobile devices. Where the mobile device is in an area of weak GPS signal reception (e.g. woodland or urban canyons), the GPS receiver within the mobile device can use the data stream from the network to improve the signal-to-noise ratio and achieve a position fix. Another approach involves having some prior knowledge of the rough location of a mobile device before deploying GPS to fix the position. The approximate location of a device can be determined using network-based positioning methods such as Cell-ID, TDOA or AOA. The satellites that should be visible at any particular time to the mobile device can then be identified and the mobile device can be directed by the network to look for those specific satellites

instead of searching exhaustively for any GPS satellites within range. This approach to A-GPS can reduce the time taken to identify and lock-on to the best four satellite signals, thereby reducing the time required to fix a position (Zeimpekis *et al.*, 2003). A further approach to A-GPS can assist in overcoming the limited processing power of a mobile device. This is achieved by having a mobile device stream the data from its GPS receiver to the network where the calculations for a position fix (including DGPS correction) can be carried out on a server. This mitigates limitations to processing power and reduces the battery consumption of the mobile device while positioning.

In general, A-GPS can be classified in two ways: one is device-based, the other is device-assisted. In broad terms, for device-based mode, the position of a device is computed in the device with the support of information provided by signals broadcast via a mobile telecommunications network. The integrated GPS unit inside the mobile device needs to be a fully functional GPS receiver. On the other hand, the device-assisted mode processes the signal data transmitted from a device to the network where the position is then determined. The integrated GPS receiver within the device needs only have basic functionality necessary for collecting the signals, which will reduce the size and cost of the device. In both modes of A-GPS, the information transferred between devices and the mobile network for position measurement includes the reference time, visible satellite list and acquisition assistance data from the GPS navigation message. Device-assisted is most useful for single rapid determination of a position, whilst the device-based approach has advantages for tracking devices over a period.

The A-GPS approach has a number of advantages in positioning. It provides much higher positioning accuracy, which can be as high as 10–15 m. Some claim that A-GPS positioning can achieve the range of 1–10 m (Zeimpekis *et al.*, 2003). A  $\pm 3$  m accuracy can be achieved theoretically in an A-GPS system by employing differential GPS (Bedford, 2004). A-GPS can provide a fast TTFF in positioning and reduce GPS search time from minutes to seconds. Furthermore, it can be used in situations where the GPS signals are too weak for a stand-alone GPS receiver to work on, covering areas such as urban canyons and indoor environments (within limits – and of course providing the network signals can still be received). In general, with such advantages A-GPS positioning can benefit a range of LBS applications. However, using A-GPS requires significant hardware investment both at the network level and at the mobile device level.

## **Location-Based Services and Geo-Information Engineering**

The deployment of different types of positioning technologies for LBS depends on many factors such as: mobile telecommunications network technology, the type of mobile devices being used, operating environment, the requirements for services and applications, and costs. For LBS, it is important to understand that the use of a particular positioning technology should be closely related to the purpose and scope of the specific application.



# Chapter 7

## Context in Location-Based Services

### 7.1 Introduction

---

In Chapter 6 locating the user – a key aspect of LBS – was discussed. Another important aspect is to understand how context is used in LBS. In order to provide tailored data and information services to users in mobile situations, most LBS applications have some level of context awareness. Context-awareness can be used for improving system design, identifying relevant content, enhancing communication and in delivering services. Therefore, understanding context is central to LBS, and there are a number of concepts that need to be addressed. Important context information can include user location and the surrounding environment, user situations during activities, time of day and date, and the technologies involved (devices, networks and systems). Context can be spatio-temporal (Section 8.6), and can also be related to user characteristics, personal preferences and behaviours. The information and services delivered to users in LBS are strongly influenced by and linked to these relevant contexts.

LBS applications are, from a general perspective, user focused and task specific. The main purpose of understanding context in LBS is to enhance the ability of service providers to supply users in mobile situations with information and data services that are viewed as having a high level of utility. It is, therefore, important for application designers to choose what context(s) to use in their applications and

to determine what context-aware behaviour(s) to support in their applications. It is generally viewed that applications provided via a mobile device can support better services for users when the mobile device being serviced can identify more contexts through its usage. Broadly speaking, using mobile devices (and mobile computing in general) can benefit from awareness of context in two ways: firstly, through adaptation of content to changes in the environment; and secondly, through improvement of the interaction with users. Context information provides a means to filter the flow of information from service provider to users, which can help address the problem of information overload and can be used to give additional meaning to the information request made by a user.

Research into context awareness, in general, focuses on location, mobility and time (Markopoulos *et al.*, 2007). Location can be a particular  $x,y$  position, a surrounding environment or a meaningful place such as home or office. Mobility can reflect users in situations where they are continuously on the move or users being in a certain place. Time can add another dimension to context, either as a momentary instant or as a fixed or floating period; time can be absolute or relative. Context can also relate to user ability and preferences, and to the nature of activities being undertaken. Furthermore, technologies employed in LBS contribute context as well. For example, this can include type and functionality of a mobile device, network connectivity and communication media supported. Context has a direct influence on many aspects of LBS, such as communication between users and mobile devices (Chapter 9), content provided for specific applications, system design, service delivery and much more. Therefore, context-awareness is one of the key means of enhancing LBS in order to better meet user requirements. Use of context will inevitably need to be pragmatic, the general challenge being the identification of relevant context(s) in terms of situation, environment and time that can be sufficiently and usefully captured. Both situation and environment are to a large degree characterized by continuity over time. Thus the context history of a user, the time-line, can itself be an important means of mining contextual information for given situations or environments (Schmidt *et al.*, 1999).

Firstly, in this chapter, context and context-awareness are defined from a broader computer science perspective. What this means for LBS is then considered and a view of context as an interaction concept between user, technology and environment is developed. Next, environment as context (location and time), technology as

context and the user as context are looked at individually. In the final section of this chapter the dynamics of context are considered.

## 7.2 Context and Context-Awareness

---

In our everyday life, we communicate with each other within a certain context; our actions are also within a certain context. Events happen and are understood within certain contexts. Context is important to every aspect of human life. The importance of context also influences how we interact with technologies and how technologies provide services for us. Since the early 1990s, the term ‘context’ has drawn the increasing attention of researchers in mobile and ubiquitous computing (e.g. Weiser, 1993; Schilit *et al.*, 1994). With the fast development of wireless communication and increasing use of mobile devices, context-awareness has become an increasingly important consideration in developing systems and applications (Dix and Abowd, 1996; Long *et al.*, 1996; Fickas *et al.*, 1997). Being aware of the context(s) in which systems and applications are run can improve their ability to adapt and react to user situations, such as surrounding locations, people and objects such as other devices. Thus, such context-awareness is better able to support computer use in applications of mobile and ubiquitous computing, and in different environments. The early focus of research was mainly on the design of devices but has broadened to a consideration of context and context-awareness in systems and applications.

*Context* is generally regarded as data or information which describes the situation that is relevant to and has influence on the state of users, systems and applications. ‘Context is any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves’ (Dey, 2001 p. 5). Accordingly, any information can be viewed as context if it can be used to characterize the user situation in an interaction. Where you are, who you are with and what resources are nearby have also been highlighted as three important aspects of context in mobile distributed computing (Schilit *et al.*, 1994; Pascoe, 1998).

Location and its related features have tended to dominate the use of context, although context is much more than location. However,

other elements of context tend to be comparatively more difficult to identify or measure. Context(s), other than location, can include physical features such as lighting condition and background noise level, resources such as networks and communication capability, surrounding people and objects. Furthermore, relevant human factors are used to describe the context alongside those related to the physical environment (Schmidt *et al.*, 1999). Such context can include information about users, user social environment and user tasks. Time is another factor which should be considered as part of context. Context can also be defined as active context and passive context depending on its influence on the behaviour(s) of an application or whether context is only relevant (but not critical) to an application (Chen and Kotz, 2000).

It cannot be stated definitively which aspect of every situation is going to be important as context; there will always be subtle changes as no two situations are likely to be the same. It should nevertheless be possible to define and measure contextual elements for broad classes of situations. There are, however, many definitions and surveys of context presented within the literature. Some of them tend to be broad ranging and vague, resulting in 'context' being used as a general word that is loosely defined, whilst others view 'context' as being more narrowly defined and specific about the types of information included (Chen and Kotz, 2000; Dix *et al.*, 2000; Dey, 2001). The importance of context is, nevertheless, growing as interactive mobile devices need to have situated contexts reflected in their system design if they are to achieve user acceptability. Another important characteristic of context is that it is dynamic (Li and Willis, 2006). This is discussed further in Section 7.7.

*Context-aware* is often described as the ability to discover changes in user situations (context), being responsive to them and using acquired context to provide services in meeting user needs (Schilit and Theimer, 1994; Dey, 2001; O'Hare and O'Grady, 2003). In context-aware systems, contextual information can either be obtained by requiring users to enter data into the system, or can be acquired by sensing devices, monitoring patterns of use and the surrounding environment. Such contextual information is then used to dynamically tailor the response of the system. By being context-aware, systems and applications can better support their users with features such as supporting the presentation of information to users, triggering the automatic execution of a service, or tagging information with some aspect of context (such as location or time) to better facilitate later retrieval. Many context-aware systems and applications use physical

location as the main contextual element (Chen and Kotz, 2000); there are some that combine location with other features such as user preferences. As stated above, context history stored within a context-aware system is widely believed to be useful but in reality is rarely used. Instead, most applications tend to use small amounts of contextual information in a piecemeal way.

### *Some Application Examples of Context-Aware Computing and Systems*

In Section 6.5.4 reference was made to the Active Badge system (Want *et al.*, 1992), which could sense a user's indoor location using infrared positioning and could thus automatically adapt applications such as call forwarding. In the following selection of experimental systems, the location where the device is being used strongly influences the information presented to the user and the interaction possibilities with the device:

- **Cyberguide** (Long *et al.*, 1996; Abowd *et al.*, 1997): context-aware information retrieval. It provides information to tourists about their current location. Tourist travel tracks are also recorded over time, and used to provide tourists with further information about points of interest.
- **City Guide** (Kreller *et al.*, 1998): multimedia city guide services on a 3G network.
- **Mobile Guide** (Pfeifer *et al.*, 1998): investigation into generic platforms for supporting location-aware services. Emphasis was on middleware needed to support heterogeneous data sources and adapting content for different devices.
- **C-MAP** (Fels *et al.*, 1998): supporting communication between exhibitors and visitors at an exhibition. It provides the context-aware mobile assistance for visitors at exhibitions. Primary concern was interface design issues.
- **Travel MATE** (Julia and Bing, 1999): multimodal user interface aimed at active mobile tourists.
- **ComMotion** (Marmasse and Schmandt, 2000): location-aware information delivery. Using current location and time as context, the application delivers relevant messages to users.
- **A nomadic exhibition guide** (Oppermann and Specht, 2000): Adaptive guide to exhibition visitors with location information.

- **GUIDE** (Cheverst *et al.*, 2000): a context-sensitive tourist guide for the city of Lancaster. The application combines the contexts of location and user preference in adapting the service provided across a wireless network.
- **Gulliver's Genie** (O'Hare and O'Grady, 2003): a context-aware multimedia tourist guide to assist roaming tourists. The system considers context in terms of location, orientation and user profile.
- **Enhancing spatial literacy skills** (Priestnall and Polmear, 2006): a location-aware mobile system for exploring and enhancing the spatial literacy skills of students during field work.

LBS applications (current and potential ones) discussed in Section 4.7 can be viewed as context-aware applications. Location is obviously the main context used in most of these applications. However, other context information does need to be brought into LBS research. In the following sections, the discussion of context is focused on LBS.

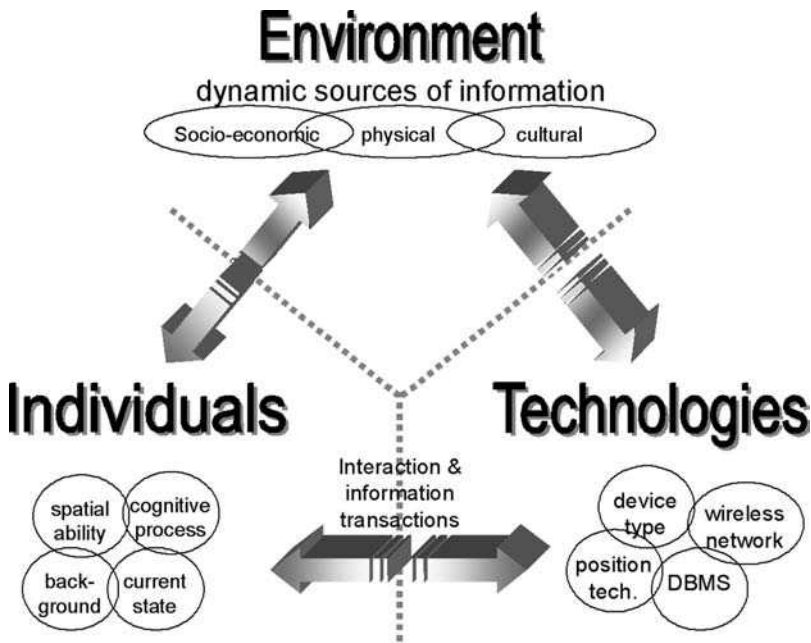
### 7.3 Context in LBS

---

LBS applications aim to deliver spatial and spatially-related information and services to individuals on the move. As users become familiar with LBS, they will increasingly require more relevant and timely services via their mobile devices in order to perform the tasks on which they are engaged whether these are for business or are purely social. As seen in Chapter 1, increasing mobility is a feature of our post-industrial society (Section 1.4). Mobility in LBS needs to take account of not just the user, who is on the move, but the environment in which the user is embedded, as well as the devices themselves and the nature of content to be delivered across a network. It hardly needs be said, but nevertheless should be emphasized, that mobility underscores all features of LBS. Furthermore, the mobile devices and the systems that are the means of accessing LBS are becoming more embedded into our daily activities and surrounding environments. The situations in which LBS are accessed are not only diverse but they are dynamic. To provide relevant data and information services to users in mobile and dynamic situations, it is critically important for LBS to be able to sense and respond to user's situations and surrounding. Therefore, a better

understanding of context and context-awareness in LBS is crucial in improving LBS capability to provide more effective services to meet user needs. The greater the utility of the information provided, the more likely users will refer to such services and be willing to pay for the services (Chapter 10).

In the previous section, a range of definitions of context were presented. Whilst there are clearly some important dimensions to context, we would argue that these need to be considered not only in their own right but also in how they interact during the information transactions that occur during LBS usage. Dourish (2004), for example, has argued that context should be defined as a dynamic interaction issue, not just as presentational issue. Context in LBS is relevant to particular settings (technological and environmental), the actions that take place and the parties involved in the action along with their particular goals and preferences. This is a much broader focus and imposes more complex demands than for many other types of IT services. Hence, the discussion below will stress an interaction view of context for LBS (Figure 7.1).



**Figure 7.1** Conceptual interaction model for context in LBS (see text for explanation; based on Li, 2006).



When LBS are delivered and used, there are real-time interactions and information flows between users, environments and devices. This interaction concept embeds three main dimensions: environment, individuals (as users) and technology. Delivering and using LBS involves dynamic interactions and information transactions between them. Based on this dynamic interaction concept, environment, individuals and technology can be used as three main strands to study context in LBS, with a focus on the dynamic real-time interactions. Context from these three strands are thus interwoven:

- Environment, as context, can be viewed as comprising the physical, socio-economic and cultural aspects of our surroundings. The environment context should also consider the context(s) of the various situations encountered by users. Context identified in this strand can be locations, surrounding settings and objects, and encountered situations. Time, either in an absolute sense (e.g. when, day/night) or in a relative sense that would include elapsed time (duration of time since an activity started or over which an activity is expected to take place), is also an important aspect of the environment context. Related social and cultural characteristics of location can also contribute to this aspect of context.
- Context related to individuals (as users) can have a direct influence on how data and information services might be requested, accessed and used via a mobile device. In the interaction discussed above, users can gain information both from the environment and LBS while they act and move within an environment. User context reflects personal differences in terms of age, gender, background, prior knowledge and their preferences. In addition, user situations (e.g. emotional state or physiological condition) should be taken into account as context.
- Technology, as context, concerns the technologies used for LBS. Such context can include wireless telecommunication networks, positioning technologies and mobile devices as well as GIS, and relevant data resources. In LBS applications, there are a range of technologies involved in delivering data and information services to users on the move. Mobile devices can be designed and customized to adapt to the mobility of users and the environment. The technological context will be reflected in the design, deployment and use of LBS.

Context from the perspective of the interaction between these three strands will influence both content and service delivery in LBS. For example, how geographical data in a LBS application are selected, analysed and delivered to a mobile device should be varied according to the specific situational context. This is not limited to the design of a context-aware device but should extend, for example, to the way GIS are used in handling geographical data.

One important issue in context is the dynamics of the interaction. Contexts in mobile situations are likely to be continuously changing during the period over which a service is used. Thus context cannot be treated as a stable, external description of a setting in which an activity is occurring, nor can it be assumed *a priori*. The context of a service request emerges and evolves during the interactions that take place. However, often in context-aware applications, the dynamic characteristics of context are not taken into account. Take using location context as an example. Although location context is critical to LBS and is used in many applications as the only form of context, location is often considered merely as an index from which to infer the overall context influencing a mobile application (Dix *et al.*, 2000). It appears that the dynamics of the environment, whether this is location or some other characteristics of the physical space, are not sufficiently taken into account in the interaction, as it is assumed that the sole interaction is between the user and the device. The consequence is that the surrounding environment has not been regarded as a mutable information source with which users interact. Thus the dynamic characteristics of context have generally not been explicitly addressed when defining context and context-awareness of systems and applications. With the highly mobile and geographically situated nature of LBS, the dynamics of context need be better reflected in the design of LBS applications.

## 7.4 Environment as Context

---

Environment context is essential for LBS applications, having an influence generally on the nature of services provided but more specifically on content, presentation and delivery of the information. Users use mobile devices and carry out activities within a broader environment. Interactions between users and mobile devices are influenced by the nature of the environment (e.g. urban or rural) and what is happening

within the environment (e.g. raining or traffic congestion). Also users interact directly with the environment and obtain information from it (e.g. observing where some landmark is). Furthermore, mobile devices interact with the environment in a way by which the information and services provided through the devices can reflect and be adapted to the environment situation (e.g. tailoring the information and its presentation by being location-aware, or in presenting a relevant pushed alert about some issue to do with the environment). Thus improved identification and better understanding of environment context is likely to enhance a service provider's ability to deliver content pertinent to the environment of a user, and to represent and deliver such content in a way that is also pertinent to the user's mobile situation within that environment. An example has already been cited in Section 4.8, where the time of day and prevailing weather conditions ought to influence the appropriate LBS response to a query. Being directed to walk across a wooded park at night as the shortest route to some destination would be poorly viewed by most users – just as much so if the park gates are locked! Environment as context can also extend to the physical, social-economic and cultural setting. For example, different social and cultural settings may require information to be sensitively presented according to established norms. Within the scope of this book, the focus is more on environment context as relating to location, time and situation within the physical environment.

Location is an essential and unique aspect in LBS arising from its definition. Many mobile devices have been developed to be location-aware by deploying a range of positioning technologies (Chapter 6). Thus location information is fast becoming an integral part of mobile device technology, though by no means used in many of the services available through them. Key as it is to the provision of LBS, location information tends to be viewed narrowly as referring only to the  $x,y$  location of a user as a co-ordinate pair rather than taking a broader view of the environment context that emerge from it. Thus, location context in LBS can be viewed from the following perspectives:

- Location information as a physical position in space with exact coordinates either in 2-D or in 3-D. Certain types of mobile devices equipped with GPS or A-GPS can provide such location information. For those LBS applications requiring a relatively accurate position fix of the user, such precise location information is necessary for establishing the location context. Emergency services in mobile situations can

be one such application. In some incidents, callers may not be able to describe their exact location. The location obtained from mobile devices at an incident could be essential for the emergency services to respond and render prompt assistance. Another example is wayfinding assistance, such as in-car navigation systems where an accurate location context is necessary for generating timely route instructions.

- Location information providing an approximate physical area. Such areas can be a cell area obtained from a mobile network Cell-ID (Section 6.4.1), a certain area covered by Bluetooth signals or a certain segment of road network. Such location context can be used by LBS applications providing information and services that are geographically specific, such as traffic condition alerts, road tolls and local maps for travellers. This approximate area location context can also benefit wayfinding assistance.
- Topological location information which can be given in relation to other locations or objects. For example, location information can be sensed and provided as being near or close to a certain place, being north or south of a location. Topology is discussed in detail in Section 8.3. Applications such as buddy systems to know when friends are nearby are one example using this type of location context. Receiving pushed alerts (e.g. notification, advertising) to opportunities for transactions that are location specific (such as at a nearby shops) is another example using topological location context.
- Location information with semantic meaning. Such location information can be a location identified with an attached meaning, such as home or work place, a town centre or a rural area, a pedestrian area or a road type (e.g. motorway). The semantic meaning of a location can be combined with the other forms of location context being discussed here.
- Location information as a set of dynamic physical positions. This can represent changes of locations, such as when tracking a user (Section 7.7).

These different ways of viewing location context are not mutually exclusive categories. For example, a home can be presented as a position with  $x,y$  co-ordinates, and can also be described with topological location information as being near a certain location, and of course it has semantic meaning. Thus, most of the above perspectives of location as

context can be used in LBS as inter-linked contexts. On the other hand, some LBS applications may only need to have a narrow view of location as context, such that proximity area location or topological location could be sufficient to know which geographical areas are pertinent rather than generating a precise position. For example, providing information on the availability of parking spaces in a city centre needs only an approximate position fix of the user to the relevant area. Each of above perspectives of location as context can either be used in combination or independently according to the applications being delivered and the user activities involved. In addition, the level of user mobility may also affect the way location is used as context.

Time is another environment context. Time and geographic location are closely connected, as users are located in a particular place at a particular time, or for a particular duration of time. Thus, user participation in tasks and activities has both a spatial and a temporal dimension. Time context can be absolute time (e.g. opening time of a museum), may be a period (e.g. day or night, rush hour), may be an elapsed time (e.g. the time taken to travel from one location to another) or may be expressed in relative terms (e.g. the museum will be open for another five hours). Time context can be generalized and expressed as a time of year to reflect seasonal effects. The change of location per unit of time can also derive context, such as the speed at which a car is travelling, and can affect the type of information and level of detail that is relevant. Furthermore, time context as elapsed time can be used to establish a record of contextualization over a period that can help in profiling users and their preferences. Although time context in LBS is not emphasized as much as location context, it should be equally considered. Apart from the examples of time context above, out-of-date information can be irrelevant, even dangerous in some cases. This is particularly so in mobile situations. It is not just the dynamics of the user that need to be considered but the dynamics of the data resources being used and their update frequency (Section 5.3.5 and Section 8.6.5). Spatial-temporal contexts can be considered as an integral form of context. Studies have been carried out to adapt information content based on spatial-temporal context so as to maximize information utility to users (Brimicombe and Li, 2006).

Situation context as related to the environment is based on the concept of users' interactions with their surroundings, mobile devices and other objects. The situations that users encounter while using their device can have an important effect on what information and services are requested, the way they should be presented and, indeed, the way in

which they may be used. For example, LBS applications could vary levels of screen backlight or font size according to the ambient light levels. Other physical situations might be the level of ambient noise and the weather conditions. This can also be related to the tasks or activities that users are engaged in. For example, wayfinding information provided to the driver of a vehicle needs to take account of one-way streets whereas this would not be necessary for a pedestrian. Situation context can be interwoven with other contexts, such as location, time and user context; this is discussed in Section 7.6. Situation context can be mobile and dynamic, changing with location and with time.

In LBS, awareness of environment context can be achieved through a range of technologies. Location information can be obtained using various positioning technologies (Chapter 6); speed and direction of movement can be calculated from position changes over time; sensors in mobile devices can be used to measure levels of noise and light; weather conditions from on-line resources and so on. Furthermore, environment context is spatio-temporal and the history of context over time can be analysed into profiles using data mining techniques. However, the level of such awareness needed by LBS depends on the requirements of specific applications.

## **7.5 Technology as Context**

---

Technology as context plays an important role in the development of LBS applications as it can affect the interaction between the technology and the user and, based on the user's consequent decision making, the subsequent interaction between the user and the environment. Technology context needs to be considered at an early stage in the system and interface design process. On the other hand, technology as context can also have a considerable effect on LBS applications during their usage. The types of information and the nature of services provided by LBS may vary with the effect of technology context. For example, how responses are formatted may depend on the different display capability of users' mobile devices. LBS applications should reflect the state of technologies being used. The variability of technology context can influence the services requested by users of LBS, particularly depending on user interfaces and interaction styles. The communication established between users and devices can thus be influenced by the technology context. Because LBS are heterogeneous

technologies there can be an influence from the range of embedded technologies – hardware, software and data resources – that are brought together to produce a service. In this section, the focus is on the technology context related to devices, the connectivity of those devices through the network to a provider, and the capability of providing pertinent spatial information to users in mobile situations.

The technology context based on mobile devices can be related to the type of mobile device, features of devices and the technologies embedded within devices. This context can reflect on many aspects of LBS applications. Mobile devices used for LBS applications can be diverse, such as mobile phones (cell phones), PDAs and a range of palm-top and tablet computers. Although the boundary between different categories of mobile devices is becoming blurred with the convergence of devices (Chapter 2), different features of devices can have an influence on how LBS can provide data, information and services to users. The screen of a mobile device can be one most noticeable and direct feature users interact with. The range and quality of LBS applications that can be provided are directly influenced by the screen characteristics. The amount of information that can be provided and the way it is presented on a per screen basis will vary depending on the size and quality of screen.

Applications can be designed to adapt to the features of devices used, such as multiple-device authoring allowing tagging for different levels of abstraction, device-specific authoring of Web pages, navigation and automatic re-authoring from client side (Bickmore and Schilit, 1997). The mode of communication appropriate for specific applications will also need to respond to device types and features. For example, a larger screen device with higher resolution may be more suitable for the communication of information through maps or 3-D visualizations containing more detail, whilst for other devices with small screens there would be benefits in using alternative modes (e.g. text, voice and symbols) or with more abstract content (e.g. schematic map) for delivering information and services. The utility of the information provided to an end user is not just a matter of content but of the opportunities afforded by their mobile device to select and display information and, more generally, to have a sufficiently intuitive and information-rich means of communication between user and service provider.

Technologies embedded in mobile devices can also contribute to the device-based technology context. Consider the available positioning technology used in a mobile device which can have direct effect on LBS applications. A GPS-enabled mobile device can provide a more



accurate position fix than if it were achieved using only network-based technology. However, the accuracy of positioning can change in areas with dense high buildings where GPS signals can be of insufficient strength. Other technologies used in devices, such as those related to battery life and available computational resources, can also have an effect on what LBS can offer. Furthermore, technologies employed in devices can determine what types of network services can be accessed by users and levels of data transfer speed that can be achieved. For instance, 2G mobile phones cannot use 3G networks, and therefore cannot access certain LBS applications designed for 3G phones. To provide continuous and relevant information and services, LBS applications need to aware and respond to this type of context.

Another important aspect of technology context in LBS concerns network connectivity, which has a direct effect on the interaction between applications and underlying technologies and between users and applications (possibly even affecting the application's user interface). Such context can be viewed as two aspects: one from a systems perspective, the another from a network perspective. Let's first consider connectivity from a systems point of view. Functionality in LBS seldom reside exclusively in a mobile device. The functionality is more likely provided by a whole system and distributed over networks. It can be viewed as the whole system supporting LBS applications. This is particularly the case for applications in mobile situations. The interactive properties of the system need to be considered in relation to the distributed nature of LBS applications and the different levels of functionality they may offer. In mobile situations these interactive properties of the system become amplified because the properties of the infrastructure may vary dynamically as the application is in use. This is often called system context and 'refers to all the interconnected devices and their applications which constitute the system as a whole' (Dix *et al.*, 2000). This includes hardware, software and data resources. There is a need to be aware of this context for the communication modes and presentation methods used in LBS applications (Chapter 9). Context derived from a system can also relate to the connectivity between a mobile device and other devices in its vicinity that might be available to support an application. System context-awareness will allow providers to design and deliver more advanced services to users.

The network perspective of context has a direct influence on connectivity. It reflects the integration of devices, telecommunication and other wireless networks, and applications. In Chapter 2, wireless networks were discussed as one of the convergent technologies that are

making LBS possible. Networks as context can have an influence on the validity of the information provided through LBS applications in terms of timely response or updating of the information delivered to users and even on the accuracy of the information delivered. Irrelevant and late information due to problems in the connectivity may cause users to misunderstand and misinterpret the situation and the environment they are in. This is particularly critical in LBS because the information delivered is expected to be relevant to a user's location in real-time. Problems such as breaks in connectivity between a mobile device and network will prevent users from receiving updated location-related information or feedback from requests (Raptis *et al.*, 2005; Section 10.2.2). In emergency situations, delays in receiving alerts and critical data for response or avoidance could have devastating consequences.

In being aware of such contextual possibilities, application developers need to find ways of minimizing the possible impacts. The level of connectivity (e.g. bandwidth and speed) may alter the way users interact with the technology and with the environment, and therefore may affect types of services that users are willing to engage with. Connectivity can also influence the application's user interface. For example, an in-car navigation system not only has functionality to generate and display maps and route information for navigation, but also has integrated positioning technology (e.g. GPS) that is essential for providing accurate navigation information. In the meantime, other elements and resources from the network (e.g. traffic monitoring) also contribute functionality as a whole to provide up-to-date, accurate information. Problems occurring in the connectivity of a system (either the mismatch between positioning obtained and navigation instructions given; or delays in information across the network) can result in loss of information utility, loss of credibility in the product and may even put users at risk.

A particular technology context for LBS is the subsystem(s) for handling geographic information; that is the capability of an LBS provider to deliver pertinent and timely location-specific information to users on the move. The subsystem also needs the necessary functionality to communicate and present information in appropriate modes and formats. GIS and spatial databases provide the ability to store and integrate spatial data, and to process, analyse and disseminate geographic information (Chapter 3). User requests for information need to be augmented with a user's location and formalized as a spatial query. The query is then processed by a spatial database that is either an

integral part of GIS or is a stand-alone spatial database that can be coupled with GIS. Such queries and the response time to be processed are critical to LBS applications. A query can be a simple request, such as where is the nearest restaurant, or more complex, such as directions on how to get from a current location to some other target location via the shortest or easiest way, that might involve different modes of transport (walking, bus, train). Spatial queries are discussed in detail in Chapter 8. The ability of the GIS and/or spatial databases to process nontrivial requests in a timely manner and the scalability of the system to handle large numbers of simultaneous requests is clearly a key technology context in LBS. To an important extent, the range of services that can be offered by an LBS provider and the level of utility that can be achieved are very much dependent on the speed and sophistication of the spatial data handling capability within the overall system.

## **7.6 User as Context**

---

User context is one key aspect that can enhance an LBS provider's capability of delivering pertinent data and information services to users. LBS applications can be made more personalized and resonate with users by adapting content to user needs through system design, choice of content and means of presentation. User context can include many aspects of the user, such as personal characteristics, preferences and frequent situations or activities (e.g. commuting to work). In addition, user context can directly influence the way users interact with their mobile devices, and with the surrounding environment. Furthermore, interactivity between user and mobile device can be studied from the perspective of the device's ability to obtain and interpret contextual information (Dey, 2001; Schmidt and Van Laerhoven, 2001). Depending on the application, it may not always be feasible to clearly distinguish each individual user in order to provide personalized services. Such personalization could be viewed as intrusive and it may also be inconvenient to the user to provide every detail. Therefore, user context aims to provide relevant contextual information of a user and/or class of users. User context is discussed from four main perspectives: user personal characteristics, user knowledge, user preferences and behaviour, and user situation.

User personal characteristics can include demographic information about users (e.g. age, gender), user physiology and user cultural

background. Such a user profile could be established at the point of setting up a subscription to a service or could be stored in a file (encrypted if necessary) on the user mobile device and read by the system when a service is requested. Thus information delivered to the mobile devices of elderly users could be presented with bigger fonts in response to the specific user context of age. This could also be the case for visually impaired users for whom voice might be the preferred mode of receiving information. Thus the mode of communication can be adapted to users with specific needs or preferences. The cultural background of users can cover a broad range of user characteristics and preferences. This could extend to the choice of language used in communication, the nature of the information content and the way they are presented (Section 3.4.3). Furthermore, the personal characteristics of users can also have an influence on the way they behave both in requesting information and in using the response – remember, not all members of the public are necessarily competent map users, or indeed necessarily like using maps. How personal characteristics are used as context will depend on the nature of the specific LBS application.

User knowledge can be an important aspect of user context. Included here is the education and experience of the individual. For example, those whose education and experience has endowed them with strong IT skills and familiarity with certain types of technology should feel confident in using mobile devices for LBS; those with weak IT skills (particularly the elderly who have not grown up with these technologies) may feel less confident and may be put off by the apparent complexity. Key for LBS applications is the spatial ability of users – their ability to learn geographical situations and reason spatially. But this aspect of user context reflects on the level of prior knowledge of locations that can be assumed within applications. In Section 5.2 it was discussed how, in requesting route directions, it is in most cases inappropriate to give detailed instructions around the user's home (an area they are usually well familiar with) but that it is the 'last ten kilometres' (LTK) of a journey to an unfamiliar location that are the most critical in terms of having precise and unambiguous instructions.

Both user preferences and behaviour can contribute to user context. Preferences often describe the relatively general interests and partiality of users. They can be interests in particular sites and events, or be preferences of using one type of technology or one mode of communication for retrieving information. Thus, LBS applications can be adapted to specific user needs by being aware and responding to user preferences. For example, there are already non-LBS applications to

mobile devices for a range of sporting enthusiasts for ‘push’ information to keep them up-to-the-moment on events and results. Such applications could easily be made to have a spatial dimension. However, user preferences can change over time, and can also vary in different locations according to the situations faced by users. Therefore, it is a challenge to personalize a system in LBS according to user preferences in mobile situations; this is something that is discussed further in Chapter 9. User behaviour is the general conduct and demeanour of users; it also leads to patterns that can be identified. These might revolve around daily activities such as work and attendant travel routines as well as social activity patterns and social networking. The frequency with which certain activities are repeated can contribute to user context. The history of activity over time or at location(s) can be built into a personal behaviour profile by service providers and used as context. Artificial neural nets could be used for this purpose in order to either classify users into groups according to their characteristics and preferences or to anticipate users’ intentions and information requirements. Thus personal profiles can be used to help service providers increase the utility of the information delivered. The use of location and user movement history has been researched as a means of ensuring relevant information provided (Mountain and Raper, 2001; Fogli *et al.*, 2003). User-adaptive content for LBS applications has been studied to ascertain users’ interests and preferences in areas such as map representation with different levels of generalization and different types of content (Zipf and Jöst, 2006).

User situation is another aspect of user context that should be taken into account. User mobility in LBS can be understood from the perspective of the situational reason of why a user needs to be mobile at a particular time and from the perspective of the situation(s) that mobility brings them to in terms of the surrounding environment. Situation context as related to users’ physical environments is described in Section 7.4 above. User situation context is more focused on the state of users in different mobile situations. It can be the emotional state and physiological condition of users at the time of using LBS. It can be also related to the nature of the tasks and activities currently being carried out, such as the level of urgency, and objectives of the tasks. Taking a wayfinding scenario, the situation can be very different when a user needs urgently to find a meeting place or a user is exploring a place at leisure. The former would not be interested in information pertaining to historic facts of certain buildings, whilst the latter might be keen to receive such information. Thus, LBS can adapt

both content and presentation to user situation context in order to deliver more pertinent information to users on the move.

Personalization of LBS responses based on user context is one way to enhance the utility of information delivered by providing only essential and pertinent information that the user probably requires. However, there are challenges in dealing with changes in personal profiles. The built and stored preference profiles could vary over time, in different situations and for various services. There is also an issue concerning the level of control users prefer to have. Users can be given the choice to personalize the system or not, depending on how much effort they are willing to put in. Despite the fact that users would undoubtedly benefit from a personalized service, they may not be ready to put in the necessary effort to define a profile either for each service or for each context of use. On the other hand, in line with many changes on the Internet and the emergence of Web 2.0 (Section 2.2.2), users may like opportunities to participate in content creation rather than being passive information consumers.

Thus LBS with user context-aware features should not necessarily result in a predestined or over-controlled service environment. The user should feel and be in control. It is suggested that users should be able to override the defaults of a system (Cheverst *et al.*, 2000). In one study (Barkhuus and Dey, 2003), users' perceived sense of control over the content within mobile devices was explored by providing users with context-aware features at three levels of interactivity: users specifying their own settings for how the application should behave in a given situation; users being presented with updated context information and allowed them to decide how to change the application behaviour; and systems changing the application behaviour autonomously according to the sensed context information. A sense of lack of control was felt by users with the more autonomous approach. Users do prefer some degree of context-adapted automation, as long as the application's usefulness is greater than the cost of limited control. Importantly, privacy of users is another issue which should be considered in using user context, and is discussed in Chapter 10.

## 7.7 Dynamics of Context

---

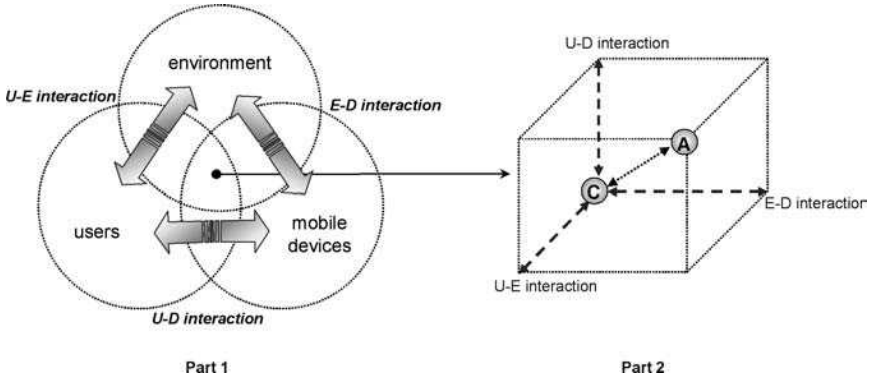
One important aspect of context in LBS is its dynamic characteristics that are embedded within the context. As discussed in Section 7.3,

context in LBS is understood through a dynamic interaction concept (Figure 7.1). The three main elements in the concept (environment, technology and user) have been used as three main strands to define environment context, technology context and user context in LBS. Due to mobility being a fundamental aspect, LBS needs to be studied through the real-time dynamic interaction between user, environments and technologies. Therefore, it is important to understand the context in LBS from this dynamic perspective.

Context is not constant unchanged information. One obvious and unique dynamic aspect in LBS is the environment. The environment is full of changes, such as changes of location, time and situation. Such dynamic characteristics in environment context require LBS applications to be aware of the continuous change in the context. Another aspect of dynamics in context is in user context, which includes variation in user preferences, in user emotional and physiological conditions, all of which can change over time and from place to place. Personal preferences and behaviours can also vary according to different activities and usage situations. The technology context contributes to a further aspect of the dynamics in context. The availability and usage of a range of technologies to deliver LBS is more likely to change in mobile situations, which will affect the ability of a provider to deliver consistent and uninterrupted services. Although these three forms of context are often only considered in application development individually, they are actually interwoven. And so it is with the dynamic nature of these contexts. Such interwoven dynamics have effects on the way users interact with devices and with surrounding environments. It is essential in the development of LBS to establish a clear understanding of contexts within a dynamic real-time interaction framework (i.e. how users interact with technology and with environment in practice).

To address these dynamic aspects, Figure 7.2 illustrates a concept of how dynamic contexts can be considered during interactions. The first part (Part 1) of Figure 7.2 focuses on the context-aware interaction. The three main elements in the interaction (environment, mobile devices and users) form the environment, technology and user contexts. Three sets of interactions are shown and directly represent the dynamics in contexts. These are: user–environment (U–E) interaction, user–device (U–D) interaction and environment–device (E–D) interaction. The aspect of the U–E interaction can reflect particularly the dynamics in context of environment and users, for instance, the dynamics in a location-related situation. The aspect of U–D interaction can reveal the dynamics in user and technology context, for





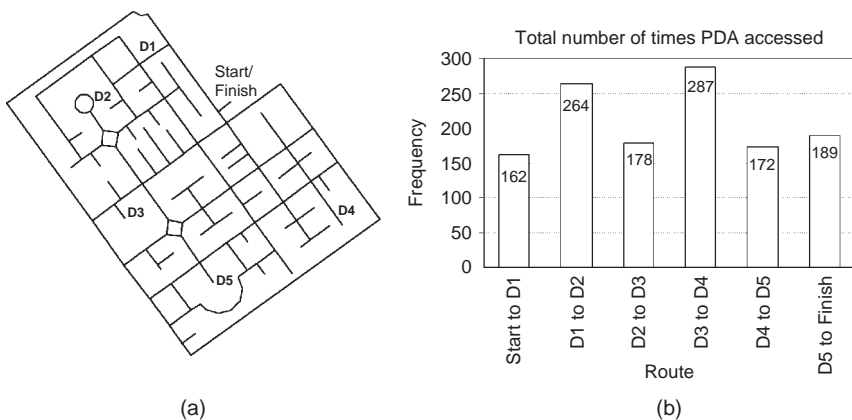
**Figure 7.2** An interaction concept of dynamic contexts (for explanation see text; based on Li and Willis, 2006).

example the adaptive user interface in a mobile device or changes in the degree of automation within a device for resolving queries. The aspect of E–D interaction can reflect the dynamic changes in both environment context and technology context. Representation of a surrounding environment (a place) through mobile devices could be either very specific (highly detailed) or more generalized. Such representation can change across different environments and different technologies employed in a mobile device. It is important to bear in mind that these three dynamic aspects of context should be considered in their totality in the interaction even though they are described individually.

The three dimension cube in Part 2 of Figure 7.2 addresses the dynamic nature of context from the interactive concept illustrated in Part 1. The three axes represent the dynamic aspects of the U–E interaction, U–D interaction and E–D interaction respectively, with the space in the 3-D cube representing the intersection of the three aspects. In all three identified aspects of the interaction, the interaction is likely to be more active (moving away from ‘C’ on each axis) as the contexts involved become more dynamic. The position ‘C’ symbolises situations with greatest certainty and less varying contexts, whilst position ‘A’ symbolises those with more dynamic contexts and high level of interaction and therefore greatest ambiguity. The certainty of users varies in most real world situations, hence the way in which users interact with devices and the information required from applications also differ.

Here two examples are shown to demonstrate some aspects of dynamics in context. The first example is a pedestrian wayfinding study

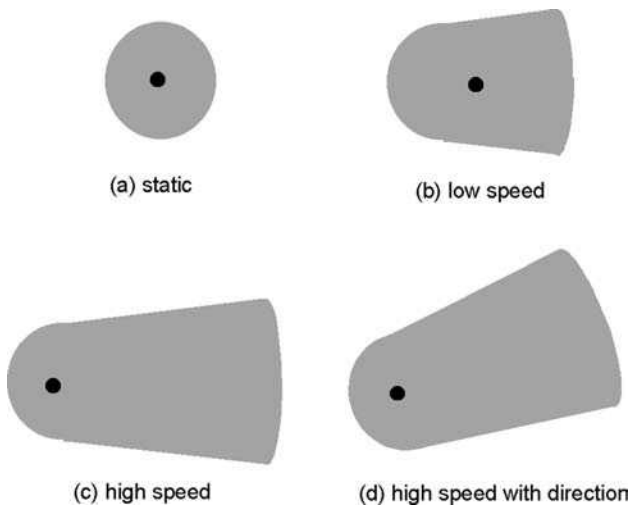
in a residential area (Li, 2006; Li and Willis, 2006). In the study, 27 participants, all in an unfamiliar area, were required to find five destinations from a starting point and then return to the starting point (Figure 7.3a). All participants could access available wayfinding assistance information from a PDA at any location and at any time. Each route between successive destinations was chosen to have different levels of complexity in terms of numbers of turnings and number of decision points passed. Captured in the experiment is the usage of information including frequency and time used in accessing information, and the types of information accessed. Only the number of times the participants accessed the PDA to acquire information to assist their wayfinding activity is looked at here; the total number of times of accessing information for each route is shown in Figure 7.3b. In completing the wayfinding task for each of the six routes, participants had different levels of interaction with the mobile device to get information. During route 2 (D1 to D2) and route 4 (D3 to D4), participants perceived the environment they encountered as being more complex than other routes, therefore they interacted with the device more frequently. The change in access frequency for each route reflects how environment context influences the interaction. The dynamics of context need to be studied as a real-time interaction. The study also showed that there are changes in the type of information used (such as area maps and step-by-step written instructions) along the wayfinding routes (Section 9.5.2)



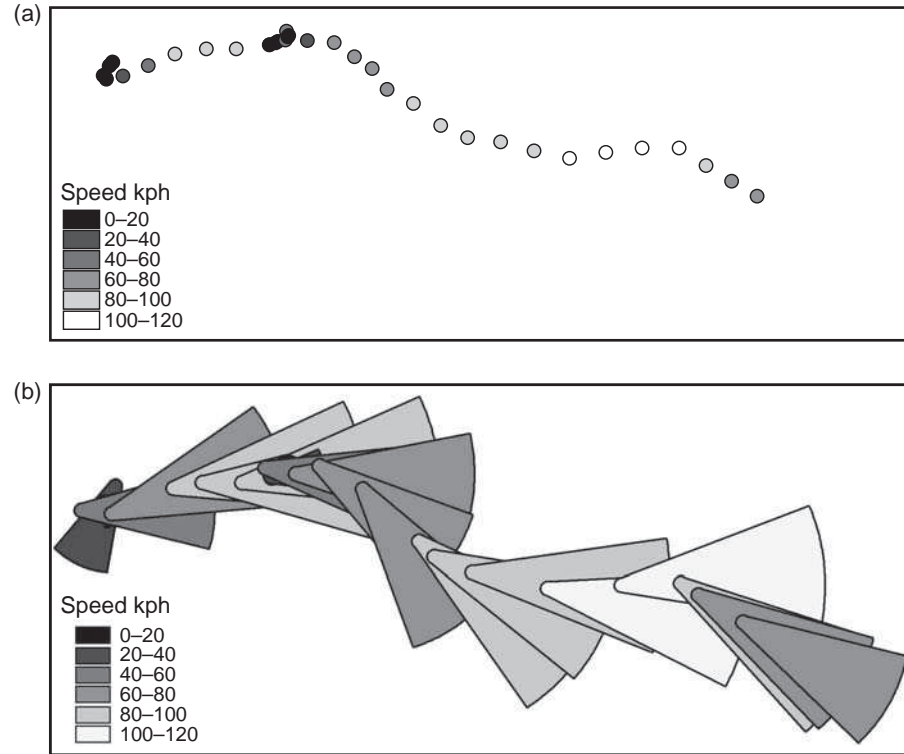
**Figure 7.3** A wayfinding case study: (a) sketch map of the residential area in which the tasks were performed; (b) the total number of times the PDA was accessed by the participants on each wayfinding task (from Li and Willis, 2006).

The second example is a mobile space–time envelope approach to ‘simultaneously provide “soft clip” pruning of candidate data sets in anticipation of queries, and provide the trigger that users are pertinently in-range for alerts’ (road works, accidents, weather conditions, etc.) in a car travel scenario (Brimicombe and Li, 2006 p. 21). The mobile space–time envelope is constructed based on an awareness of the changing context of user location, direction of movement and speed of travel, and then adapts its shape accordingly to provide more immediate and geographically more wide ranging (and yet pertinent) information than just the proposed itinerary itself. A mobile space–time envelope is constructed considering user current location, travel speed and movement direction. Where a user is stationary (travel speed is zero), the shape of the envelop can be clipped into a circular zone (shown as static zone in Figure 7.4a), which implies that the user could subsequently move in any direction, information and queries could relate to any direction of movement and pushed alerts for any event or location within this zone would be welcome.

As a user’s velocity increases, the shape of the mobile space–time envelope is modified in such a way that it stretches ahead in the direction of movement of the user (as shown in Figures 7.4b, 7.4c and 7.4d). The furthest extent of the envelope is determined by the distance travelled in  $t$  minutes at the current velocity  $v$ . When the



**Figure 7.4** A set of mobile space–time envelopes adapting to speed and direction (based on Brimicombe and Li, 2006).



**Figure 7.5** Dynamic mobile space-time envelopes for a section of journey: (a) location of car for every three minutes travel, classified according to speed; (b) corresponding mobile space-time envelopes (based on Brimicombe and Li, 2006).

velocity increases, the envelope becomes proportionally elongated; and the location of the user is positioned towards the back of the envelope (Figure 7.4c) in comparison with the location in a lower velocity situation (Figure 7.4b). This reflects the fact that there will be greater interest in what is ahead and less willingness to stop and turn around when a user is moving forward fast. A stationary or slowly moving user is at or near the centre of the envelope. The mobile space-time envelope determines the geographical extent or zone of immediate interest to a user and adapts its shape and orientation according to the changing context of user location, travel speed and direction of movement. The way these envelopes appear for a section of journey by car is shown in Figure 7.5. This technique is further discussed in Section 8.7.4.

With the understanding of the dynamic nature of context, it is necessary to be aware how LBS applications should respond and adapt to the dynamic context. The nature of applications and the situations in which they are employed have a direct influence on which features of context are going to be relevant. These need to be explicitly identified in the design of applications. It is thus important to consider the relationship between mobile devices and their users and the situated environments in a real-time interaction during the design process. In the case of LBS applications the normal design considerations are amplified by the need to consider the real-time dynamics. A context-aware user interface can be implemented to select appropriate modes and interaction styles that allow adaptation to the surrounding environment and conditions. This includes the choice of content, level of detail and mode of representation. However, there are challenges for the context-aware user interface in mobile situations due to the fluidity of contexts – contexts should not be assumed *a priori* but are an emergent feature of the interactions.

Context in LBS will be an important research area both in computer science and in GIScience. Current research has only just scratched the surface yet it is an important design aspect of engineering LBS applications.

# Chapter 8

## The Spatial Query

### 8.1 Introduction

---

Queries are fundamental to LBS in that they are the key to accessing information held in a database. No database will process its data into information without some form of instruction on which data to retrieve and how to present them in a view. The query establishes what is being asked of the database. However, simple as this may seem, this is not the whole picture. What can be asked of a database depends, of course, on what data are stored within it; but it also depends on the way the data are stored, that is the structure of the database. How quickly data can be retrieved also depends in part on how the data are indexed. So whilst from its title this chapter appears to narrowly focus on the query, it will be necessary to discuss a range of database issues not covered in the preceding chapters. Whilst the coverage in a book such as this will necessarily be selective, those wanting a more in-depth text on spatial databases are referred to Shekhar and Chawla (2003) as well as an earlier text by van Oosterom (1993).

Spatial queries are typically long transactions. By that we mean that running a query on spatial data to select relevant records and submit them to a view usually takes longer (often much longer) to transact than for those containing nonspatial data. Why this is so, is down to a number of issues. Firstly, spatial databases tend to be more complex than, say, business information systems, because there are three quite different types of data to be stored: geometric data, topological data and attribute data (in general, business information systems only have attribute data). Geometric data are what underlie the

graphic map visualizations of GIS as constructed using the four spatial primitives: point, line, polygon and cell/pixel (Section 3.4.4). Points, lines and polygons are used to depict discrete objects and are a *vector* representation. Cells or pixels are used to depict continuously varying features and are a *raster* or *field* representation. Topological data are necessary to describe the spatial relationships between the geometric data. This type of data makes queries about connectivity, adjacency and inclusion (such as: which roads connect with the M11 at junction 8? Answer: see Figure 4.3) much quicker to resolve in real-time than would otherwise be the case. Attribute data (Section 3.4.5) are all the additional data about the geometric features that describe what they are, what their function is and so on. Secondly, for any reasonably detailed, and therefore useful understanding of an area, large amounts of spatial data are required. As a consequence, spatial databases quickly become bulky (Table 5.2 gives some examples).

In any computer application that needs to read data from a store and perform some function on them, there are two main ‘costs’ that influence the time taken: CPU cost and input/output (I/O) cost. CPU costs are high where memory intensive operations are carried out on the data, such as performing calculations. I/O costs are high where data sets are large and not all the data can be held in memory at any one time, thus involving sequential read/write actions between memory and the hard disk. These costs are broadly summarized for different types of applications in Table 8.1. For applications using spatial databases both CPU and I/O costs are typically high, leading to longer transaction times in responding to queries. The implications for LBS should be clear. Long transaction times for spatial queries coupled with network costs in receiving the query and transmitting a response, to which must be added time costs associated with locating and contextualizing the user, may well mean that users run out of patience. As a rule, user patience tends to run out after about 10 seconds and this is taken as the upper limit, for example in designing Web usability

**Table 8.1** Broad CPU and I/O costs for a number of different applications (based on Shekhar and Chawla, 2003).

Application type	CPU cost	I/O cost
DBMS – Business Information System	Low	High
Spreadsheet, Statistical Package	High	Low
Spatial Database for GIS (and hence LBS)	High	High



(Nielson, 2000). For a high speed processor, 10 seconds is ample time to process millions of instructions, but for processing an LBS query it is not very long at all.

## 8.2 Geometric Data

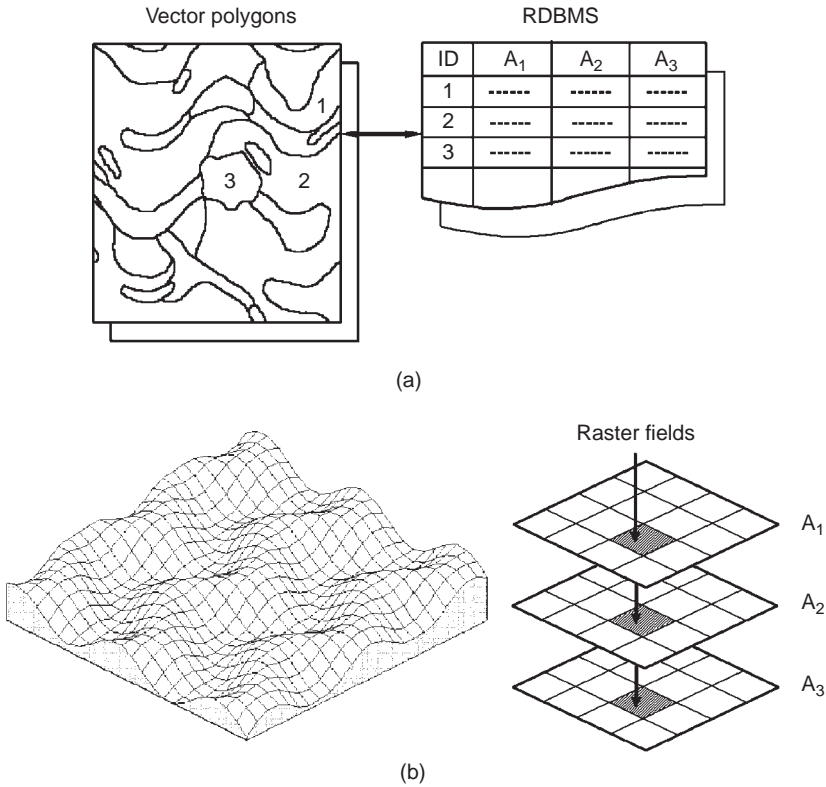
---

In Section 3.4.4 the data primitives that are the building blocks of the geometric representation of features in GIS were discussed. These were the point, line, polygon and cell or pixel. It was also noted that in practice these primitives tended to be organized into separate layers along thematic lines so that points representing, say, the location of post offices would not be mixed in the same layer with polygons representing, say, the location and extent of parks. There is also a fundamental difference between the vector primitives of point, line and polygon which are used to represent discrete objects, and the cell as a space-filling tessellation which is used to represent fields of continuously changing phenomena with few or no hard boundaries. This difference is fundamental to the way data are structured in a spatial database and therefore the way in which they can be queried.

Figure 8.1a shows conceptually that vector data are organized into a geometric component (in this example a map of polygons) coupled with a database component used for storing attribute data and managed using a relational database management system (RDBMS). Each vector object can have any number of attributes with the join between the geometric data and the RDBMS being a unique ID given to each vector object. By contrast, Figure 8.1b shows the general approach for cells organized as a matrix of raster data. Here there is no link to an RDBMS. Each attribute is treated as a field, represented as a raster layer that is distinct from the other layers. The cell geometry (position, orientation and size), however, must be identical for all layers of attributes, so that any cell in one layer can be unambiguously related to the equivalent cells in all the other layers.

### 8.2.1 Vector

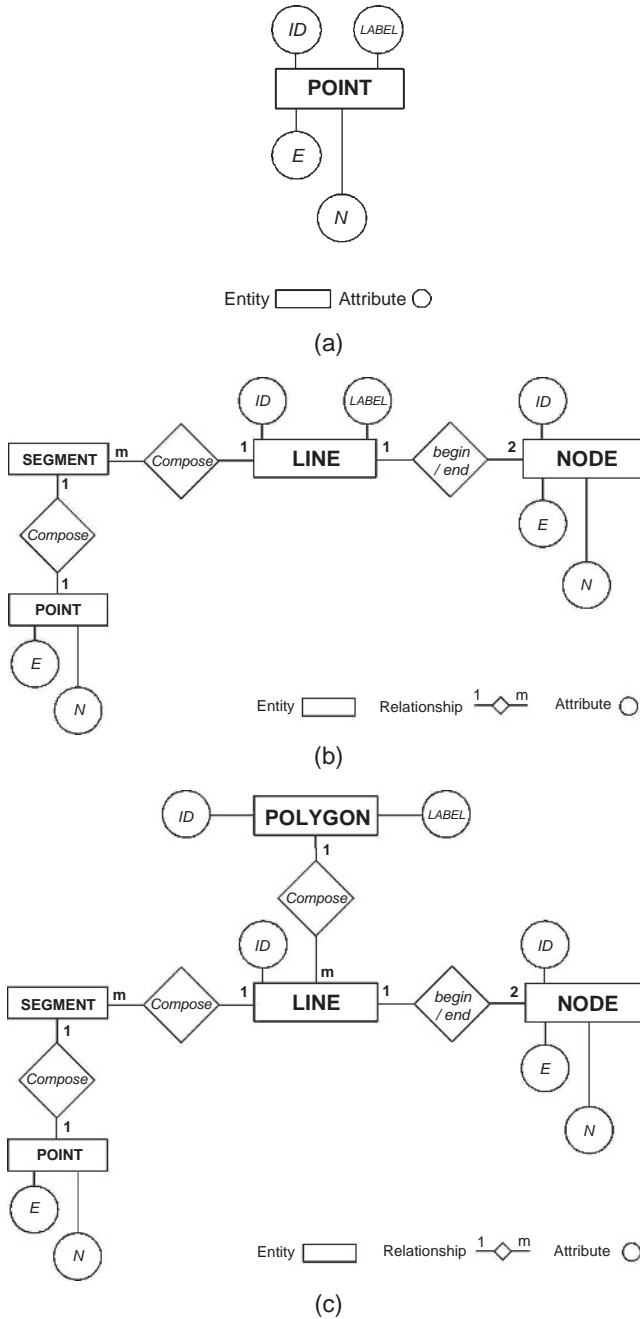
Vector geometry allows the mathematical properties of an object to be explicit. Such properties can be location, size and relationships between objects (Section 8.3). As regards the structure for encoding the



**Figure 8.1** Conceptual structure for organizing and storing: (a) vector data; (b) raster data (adapted from Brimicombe, 2003).

geometry of the point, line and polygon primitives, it is useful to use an entity-relationship model (Chen, 1976). In an entity-relationship model, the geometric components are the entities and the nature of the relationships between them is made explicit through the structure diagram (Figure 8.2). Key attributes of the entities can also be added to the diagram, which is then a blueprint for the database design.

From a purely geometric perspective, a point is a zero-dimensional entity as it has no extent (Figure 8.2a). Its location is described by the two attributes easting (E) and northing (N) in the relevant units (degrees longitude and latitude, metres from origin and so on). A point needs a unique key (ID) for joining with an RDBMS, where further attributes may be stored, and it can also optionally have a label (e.g. lamppost) to distinguish different classes of points in a layer – but this could equally well be stored in a RDBMS.



**Figure 8.2** Entity-relationship diagrams for (a) point, (b) line, (c) polygon (adapted from Brimicombe, 2003).

A line is a one-dimensional entity (Figure 8.2b); it has length but no breadth. Each line must begin and end at points; these are termed *nodes* to separate them from other intermediate points that determine the shape of the line between the two nodes. The geometry of a line is thus determined by the location of the beginning node, a series of *segments* joining successive points and ending at a node. A line therefore has an implicit direction from the beginning node to the end node and can, therefore, have a left and a right, properties that are exploited in encoding topology (Section 8.3). A straight line does not require any additional segments. Lines that share nodes and are therefore connected as a series of lines can form a *network*.

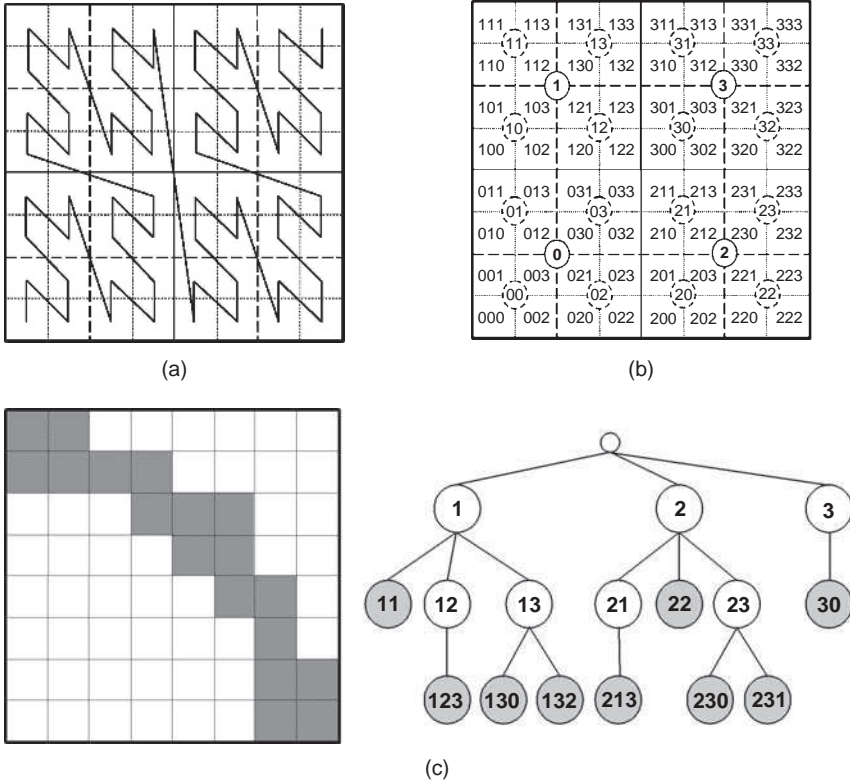
A polygon is a closed two-dimensional feature comprising one or more lines (Figure 8.2c). Where there is only one line, the beginning node and the end node must be at the same location. Where there is more than one line, all the beginning nodes must each have a shared location with an end node to form a chain that is connected back on itself without break.

This then is the basis on which point, line and polygon coverages are constructed as *shape files* in desktop GIS. As will be seen in Section 8.3, more complex structures are derived where spatial relationships between entities are made explicit within the data structure, ostensibly to add some in-built intelligence and speed up queries.

### 8.2.2 Raster

As noted in Figure 8.1, raster data are structured differently. In their raw form they are a matrix of attribute values, each value corresponding to a cell. This matrix can be simply stored row by row, but for large areas covered using small cell size this can amount to large matrices. For example, using 10 m by 10 m cells to map, say, noise levels across Greater London would require nearly 16 million cells. However, it is likely that across large swathes of residential London, away from main roads, ambient noise levels are going to be very similar, such that considerable numbers of neighbouring cells are going to have the same attribute value. It is clearly wasteful of disk space and, most importantly, I/O costs to store exactly the same attribute values maybe thousands of times. There are smarter ways of structuring raster data to reduce the I/O costs.

One family of data structures focuses on space filling curves and their derivatives (Figure 8.3). The space filling curve based on a



**Figure 8.3** Data structures for raster data: (a) Peano N-Scan; (b) Morton ordering; (c) an example of coding using quadtree-based Morton ordering.

recursive N-shape given in Figure 8.3a is known as a Peano scan after the original work by Peano (1890). This type of ordering allows the use of a *key* to find cell locations on the curve rather than a coordinate system. One such set of keys is Morton ordering (Morton, 1966) as shown in Figure 8.3b. Whilst following the Peano N-scan, it is also hierarchical based on quadrants. Thus the upper level – 0,1,2,3 – is the total area divided into four equal quadrants numbered following the N-scan. At the next level down there is a further four-way split so that quadrant 0 is subdivided into 00,01,02,03 following the N-scan and so on down to the lowest level of cell. This successive division by four from the total area down to the atomic cell (smallest cell in use) is called a *quadtree* (Samet, 1984). Thus Figure 8.3b is a quadtree with keys derived from Morton ordering with the underlying structure of a

Peano N-scan. This can now be applied to a simple example in Figure 8.3c where 18 (dark) cells have a binary value 1 and the remaining 46 cells have a binary value 0. To store this as a matrix would require 64 [0,1] values. Using the quadtree structure and Morton ordering, only the keys for the 18 dark cells need be stored as shown in Figure 8.3c. Where a higher order quadrant is completely filled by the same attribute, then the key for that quadrant is stored rather than the further sub-divisions within it. Thus in the upper left corner of the raster, the first four dark cells can be simply stored as key 11 and not as 110,111,112,113. This reduces I/O costs even further as the final structure only requires nine keys to be stored – a data reduction of 86%. Further advantages of this type of structure for raster data are discussed in Section 8.3 below.

### 8.2.3 Object-Oriented

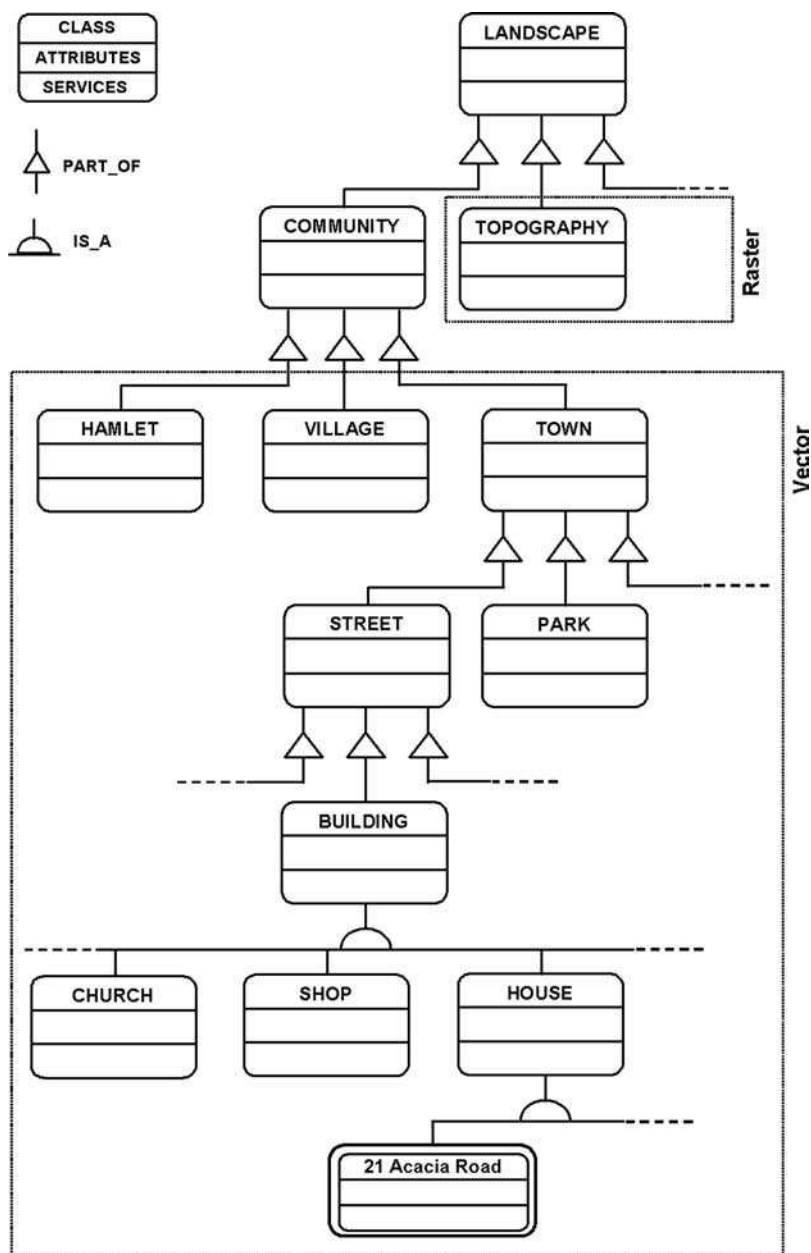
During the 1990s object orientation became a dominant paradigm in the computer sciences and an industry standard for software (e.g. C++, Java). It is not surprising, then, that object orientation has found its way into GIS and spatial databases (e.g. Worboys, 1994). Object orientation is not to be confused with object based, which is the use of vector primitives to geometrically describe discrete objects in the environment. Object-oriented (OO) analysis seeks to decompose a phenomenon into identifiable *classes* of objects and to explicitly relate them within a structured theme (Coad and Yourdan, 1991). Classes are used to represent groups of objects that have similar or shared characteristics. These shared characteristics are made explicit in the *attributes* and *services* of a class, where the attributes are the common descriptors and the services (or *methods*) are computer code for handling that class (e.g. how to visualize the class, say, in a map with appropriate symbology).

One distinct advantage of the OO approach for GIS is the ability to have classes that represent abstract concepts for which a concrete geographical representation is difficult (e.g. see Brimicombe and Yeung (1995) for an object-oriented view of Chinese *feng shui*!). It is possible to have a class `BUILDING` representing all those physically constructed objects above ground of a nontemporary nature that form enclosed structures which offer protection from the elements. Traditional object-based GIS have no difficulty in encoding the geographical footprint of any building as a vector

polygon along with the means to store attributes such as ownership, date of completion, height, type, main construction material, use of the building and so on. On the other hand, it is possible to have a class `COMMUNITY` representing a social construct around groups of people in their respective anthropomorphic environments, some of which may have urban characteristics (e.g. in towns) and some of which may have more rural characteristics (e.g. villages and hamlets). Attributes may concern degree of social cohesion, political outlook and so on. Neither the geographical extent nor the attributes are easily measurable; consequently nor are they so easily portrayed in map form. Traditional vector and/or raster GIS have not been good at including abstract, conceptual features that may not be spatially distinct or measurable, though one strand of the GIScience agenda has focussed on handling fuzzy and indeterminate objects (e.g. Burrough and Frank, 1996). In OO, however, it is perfectly feasible to include such classes of features within a data model. Figure 8.4 provides just such an example. This shows the decomposition of a landscape into its various components in the way that attributes are inherited from the uppermost class, `LANDSCAPE`, down to an instance of `HOUSE` – in this case a specific address. An attribute of `LANDSCAPE` to be inherited by all features within it might be the climatic type such as ‘humid temperate’. The class `TOPOGRAPHY` can have attributes and services for a raster representation, whilst various aspects of the built environment can have attributes and services for a vector representation. This leaves the classes `LANDSCAPE` and `COMMUNITY` with global attributes without requiring the need for a specific map representation.

Object orientation holds considerable potential for GIS and LBS, but there are very few truly OO GIS on the market. Oracle Spatial uses a hybrid Object-Relational Database Management System (ORDBMS) to ‘combine the best of both the relational and object-oriented databases’ (Sharma, 2002 p. 3). The advantage is that spatial object data types can be defined and stored, yet accessed and manipulated using spatial index methods developed for handling the spatial primitives of GIS. This allows the Structured Query Language (SQL) of relational databases to be extended to include spatial primitives (referred to as ‘geometry types’) and, therefore, to be able to perform operations in queries such as defining an area-of-interest or performing spatial join. Spatial indexing is discussed further in Section 8.5 and queries in Section 8.7.

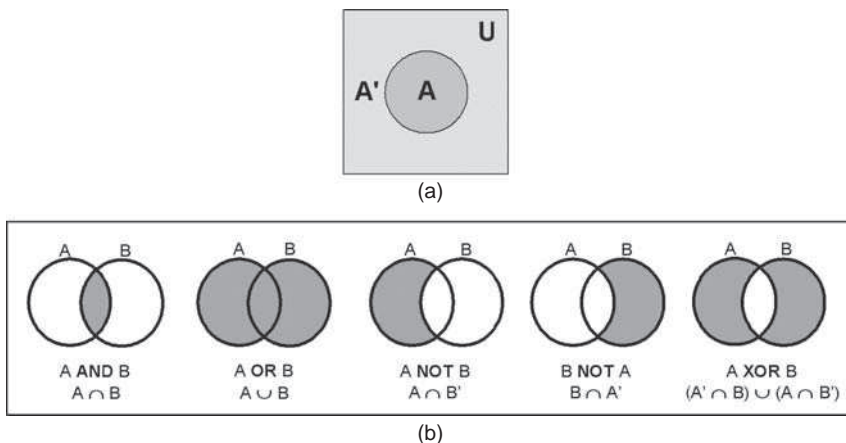




**Figure 8.4** Object-oriented modelling of geographical features (adapted from Brimicombe, 2003).

## 8.3 Topological Data

Topology is a basic discipline of pure mathematics that deals with different kinds of spaces (Simmons, 1963), whether they are Euclidean, spherical, manifolds or even Peano scans. The focus is not just on the static properties of these spaces but also on the rules governing their change of form. At the heart of topology are sets and the algebra of sets including Boolean logic. In Figure 8.5a is a universal set  $U$ , also known as the frame of reference. Within  $U$  is a subset  $A$ , usually denoted as  $A \subset U$ . The remaining part of the universal set that is not  $A$  is denoted by  $A'$  such that  $A$  together with  $A'$  form  $U$ . In Figure 8.5b the situation is extended to having two subsets  $A, B$ . Illustrated are some basic functions and their notation relating to subsets  $A, B$ . The first is the *intersection* of subsets  $A, B$  denoted as  $A \cap B$ , also denoted as  $A$  AND  $B$ . The second is the *union* of subsets  $A, B$  denoted as  $A \cup B$ , also denoted as  $A$  OR  $B$ . From this follows two fairly obvious *difference* functions:  $A$  NOT  $B$  ( $A \cap B'$  or  $A \setminus B$ ) and  $B$  NOT  $A$  ( $B \cap A'$ ). The last function illustrated is  $A$  XOR  $B$  (where XOR is 'exclusively OR'), which by the algebraic expression  $(A' \cap B) \cup (A \cap B')$  gives the union without the intersection of subsets  $A, B$ ; this is referred to in some texts as *symmetric difference*,  $A \Delta B$ . There are some functions such as selecting only  $A$ , selecting only  $B$ , selecting neither of them, which it is not necessary to illustrate here. In fact there are  $2^{(2^n)}$  Boolean



**Figure 8.5** Introduction to sets: (a) the universal set  $U$  and subset  $A$ ; (b) basic Boolean functions for two subsets  $A, B$ .

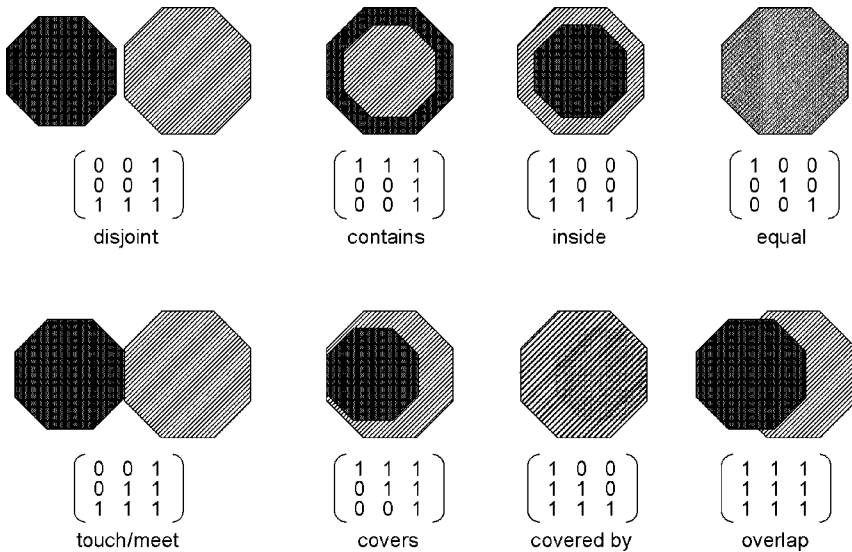
functions, where  $n$  is the number of subsets (Harris and Stocker, 1998). Thus for two subsets  $A, B$  there are 16 Boolean functions that can be derived from the operators AND, OR, NOT.

These principals can be extended to the possible relationships of two polygons  $A, B$  in a plane, where the plane is commonly denoted by  $\mathbb{R}^2$ . For any polygon  $A$ , we can consider its interior  $A^\circ$ , its boundary  $\partial A$  and its exterior  $A'$ . For two polygons  $(A, B)$ , their three parts (interior, boundary, exterior) can be used to construct what is known as the *nine-intersection matrix* (Egenhofer and Herring, 1990) as given in Equation 8.1. Using a binary notation  $[0,1]$  where 0 denotes empty and 1 denotes nonempty, this matrix allows the realization of eight fundamental topological relations in  $\mathbb{R}^2$ . These are: *disjoint*, *contains*, *inside*, *equal*, *meet* (touch), *covers*, *covered by* and *overlap*, as illustrated in Figure 8.6. These topological relations can be extended to combinations of points, lines and polygons.

$$\Gamma_9(A, B) = \begin{pmatrix} A^\circ \cap B^\circ & A^\circ \cap \partial B & A^\circ \cap B' \\ \partial A \cap B^\circ & \partial A \cap \partial B & \partial A \cap B' \\ A' \cap B^\circ & A' \cap \partial B & A' \cap B' \end{pmatrix} \quad (8.1)$$

An understanding of these and other spatial-analytic relations (see below) is fundamental to spatial queries. Topology-based queries typically have high CPU costs and I/O costs as they require extensive use of coordinate geometry to establish object intersections. Two approaches of speeding this up are to pre-compute some of the topological relations and store them within the data structure, and/or to use spatial indexing. The latter is discussed in Section 8.5. In Figure 8.6, the only two topological relations possible within a single, conflated object-based layer are disjoint and touch/meet. This is because within a conflated GIS coverage, features cannot overlap or cover each other: one building cannot be covered or overlapped by another, road sections in a network meet exactly at shared nodes. Nevertheless, the fact that two polygons or two lines in a network meet is an important topological relation used in many types of queries, especially, for example, in route finding and calculating travel times (see Figures 3.3 and 3.4).

A topological data structure for lines and polygons explicitly records connectivity (as in a network) and adjacency of polygons where they touch. An example of such a data structure is given in Figure 8.7, in which there are four relational tables for nodes, lines, segments and polygons, and reflects the entity-relationship diagrams in Figure 8.2. The central table is the 'line list' which points to the start



**Figure 8.6** Eight fundamental topological relations in  $\mathbb{R}^2$  derived from the nine-intersection matrix (based on Egenhofer and Herring, 1990).

and end nodes in the ‘node list’. Because a line has a direction from start to end node (with or without intervening segments), pointers to the polygons to left and right of the line are also stored in the ‘line list’. The ‘segment list’ stores the point coordinates that form segments and the ‘polygon list’ stores pointers to the lines that form its boundary. In this way polygon adjacencies are quickly identified, as is connectivity between lines. Although the example of a topological data structure has been illustrated using polygons, it is straightforward from Figure 8.7 to see how, in the absence of polygons, the structure can be simplified to represent the connectivity of lines that form, for example, a road network. The data structure then very much simplifies routing type queries (Section 3.4.6 and Section 8.7).

With raster data, topological relations are established in a different way. The space-filling structure of square cells means that neighbouring cells either share a side or touch at the corner. This leads to a consideration of three-by-three cell neighbourhood with the cell of interest at the middle (Figure 8.8). Topology can be implicit as shown in Figure 8.8a through displacement by one row ( $i$ ) and/or one column ( $j$ ). The topology can also be made explicit, for example through Morton ordering. Consider the cell in Figure 8.8b (from an

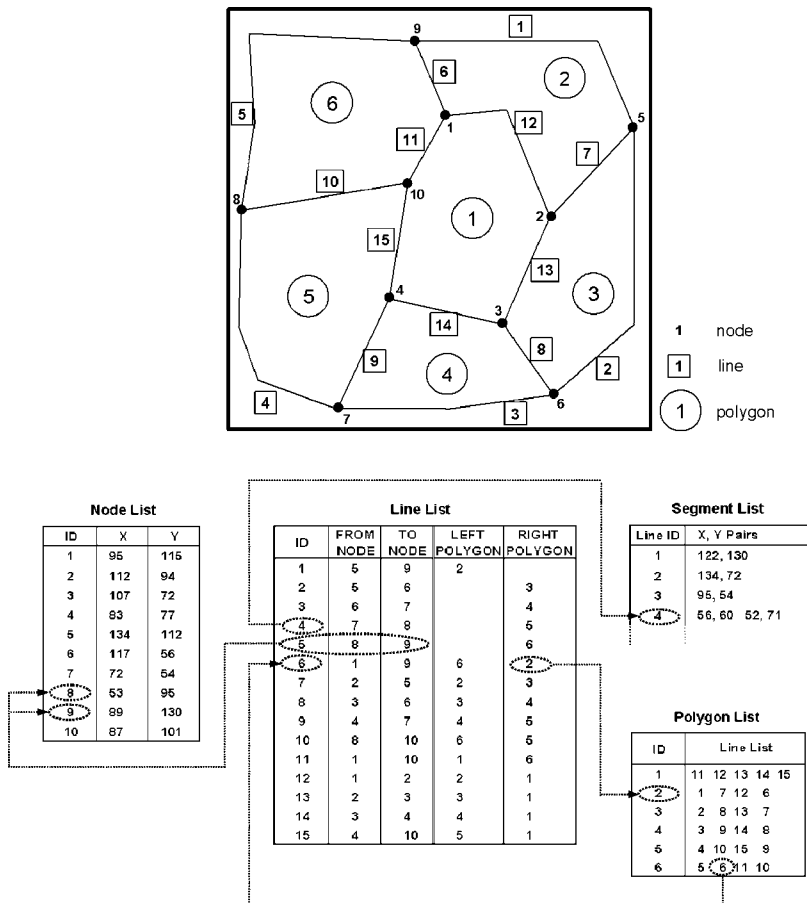
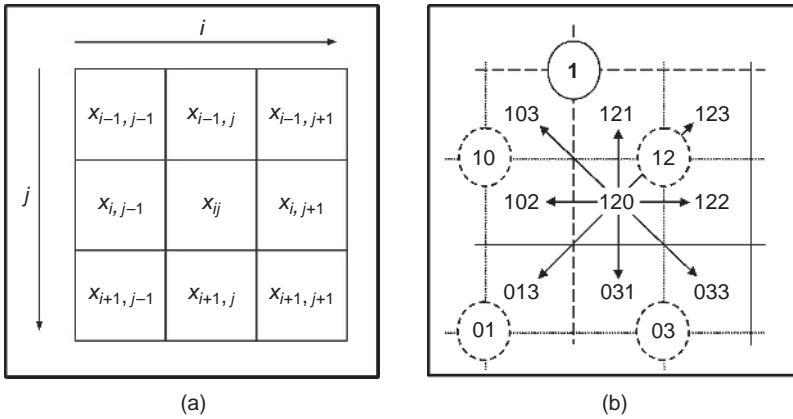


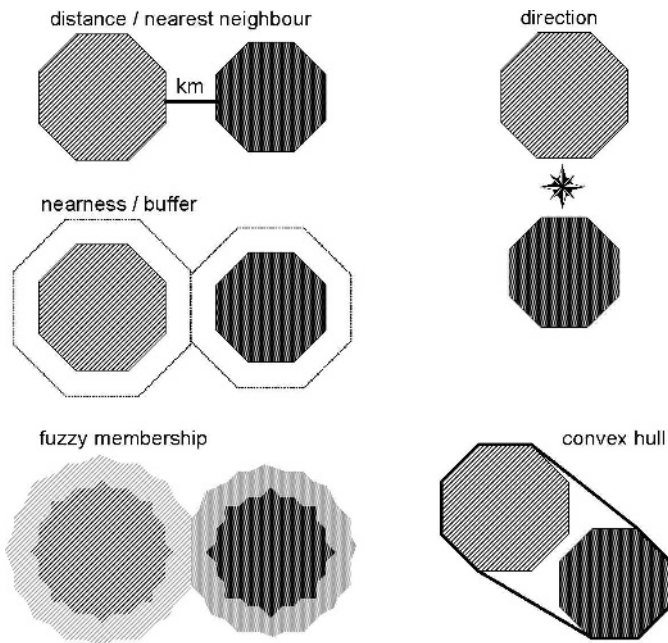
Figure 8.7 Building topology into the vector data structure (from Brimicombe, 2003).

enlargement of Figure 8.3b) with the key 120. Since the key ends in 0, all eight neighbours will have keys ending in 1 (above and below the cell), 2 (to right and left) or 3 (on the diagonals) and not 0. This topological relationship of the ordering holds for higher levels of the quadtree (e.g. 01 is below 10, 12 is the right of 10, 03 on the diagonal to 10). Scans, such as the Peano N-scan, are topological spaces in themselves.

A further set of spatial relations (Figure 8.9) might best be termed spatial-analytic as they wholly or partly rely on calculations of distance and/or direction in  $\mathbb{R}^2$  rather than exclusively relying on the invariant intersection of geometry. These are again fundamental to the types of queries that arise in LBS. The *nearest neighbour* relationship is



**Figure 8.8** Topology in raster data: (a) implicit; (b) explicit in the Morton ordering.



**Figure 8.9** Spatial-analytic relations between objects.

a distance measure that seeks to find an object where the distance is a minimum compared with other similar objects. Queries, or responses to queries, may take a *directional* form as in, for example, a destination being south of London. Sometimes topological relationships, such as

meet or overlap, can be performed for transformed objects, such as through a buffer. This establishes a *nearness* property. Similar to this are relations based on *fuzzy membership*, where objects such as polygons may have vague boundaries resulting not in binary intersection but in a degree of belief that an intersection occurs. Lastly, the *convex hull* relation can be used to establish a convex polygon of minimum area completely enclosing a set of points, lines or polygons. An important property of the convex hull is that it can be used to find the centre of gravity in  $\mathbb{R}^2$  of the set (de Smith *et al.*, 2007).

### 8.4 Attribute Data

---

In Section 3.4.5 the introduction was made to attribute data as providing an opportunity for attaching nongeometric (i.e. nonspatial) data to spatial objects. In the more straightforward desktop GIS, as illustrated in Figure 8.1a, attribute data are stored separately in either flat files or single table database files that can be indexed to improve searching. In enterprise level GIS these will commonly be stored in relational databases as a series of tables having a relational schema. Almost any type of data can be stored. Traditionally, RDBMS were limited to numbers (integer, float), text and dates but advances in interoperability mean that data fields can now include images, sound clips and hyperlinks to, for example, external documents, spreadsheets and Web URLs. All of these then can be linked to the spatial location and geometry of an object using GIS and queried. The design and implementation RDBMS schemas are not covered here and the reader is referred to texts such as Date (1990). However, it should be noted that whilst attribute only databases are successively normalized to improve the overall simplicity of the schema and reduce the level of redundancy in the relational tables, this is often not carried out in GIS to its fullest extent as joining numbers of large tables can substantially increase both CPU and I/O costs. GIS attributes are by and large stored in non-normalized database schema.

### 8.5 Indexing Spatial Databases

---

Because GIS databases tend to be very large with high CPU and I/O costs, queries need to be speeded up. One way, which has been seen above, is to pre-compute and store certain topological data within the



database. Another approach is to index relevant fields so as to reduce the search time to locate specific records within the database. This avoids the laborious approach of searching for records sequentially from the first record in the database through to the last record. Instead, indexing allows the search to move quickly to the approximate region of the database where the actual record being sought is going to be found.

There are similarities and yet differences in the way attribute and spatial (geometric) data are indexed. One standard approach to indexing attribute data in commercial RDBMS is the Balanced Tree or *B-tree*. This is illustrated in Figure 8.10, which lists the 32 Boroughs of Greater London. If the goal is to search for a particular Borough without there being any indexing, it would be necessary to sort the data alphabetically and search sequentially down the list. Of course, if 'Barking and Dagenham' was being looked for it would be found immediately (at the first test), but for 'Westminster' it would be necessary to search to the bottom of the list and find it at the thirty-second test. For a large number of searches involving all the Boroughs, it can be expected that the average number of tests that needed to be carried out would be 16.

The list of Boroughs with a B-tree index is shown in Figure 8.10a. The indexing has two layers: first the list is split in half at 'Hil' (for 'Hillingdon') and each half subsequently split into four. Figure 8.11b shows how an indexed search for, say, 'Havering' progresses. The first test is whether 'Havering' comes before or after 'Hil' alphabetically. If before 'Hil', it moves to test against 'Bre' in the second level and so on down to 'Hil' again, which causes the search to go to the lowest level and tests the actual Borough names in turn to find 'Havering'. This has required eight tests instead of 15 tests in an un-aided sequential search. The minimum number of tests using the indexed approach is three (compared to one for the sequential) but the maximum is only nine (compared to 32). For a large number of searches of all Boroughs, the average number of tests would be six compared to 16 for the sequential search. Indexing can thus dramatically reduce both CPU and I/O costs. The example given should be taken as illustrative of the principle, as in reality text or alphanumeric data would be given an integer key and thus indexed using numbers, thereby further reducing the CPU cost of each test.

The B-tree is one dimensional and therefore not best suited to indexing two-dimensional spatial data. For raster data, the use of space-filling curves such as the Peano scan in Figure 8.3a forms the basis for indexing as a spatially hierarchical tree as illustrated in Figure 8.3b. For vector data a number of indexing schemes have been devised

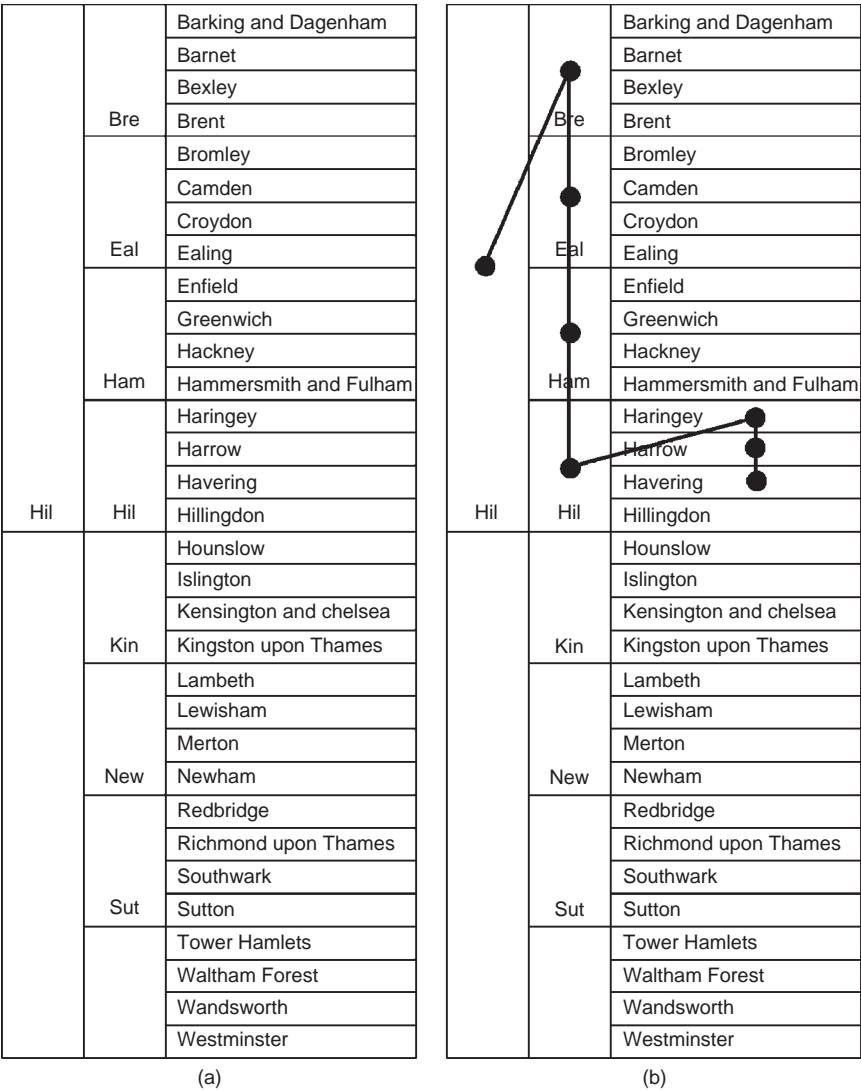
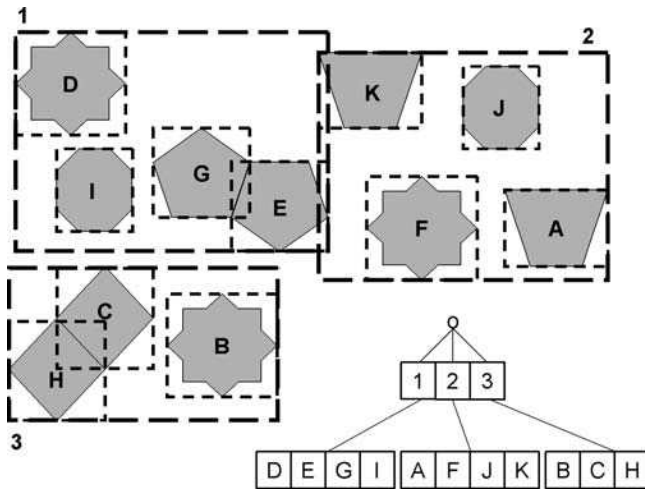


Figure 8.10 Illustration of a B-tree for searching a list of London Boroughs.

(see, for example, van Oosterom, 1993) but the basis of these approaches can be best illustrated using the Region Tree or *R-tree*. In Figure 8.11 there are a number of polygon features labelled from A through to K. Because each feature can have a complex shape, the CPU and I/O costs are reduced by taking only the minimum bounding rectangle (MBR) of each polygon. The MBRs are then indexed in a



**Figure 8.11** Illustration of an R-tree using minimum bounding rectangles for indexing vector data.

spatial hierarchy of higher order MBRs that can either contain roughly the same number of polygons or are partitioned according to the size of data set that the polygons represent. The R-tree allows fast searching of geometric features for attribute selection, carrying out overlay and for fast refresh when zooming or panning the display. Thus a search to find all polygons that fall within an area of interest  $\mathcal{R}^2$  starts by testing the MBRs at the highest level of the R-tree hierarchy with the extent of  $\mathcal{R}^2$ . If there is an intersection then the search moves down to the next level of the R-tree until at the atomic MBR (i.e. at the finest resolution in the hierarchy), individual polygons can be selected on the basis of their MBR. The true shape of each polygon selected in this way can then, if necessary, be intersected with  $\mathcal{R}^2$  as a final test.

## 8.6 Issues of Data Temporality

GIS have historically been designed to store and analyse static representations of landscape and the built environment. Of course, very few phenomena that have spatial processes are truly static and GIS have had an important role in the study of both social and environmental change. Databases may be edited as feature changes are detected, with an inevitable time lag (Section 5.3.5), usually resulting in the outdated information being overwritten. Important snapshot data sets that are

re-collected in their entirety (such as censuses or land use mapping) tend to happen cyclically (sometimes decades apart) and are usually kept as stand-alone data sets. Very often these data sets are incompatible with changes in the spatial units used to report the data and/or changes in the classification of attributes used. Thus, for Greater London, 15 366 Enumeration Districts were used to report the 1991 population census and 24 140 Output Areas were used to report the 2001 census, which, together with changes to some of the questions asked in the questionnaire, makes analysis of social change more challenging. Again, national land use surveys of the United Kingdom carried out using satellite imagery in 1990 and 2000 have different land use classes for the mapping, thus complicating any analysis of land use change.

For LBS, however, the purpose is not to analyse medium to longer term social or landscape change; they have a more immediate concern with data dynamics. To be sure, the base data that LBS providers draw upon will need to be up-to-date with sufficient temporal and spatial granularity. Thus as changes occur to road networks, for example, so revisions will need to be inserted into the database. LBS significantly differ from GIS in that some attribute data (e.g. traffic conditions) are highly dynamic, if not continuously changing, whilst users and/or targets (Section 4.6) may themselves be mobile, that is continuously changing their spatial position over a period. Consequently, in Section 5.2 the need to distinguish three broad sets of database objects was introduced: *static* (short to medium term invariant), *dynamic* (short term with periodical updating) and *mobile* (spatially moving either very short term intermittently or continuously and thus requiring almost continual updating).

The literature on temporal GIS and time geography is steadily growing. The reader is referred to, for example Langran (1992), Peuquet (1994, 1999), Wachowicz (1999), Miller (2005) and Yuan (2008) for a fuller treatment of the subject. Only those issues most pertinent to LBS are considered here.

### 8.6.1 Space Versus Time Dominant Views

Concepts of space and time are inherently related – both emerged from the Big Bang event that created our universe and are integral aspects of gravitational fields. Both are closely interwoven as part of a four-dimensional continuum in which space and time are interchangeable over short durations. Yet conventionally we separate three-dimensional

space from one-dimensional time and tend not to treat them on equal terms. The GIS view of the world is that absolute space exists independently of any objects or fields it may contain. Firstly, an unbounded thematic layer is created and given a coordinate system and projection. Only then is it populated with objects (as vector) or discretized into a field (as raster). In this way the absolute space thus created is objective and provides a rigid, geometric structure in which objects can exist and change (Peuquet, 1994). Layers are treated as snapshots in time with updates taking place periodically. This puts the emphasis on versioning as the means of analytically handling time, especially when studying change over time. Recently, a more timely and cost-effective approach has been adopted of providing update patches in which relevant features and attributes are patched (overwritten) into the existing layer with an attribute time-stamp provided. This means that layers are no longer a single snapshot but a snapshot for a period from the oldest to the newest time-stamp attribute. The main characteristics of the space-dominant view are summarized in Table 8.2.

In a time-dominant view there is no explicit representation of space (though it may be implied by the naming of geographical locations but without the ability to map either their absolute or relative positions). Time is explicit and recorded for all data either in absolute terms (date- and time-stamp) or in relative terms (time elapsed from an arbitrary start point). As with space, time also needs to be discretized and marked out in some measured interval against which to record data and can range from a coarse granularity (year, month) to a very fine granularity (milliseconds) for near continuous recording. Also, as

**Table 8.2** Main characteristics of space- and time-dominant views (based on Wachowicz, 1999).

Space-dominant view	Time-dominant view
space is a container	time is viewed as a line
objects and fields must be associated with a layer	events and observations are linked to a time line
applied primarily to the mapping of geographical features	applied to geological and environmental sciences
layer-based object and tessellation models	interval and event models
each layer tends to be an instance in time	absolute space is not represented
periodic updating	event-based updating
analysis based on layer geometries	analysis based on lineage and time series

with changes in spatial resolution, the finer the temporal granularity, so the volume of data to be stored will increase. Data for an entity when sorted by time-stamp becomes *time-series* data and there are RBMS specifically adapted to storing such data (e.g. hydrological databases). An entity must come into existence and has *life* until its existence comes to an end. During its life an entity may change *state*, possibly many times. This change of state may occur *continuously* throughout some time interval, be *majorative* for most of some time interval, be *sporadic* from time to time (*cyclic* if regularly oscillating), or *unique* (happening only once). The point or period of time when a change in state(s) occurs can be referred to either as an *episode* for a series of state changes or as an *event* for a single change in state.

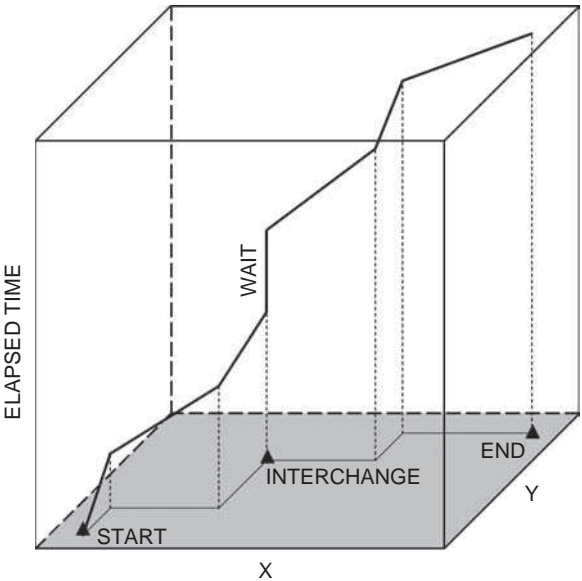
Bringing these two views together into a space–time view, objects and fields can have spatial existence in which both location and time are recorded. Objects and fields can have changes in state, either recorded through their attributes (e.g. change of ownership of a land parcel) or through their geometry (e.g. changes to the boundary of a land parcel). Importantly, though, a space–time entity can have *motion*, that is over time it changes its geographical location. Thus a weather front can be changing both its state (becoming more intensive, strengthening wind, increasing rainfall intensity) and its location (moving across central England) over time. Motion can be viewed in two different ways: Lagrangian or Eulerian. The Lagrangian view is space-dominant and, therefore, is the means of preference in GIS; the Eulerian view is time-dominant and, therefore, of preference in, say, environmental simulation. Both Euler (1707–1783) and Lagrange (1736–1813) were mathematicians. A Lagrangian view of motion focuses on the object that is moving, such as tracking a car being driven across a city. A Eulerian view of motion focuses on a fixed portion of space through or across which some motion is taking place. This is much like standing on the side of a road and watching the vehicles pass through a fixed field of view. Both are relevant to LBS. Whilst there is obvious interest in tracking individual users and targets (Section 7.7), data are also collected based on frequency of movement across sensors (Section 5.3.4) of the type that are now on many primary roads and motorways in order to automatically monitor the rate of traffic flow. The space–time view for LBS is made the more complex due to, for example, the need to know changes in state of individuals or objects whose motions are being tracked, changes in the state of the network upon which the motion is taking place and more general changes in the context about and for which the motion is taking place.

### 8.6.2 Space–Time Paths and Topology

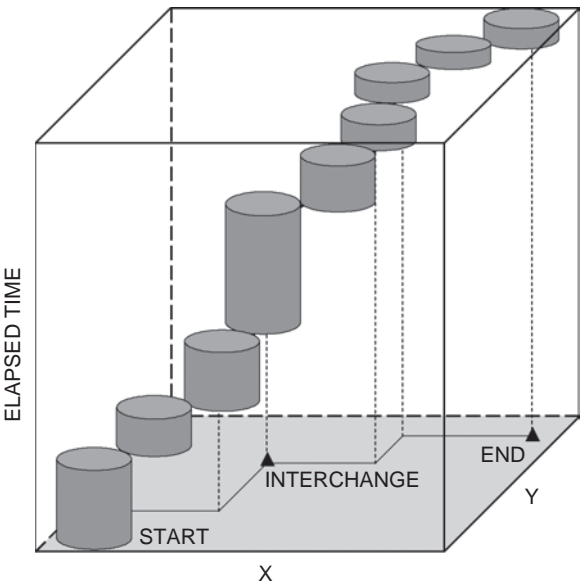
The concept of a *space–time path* has been central to time geography since the 1970s. Basically, a space–time path graphs the trajectory of an individual (e.g. human, moose) or object (e.g. car) as it moves across a bounded region of geographical space over time. A space–time path using continuous measurement on all axes (such as might be achieved using a GPS) to record a trip in which an individual travels across town but has to wait at a transport interchange for a connection is illustrated in Figure 8.12. If such a journey for many individuals were to be recorded, convergence of many space–time paths would be found at the transport interchange, some of which would be co-located in space–time (being in the same place at the same time) and some of which will only be co-located in space (coming to the same place but at different times). The track shown in Figure 8.12 is idealistic as it assumes continuous  $x, y, t$  measurements are recorded. From an LBS perspective the amount of data storage required is likely to be problematic and the amount of bandwidth used to obtain such data is likely to adversely affect the amount available for voice/data services. On the other hand, if only Cell-ID for a network is known, then the space–time path for the same journey would look much like Figure 8.13, which gives the effective cell service area and the residence time in each cell. This is a much coarser space–time resolution and can lead to some ambiguity as to exactly where the individual is. For example, suppose that at the transport interchange the individual knew how long the wait was to be and decided to go off in search of a better cup of coffee than on offer at the interchange, and yet happened to stay within the same network cell. This more intricate aspect of the individual's *activity space* is not recorded. It may not seem important but if, say, in a mobile commerce application (Section 4.7 and Chapter 10) a discount voucher is to be pushed to any individual whose activity space falls within a certain distance of an outlet, the finer granularity may be important.

Wherever there is coarser resolution or intermittent tracking of an individual then it is usual to model the *potential activity space* through the use of *space–time prisms*. Between any two known points ( $x_1, y_1, t_1$  and  $x_2, y_2, t_2$ ) the space–time prism is the geographical area that an individual has the potential to visit given a maximum travel velocity. This prism is fixed at the known points and extends to a maximum mid way between  $t_1$  and  $t_2$ . Within our Cell-ID scenario (Figure 8.13) the fixed points (within the resolution afforded) are the

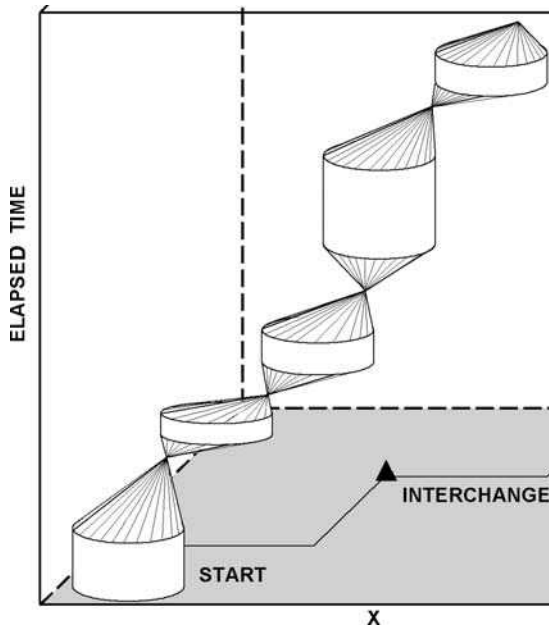




**Figure 8.12** Illustration of a space-time path with continuous measurement on all axes.



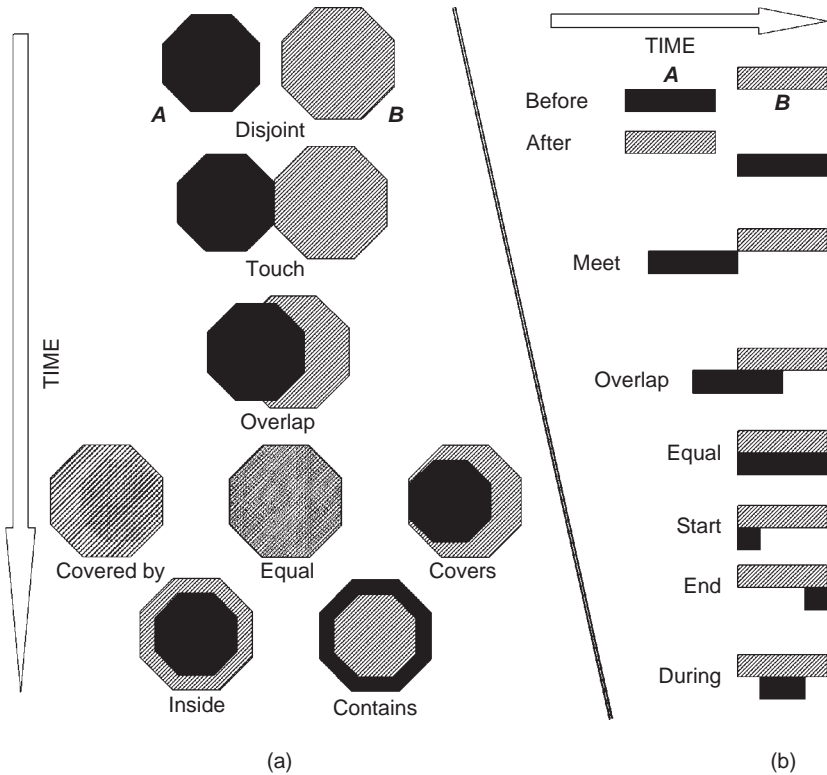
**Figure 8.13** Same space-time path as in Figure 8.12 as might be measured using Cell-ID.



**Figure 8.14** Potential space–time activity spaces based on Cell-ID in the early stages of the trip in Figure 8.13.

handovers between one Cell-ID and another. Once in the cell, the prism extends but is bounded by the limits of the cell (otherwise another cell handover would have occurred). This leads to a series of space–time prisms as illustrated in Figure 8.14 and defines the potential activity space for the trip.

The concept of spatial topology (Section 8.3) is made more complex with the added dimension of time. The nine-point intersection model of topology (Figure 8.6) can be ordered to reflect a temporal motion in the relationship between entities. This is shown in Figure 8.15a. Strictly from a temporal view (without space), topology expresses the relationship of entities in strict temporal order and is also linked to events. A scheme based on Allen (1984) is given in Figure 8.15b; these have an approximate mapping with the nine-point intersection model in terms of their Boolean definition. This still leaves some relationships that would be important to LBS undefined. For example, two objects ( $A$ ,  $B$ ) may be spatially disjoint, but because one or both may be in motion, depending on their relative direction these two objects may be converging towards the same



**Figure 8.15** Spatial and temporal topology: (a) temporal arrangement of the nine-point intersection model of topology (based on Wachowicz, 1999); (b) temporal relationships of objects (based on Allen, 1984).

location in space and time, diverging (disjoint, distance increasing) or conceivably even moving in parallel. This need to define additional space–time relationships is returned to in Section 8.7.4.

### 8.6.3 Database Implications for LBS

The adding of a temporal dimension to spatial databases provides added complications, not only from the perspective of data volume to be managed but also how the data need to be structured, how they might be queried and the response times to queries. The use of time alongside space as equally important dimensions, as is necessary for LBS, is a major departure from traditional GIS applications. Of greatest difficulty to capture and integrate within a database schema will be

mobile data types that change their position with time. For simplicity, most objects being tracked (individuals, vehicles) can be treated as points requiring only the update of a single coordinate pair, thus avoiding the difficulty of having polygons that change location and/or shape. Changes only to attributes that might occur with objects whose states are dynamic (i.e. requiring short term or periodical updates) but don't necessarily change their geographical position (e.g. change in opening hours of a shop) can be handled in standard database ways. There are two broad approaches to recording time in databases: *transaction* time and *valid* time. Transaction time records the instance of an event, when something occurs and the database is updated. Valid time is the period over which some object or its state exists. For example, making a purchase on-line causes an entry to be made into the vendor's database and the date and time of the transaction will be recorded against it; if an item has been discounted during a sale, then the database will have a record of the discounted price and the period for which it is valid.

Time may be recorded either in absolute terms or relatively from some fixed point. Thus a transaction may be recorded in absolute time, but the end of a valid period may be relative to an absolute date which marks the start of the period of validity. Date/time-stamping appears to be the most often used method in dealing with the time dimension. In all RDBMS, dates are a specific data type, usually combining date and time in the general form *31/12/2007 12:30:15* and requiring eight bytes of storage (equivalent storage requirement for an *x,y* coordinate pair). RDBMS functions to provide a day/time-stamp against a new or modified record include: `SYSDATE()` in Oracle, `GETDATE()` in Microsoft SQL Server and `NOW()` in Microsoft Access and MySQL. In some RDBMS, there is a choice of date/time data types: `DATETIME` and `TIMESTAMP`. The former has the same general form as given above whilst the latter in MySQL takes the general form *20071231123015* which has a smaller storage requirement (four bytes). Importantly for LBS, the `TIMESTAMP` associated with a record is automatically updated whenever the record is altered. From general experience we have found that storing date/time in the latter general form is more stable when passing date/time data between applications where there may not be support for a date/time data type or there are subtle variations in how this data type is handled in different RDBMS and spreadsheets. Most RDBMS also have functions for manipulating date/time (sometimes referred to as *date arithmetic*), such as calculating the difference between two dates (to give a duration), identifying

most recent record and so on. Storing dates as text in the form ‘31 Dec 2007’ (or similar) can be problematic and is not recommended, firstly because the entries will not sort properly in order of date and, secondly, because the available functions for manipulating dates will not work on this type of text entry. For strict comparison of date/time records it is necessary to ensure they are to the same time zone (particularly when a service user is roaming) and follow the same calendar (e.g. Gregorian, Hegira).

If the history of an object or its states is of little or no interest, then each time a change occurs, the new location or state can simply be overwritten in the database and date/time-stamped to know its currency. This may be alright for certain classes of objects. For example, if a petrol station were to have its fuel prices changed, then from a LBS perspective only the current prices are of any interest. However, if it is desirable to track either the movement of an object through space or the changing states of an object because this can itself generate useful information, then some form of *version management* becomes necessary. There are two broad approaches to versioning. One is to create objects that have date/time-stamped attributes (including location) so that in effect the new state is represented by a new set of attributes. An object’s attributes can then be sorted to give an historical order. The other approach is to create a new object with date/time-stamp each time a change to the attributes is necessary with a pointer to its predecessor object. In this way it is possible to temporally chain objects into an historical sequence.

Another database implication of using space-time for LBS is that where topological relationships are stored explicitly within a database (Figure 8.7) or where there is spatial indexing (Figure 8.11), any time a change in the location or geometry of an object is recorded the topological data and the indexing need to be re-calculated and stored again with high CPU costs. This would also require that any new queries be locked out for the duration of the update.

## 8.7 Spatial Queries

---

If data are to be organized and stored in a database, then there must not only be a means by which the structure (system of related tables) is set up and populated with data, but also a means by which it can be queried in order to extract information. The most widely used

language for interacting with databases is the *structured query language* (SQL) originally developed by IBM. SQL is a *declarative* language rather than a *procedural* language (such as C++ or Visual Basic) in that the user specifies the required end result and not the means (procedure) of achieving it. Underlying such a language is, necessarily, a *formal* language which provides the theoretical grounding. For query languages this is *relational algebra* (RA). In this section the formalisms of RA are outlined first, and it is then seen how these are implemented in SQL. This is followed by a discussion of extended SQL for spatial objects as implemented in MapInfo and Oracle Spatial. Other aspects of spatial queries that need to be considered are: optimization to CPU and I/O costs; and queries that require associated computational algorithms, such as the routing queries illustrated in Section 3.2.

### 8.7.1 Relational Algebra

Algebra is the branch of mathematics for finding solutions to equations and inequalities which contain only basic operations (addition, subtraction, multiply, divide, raising to powers). Relational algebra is similarly structured but is applied to database tables (relations). There is a single class of operand ( $\Omega_a$ ): the table; and eight operations ( $\Omega_o$ ): *select*, *project*, *join*, *intersection*, *union*, *difference*, *symmetric difference* and *cross-product* (also known as *Cartesian product*). The basic axiom of RA is that the result of an operation on an operand must be an operand. In other words, any manipulation of a table must result in a table. The operations *select* and *project* are applied to a single table to retrieve a subset of records (rows) and variables (columns) respectively. The *select* operation ( $\sigma$ ) requires the use of operators ( $=, <, >, \leq, \geq, \neq$ ) in order to define subsets, whilst the project operation ( $\pi$ ) merely specifies the relevant variables that are required. The *join* operator ( $A \otimes B$ ) allows *select* and *project* to be applied across more than one table where those tables can be related using primary and secondary keys (see the example in Section 3.4.5). The other five operations are set operations applied to two or more tables. These are illustrated in Figure 8.5b, except for the *cross-product* ( $A \times B$ ), which is the set of all ordered combinations of table *A* and table *B*. These set operations on tables are illustrated in Table 8.3. The operations *select* and *project* can be applied to tables both before and after *join* and other set operations in order to derive the desired result to a query.

**Table 8.3** Basic operations on tables: (a) join ( $A \otimes B$ ), select ( $\sigma$ ), project ( $\pi$ ); (b) intersection ( $A \cap B$ ), union ( $A \cup B$ ), difference ( $A \setminus B$ ) and symmetric difference ( $A \Delta B$ ); (c) cross-product ( $A \times B$ ).

A.Key	A.Var1
1	11
2	12
3	13
4	14

Table A

C.Key	A.Var1	B.Var1	B.Var2
1	11	111	211
2	12	112	212
3	13	113	213
4	14	114	214

$A \Join_{(Key)} B_{(Key)}$

B.Key	B.Var1	B.Var2
1	111	211
2	112	212
3	113	213
4	114	214

Table B

B.Key	B.Var1	B.Var2
3	113	213
4	114	214

$\sigma_{(Var1 > 112)} B$

B.Key	B.Var2
1	211
2	212
3	213
4	214

$\pi_{(Key, Var2)} B$

(a)

<table><tr><th>A.Var</th></tr><tr><td>111</td></tr><tr><td>112</td></tr><tr><td>113</td></tr><tr><td>114</td></tr></table> <p>Table A</p>	A.Var	111	112	113	114	<table><tr><th>B.Var</th></tr><tr><td>111</td></tr><tr><td>113</td></tr><tr><td>115</td></tr><tr><td>116</td></tr></table> <p>Table B</p>	B.Var	111	113	115	116	<table><tr><th>C.Var</th></tr><tr><td>111</td></tr><tr><td>113</td></tr></table> <p><math>A \cap B</math></p>	C.Var	111	113	<table><tr><th>C.Var</th></tr><tr><td>111</td></tr><tr><td>112</td></tr><tr><td>113</td></tr><tr><td>114</td></tr><tr><td>115</td></tr><tr><td>116</td></tr></table> <p><math>A \cup B</math></p>	C.Var	111	112	113	114	115	116	<table><tr><th>C.Var</th></tr><tr><td>112</td></tr><tr><td>114</td></tr></table> <p><math>A \setminus B</math></p>	C.Var	112	114	<table><tr><th>C.Var</th></tr><tr><td>112</td></tr><tr><td>114</td></tr><tr><td>115</td></tr><tr><td>116</td></tr></table> <p><math>A \Delta B</math></p>	C.Var	112	114	115	116
A.Var																																	
111																																	
112																																	
113																																	
114																																	
B.Var																																	
111																																	
113																																	
115																																	
116																																	
C.Var																																	
111																																	
113																																	
C.Var																																	
111																																	
112																																	
113																																	
114																																	
115																																	
116																																	
C.Var																																	
112																																	
114																																	
C.Var																																	
112																																	
114																																	
115																																	
116																																	

(b)

<table><tr><th>A.Var1</th><th>A.Var2</th></tr><tr><td>11</td><td>21</td></tr><tr><td>12</td><td>22</td></tr></table> <p>Table A</p>	A.Var1	A.Var2	11	21	12	22	<table><tr><th>B.Var1</th><th>B.Var2</th></tr><tr><td>111</td><td>121</td></tr><tr><td>112</td><td>122</td></tr></table> <p>Table B</p>	B.Var1	B.Var2	111	121	112	122								
A.Var1	A.Var2																				
11	21																				
12	22																				
B.Var1	B.Var2																				
111	121																				
112	122																				
<div></div>																					
<table><tr><th>A.Var1</th><th>A.Var2</th><th>B.Var1</th><th>B.Var2</th></tr><tr><td>11</td><td>21</td><td>111</td><td>121</td></tr><tr><td>11</td><td>21</td><td>112</td><td>122</td></tr><tr><td>12</td><td>22</td><td>111</td><td>121</td></tr><tr><td>12</td><td>22</td><td>112</td><td>122</td></tr></table> <p><math>A \times B</math></p>		A.Var1	A.Var2	B.Var1	B.Var2	11	21	111	121	11	21	112	122	12	22	111	121	12	22	112	122
A.Var1	A.Var2	B.Var1	B.Var2																		
11	21	111	121																		
11	21	112	122																		
12	22	111	121																		
12	22	112	122																		

(c)

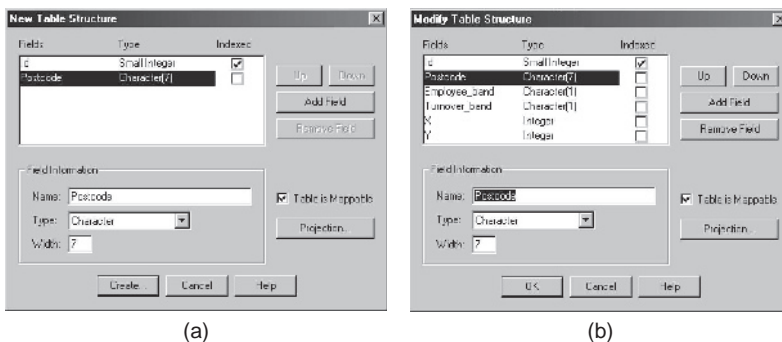
Point of notation: A.key is the primary key in Table A; A.Var1 is the first variable in Table A.



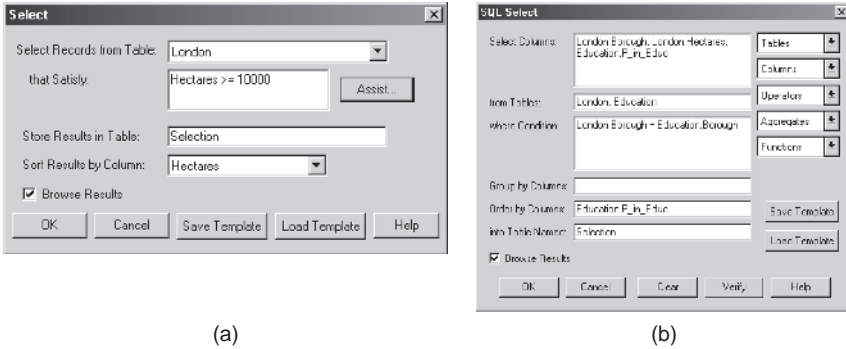
## 8.7.2 SQL and Extended SQL

The most common (*de facto* standard) implementation of RA as a declarative language is SQL. Implementations of SQL can have subtle differences even when designed by the same company. Thus, for example, Microsoft Access uses \* as a wildcard character in searching text whilst Microsoft SQL-Server uses %. Fortunately, most implementations can be run through a user interface that often incorporates software wizards to assist the user in setting up or querying a database. This will be illustrated using MapInfo. There are two components to SQL: a data definition language (DDL) used for defining database tables, and a data modification language (DML) for populating tables with data and for querying tables. The DDL in MapInfo is run through two interfaces: one to create new tables (Figure 8.16a) and the other to modify existing tables (Figure 8.16b). To create a new table a user needs only name each of the required fields, specify the data type for each field and which one(s) should be indexed. When modifying tables, the existing table structure is displayed and by selecting a field it is possible to change its name, data type and whether or not it is indexed.

The DML in MapInfo is also accessible through two interfaces: one to make selections from individual tables (Figure 8.17a) and the other to perform operations on joined tables (Figure 8.17b). Figure 8.17a illustrates the selection from a single table of London Boroughs (*London*) of all those boroughs whose attribute field *Hectares* is greater than or equal to 10 000 ha. The result is sorted in order of *Hectares* and placed in a new table called *Selection*. In RA terms the sole operation is



**Figure 8.16** Interfaces for the DDL as implemented in MapInfo: (a) creating new tables; (b) modifying existing tables.



**Figure 8.17** Interfaces for the DML as implemented in MapInfo: (a) select from a single table; (b) query of joined tables.

*select* ( $\sigma$ ) using the inequality  $\geq$ . Figure 8.17b illustrates the *join* of two tables: the same table of London Boroughs (*London*) and a table of educational qualifications by London Borough (*Education*). The *join* ( $\otimes$ ) is achieved by matching the primary key *Borough* (the name of each Borough, as in Figure 8.10) in table *London* and the primary key *Borough* in *Education*. Only three attribute columns are projected ( $\pi$ ): *Borough* and *Hectares* from *London* and *P\_in\_Educ* (percentage of resident population in education) from *Education*.

The examples of queries just given are SQL implementations of standard RA operations. There is nothing inherently spatial in the queries other than the fact that the primary key happens to be a list of geographical places (London Boroughs, as in Figure 8.10) and one attribute of these places is their size in hectares. Thus, although these illustrative examples were carried out in MapInfo, they could just as easily have been carried out in Access, SQL-Server or Oracle. To resolve queries that are inherently spatial, SQL has had to be extended. SQL has already been extended beyond the operations available in RA to meet the needs of business applications, namely the aggregation or summary of retrieved records using count, sum, average, maximum and minimum. To handle spatial objects, SQL has needed to be further extended and the way this has been achieved varies across the different implementations. MapInfo, for example, follows a more traditional GIS route in which the geometric data of objects are stored separately to their attributes. The extended SQL is both the means to perform operations using the geometry of objects and to retrieve relevant attributes. An example is given in Figure 8.18. Here the *London* table, which has polygon objects (the Borough boundaries), is jointly

queried with a *Post Offices* table, which gives the location and other attributes of post offices that are point objects. To *join* the two tables together so as to determine, say, the number of post offices in each Borough, a *spatial join* must be carried out, in this case through point in polygon matching. In order to do this there must be an extension to the database structure and to the SQL operators. The first is to allow a data type that is a geometric object (e.g. *Post\_Offices.Obj*), which may be points, lines or polygons. They are software generated (rather than created by the user) and act as a primary key linking the geometry of the map representation and the attribute table (as in Figure 8.1a). The second is to augment the number of SQL operators to facilitate *.Obj joins* which in the current example is *Within*; specifically: *Post\_Offices.Obj Within London.Obj* to create a point in polygon spatial join. Another aspect of the example in Figure 8.18 is the use of *Count(\*)* to create a new variable in the resulting table that is the count of post offices in each Borough. The extensions to SQL implemented in MapInfo are also summarized in Figure 8.18.

The approach taken by Oracle Corporation in its Oracle Spatial ORDBMS is to store both geometric data and their attributes in the same database. An object data type *MDSYS.SDO\_GEOMETRY* provides the means to specify 2-D geometric objects (points, lines, polygons) together with the coordinate system (longitude–latitude, British National grid and so on) and the *x,y* coordinate pairs that describe the geometry. The *SDO\_GEOM.RELATE()* function is an SQL extension to determine topological relationships using the nine-intersection matrix [Equation 8.1] with *INSIDE* being the equivalent of *Within* in the MapInfo example above. There are also SQL extensions for some of the spatial-analytic relations in Figure 8.9 with, for example *SDO\_NN()* returning the nearest neighbour and *SDO\_BUFFER()* returning buffered objects. Thus, whilst as argued in Section 3.4.7 Oracle Spatial would be classified as a spatial database rather than GIS as commonly understood, its ORDBMS extensions to the data structure and SQL make it a powerful engine for LBS.

### 8.7.3 Querying Graphs

Graph (or network) data structures are of particular interest in LBS because they are used to represent transport networks along which service users may require guidance for travel. Graph data structures are founded on the entity-relationship model for lines (or edges) connecting nodes (or vertices) as in Figure 8.2b. Topologically they are

SQL Select

Select Columns:

London.Borough, Count(\*)

from Tables:

London, Post\_Offices

where Condition:

Post\_Offices.Obj Within London.Obj

Group by Columns:

London.Borough

Order by Columns:

into Table Named:

Selection

☒ Browse Results

OK

Cancel

Clear

Verify

Help

Tables

Columns

Operators

Aggregates

Functions

Save Template

Load Template

#### Extended Operators

Contains  
Contains entire  
Within  
Entirely within  
Intersects

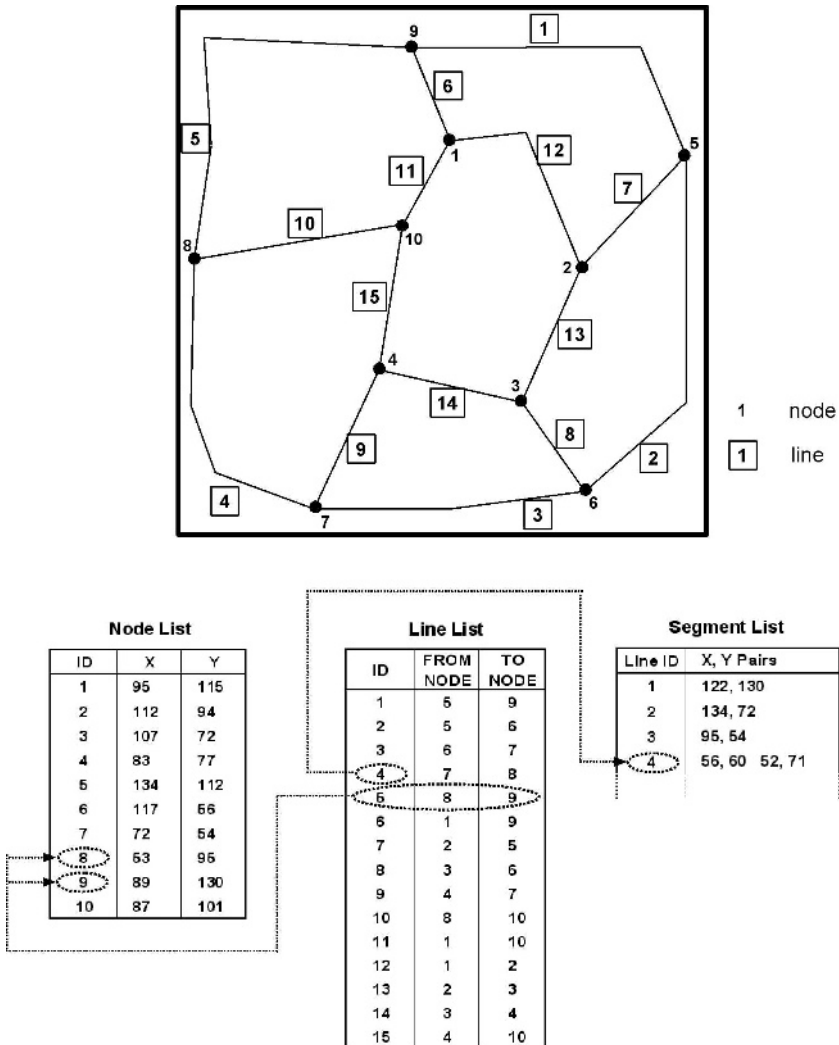
#### Extended Aggregates

Count  
Average  
Minimum  
Maximum  
Sum  
Weighted Average

#### Extended Geographical Functions

Area  
Distance  
Perimeter  
Centroid X  
Centroid Y

**Figure 8.18** Implementation of extended SQL in MapInfo.



**Figure 8.19** Topological data structure for a network (adapted from Brimicombe, 2003).

similar to Figure 8.7 without the need for the polygon elements (as shown in Figure 8.19). The fundamental relationship is one of *connectivity* – how the lines join together to form a planar graph along which movement can occur. It should also be remembered from Section 8.2.1 that each line has an implicit direction from its start node to its end node (although if a line represented say a road, traffic

could flow bi-directionally as an up-flow and a down-flow). Attributes of line networks are used to limit directional flow (as in one-way systems) and to provide impedances on the rate of flow (e.g. average speed of traffic, number of lanes, gradient). Some of these attributes may be temporally defined, such as average speed of traffic during rush hour and average speed at other times.

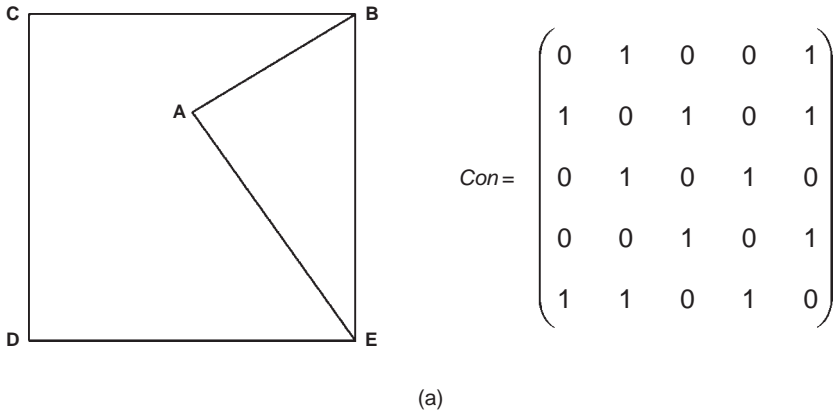
Most of the SQL and extended SQL commands discussed above can be applied to network coverages. One important class of queries that distinguishes graph data structures is *network analysis* queries. Network analysis is designed to answer a number of generic queries:

- efficient travel routes across a network that might either be the shortest distance, the quickest to complete given certain impedances or the least cost (this could be to either a single destination or a series of destinations);
- determining which, amongst a number of competing destinations, is the nearest or quickest to get to;
- generating travel directions for a chosen route;
- creating zones of equidistant or equitemporal travel from a point of origin, or creating such zones around a number of points that represent service areas.

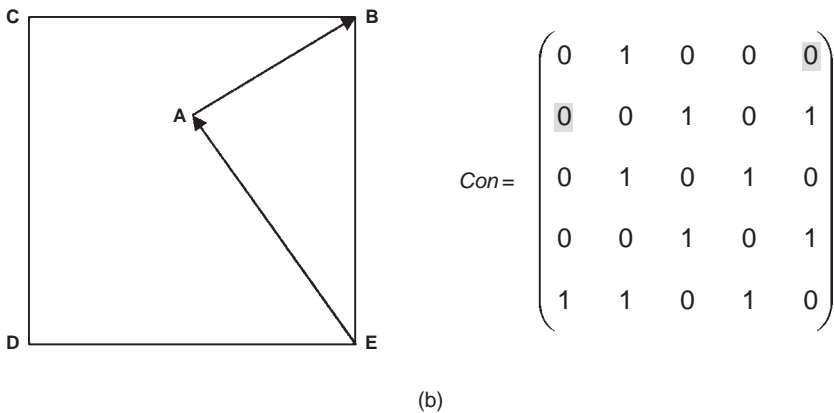
To achieve network analysis queries, graph traversal algorithms are required. These are part of a class of mathematical and algorithmic problems known as *recursion*. These are problems for which the solution is to be found within itself. For example, taking the Fibonacci numbers: 0, 1, 1, 2, 3, 5, 8, 13, 21, ... any  $N$ th number is derived from:

$$\begin{aligned} FIB_{(N)} &= FIB_{(N-1)} + FIB_{(N-2)} \\ FIB_{(1)} &= 0 \\ FIB_{(2)} &= 1 \end{aligned} \tag{8.2}$$

In other words, from any two starting numbers, the whole series can be derived from itself as each successive number is the sum of the previous two. The Morton ordering in Figure 8.3b is also a recursion, often referred to as a *recursive tessellation* as it has a space-filling hierarchical structure. In order to solve the classes of network queries just given, it is necessary to build a recursive spanning tree upon which calculations and route finding can be performed. To illustrate this the simple street network given in Figure 8.20a will be used. The topological data structure for line networks in GIS (Figure 8.19) can be



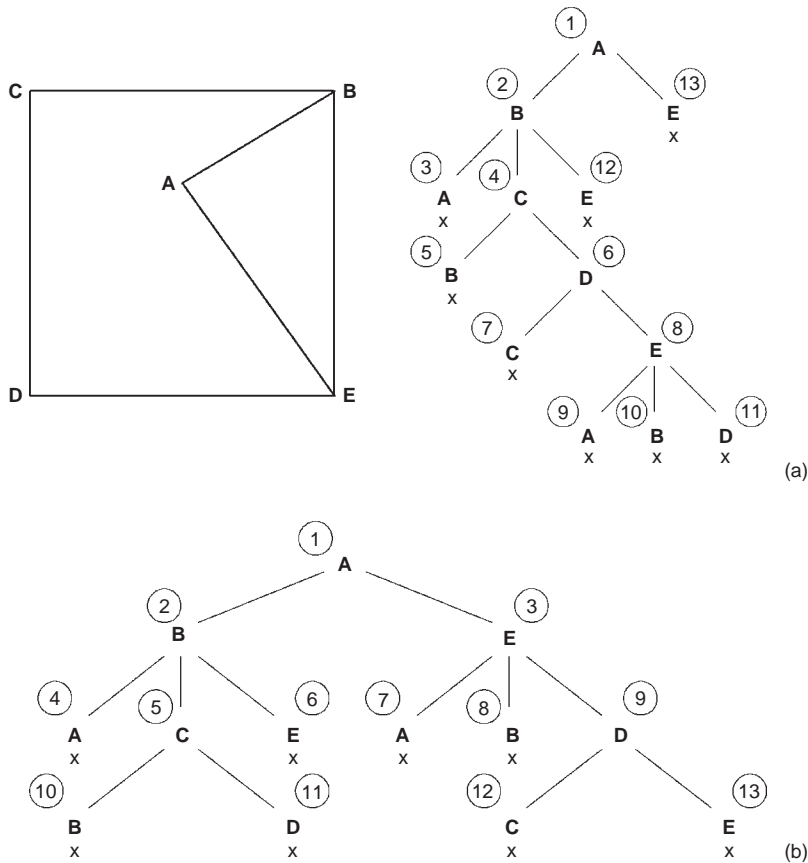
$$Con = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$



$$Con = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$

**Figure 8.20** Connectivity matrices for a simple road network: (a) assuming bi-directional travel along all lines: (b) with a section of one-way access (shaded entries in matrix *con* are those that have changed).

further summarised as a *connectivity matrix*, that is a binary matrix in which the rows and columns represent the nodes in the network and a 1 is inserted if any two nodes are linked by a line, else 0. Figure 8.20a assumes each line can be travelled either way, say from *A* to *B* and from *B* to *A*. This results in a symmetrical matrix such that entries above the diagonal are mirrored below the diagonal. If a one-way system is introduced through *A* such that access to *A* is only from *E* as illustrated in Figure 8.20b, then two connections from *A* to *E* and from *B* to *A* must be changed to 0 and the connectivity matrix becomes asymmetric.



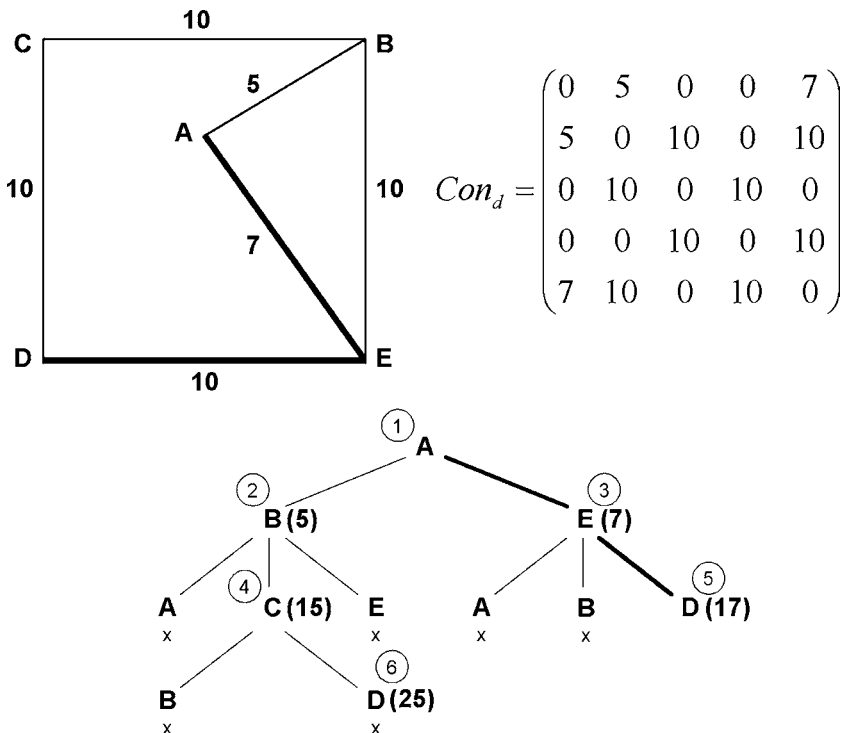
**Figure 8.21** Construction of spanning trees for a simple road network: (a) depth-first search; (b) breadth-first search.

There are a number of algorithms for establishing a spanning tree of a network from any desired starting point (node) on the network (Harris and Stocker, 1998). Two broad classes of such algorithms are depth-first search (DFS) and breadth-first search (BFS). These are illustrated in Figure 8.21 for the same simple road network with the starting point at A; the order in which the spanning trees are constructed is given by the numbers (①, ②, ...). In both approaches the spanning tree is grown recursively downwards stopping at any node that has already been included in the tree. They are both the same size, resulting in a 13-node tree; but there the similarity ends. DFS, as its name suggests, exhaustively follows a series of nodes until all that can



be visited on that particular branch have been recorded; it then resumes at incomplete branches at higher levels (Figure 8.21a). DFS finds the nodes in the following order:  $A, B, C, D, E$ . BFS, on the other hand, successively establishes all single adjacent links first before proceeding down to the next level. BFS finds the nodes in the following order:  $A, B, E, C, D$ . Both spanning trees can be pruned of all end nodes to reduce their complexity.

Such spanning trees can be used to solve a range of network problems. For example, in Figure 8.22 the distances are given for each of the lines and reflected in the connectivity matrix ( $Con_d$ ). What is the shortest distance between starting point  $A$  and end point  $D$ ? By using the BFS spanning tree and summing the distances to successive nodes, the route  $A-E-D$  is quickly established as the shortest route. To search exhaustively for all end nodes at  $D$  in a BFS spanning tree from  $A$



**Figure 8.22** Using a spanning tree to find the shortest route from  $A$  to  $D$  where  $Con_d$  is the distance weighted connectivity matrix [note: a minimum spanning tree from  $A$  would be pruned to the two branches  $(A-B-C)$  and  $(A-E-D)$ ].

would show that all other routes to  $D$  were longer. In fact, given any sort of weighting scheme (e.g. distance, average speed of travel, cost of travel) it is possible to construct a *minimum spanning tree* (Prim, 1957) in which each node only appears once in the tree where the cumulative weights are a minimum. In Figure 8.22 this would result in a pruning of the spanning tree to just two branches:  $(A-B-C)$  and  $(A-E-D)$ . *Dijkstra's algorithm* (Dijkstra, 1959) is a more complex shortest route algorithm that grows the spanning tree down successive branches, switching branches when necessary to maintain a minimum weighted route at each level of the hierarchy in the tree. The algorithm provides for a well tested, unique solution.

### 8.7.4 Query Optimization

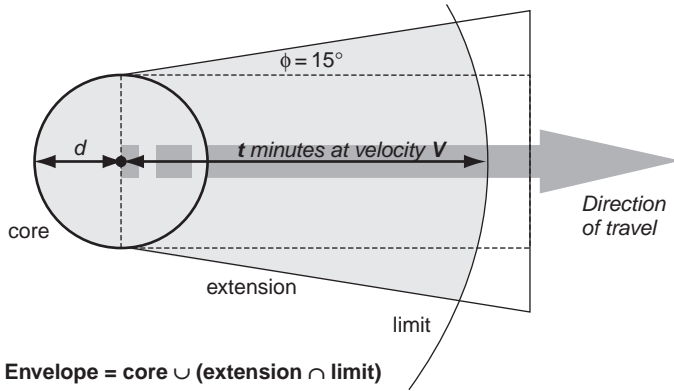
Given the high CPU and I/O costs associated with spatial queries, it is necessary to consider the issue of optimizing queries to reduce the overall time for a response whilst maintaining the utility of that response. Queries that are made in, or translated to, SQL need to be resolved by the database software as a sequence of operations that result in the desired response. In doing so the software will make use of the way the data have been structured (e.g. topologically) and the indexes that are maintained. For spatial databases where there are geometric and attribute dimensions, one common problem in optimization is whether to carry out selection based on geometry first (often CPU intensive) and then by attribute (often I/O intensive) or to do it the other way round. This is where pre-storage of topology, area, perimeter, nearest neighbour and appropriate spatial and nonspatial indexing can radically help to reduce the costs. Most spatial databases have a query optimizer which can calculate the costs of different approaches to resolving a query and which, accordingly, generates an optimal execution plan for an SQL statement. Since spatial joins of objects with complex geometry can involve a high level of computation, it is usually more efficient to reduce the number of candidate spatial objects by executing any attribute selections first. Thus even a simple query to find all pizza restaurants within three miles of a user would first select all those outlets with attribute 'pizza' from a list of all restaurants and then test their distance from the known location of the user, rather than finding all restaurants within three miles and then selecting from them those with attribute 'pizza'. To reduce the cost of spatial joins a strategy is adopted in databases such as Oracle

Spatial to filter and then refine. The filter stage uses the minimum bounding rectangle of an object (as in Figure 8.11) in conjunction with the spatial indexing (as in the R-tree) to identify candidate objects that approximately satisfy some desired relationship. So, for example to find which objects in Figure 8.11 *intersect*, the MBR would indicate objects ( $G$  and  $E$ ) and ( $H$  and  $C$ ) because their MBR intersect. Calculating intersection based on rectangles is very much quicker than the exact object geometry. Having filtered out these objects from the rest, the refine stage checks the detailed object geometry to verify that intersection occurs.

For the most part, it is assumed that database queries are executed only once for a transaction to be completed (Leung *et al.*, 2003). However, an LBS request for, say, traffic alerts whilst travelling is a query that needs to be run over again until the time validity of the service request expires. Not only would this allow dynamic data to be re-queried once they are refreshed, but that the user's changing location would need to be included in the query specification. For a large number of users subscribed to a service, the load on the database to resolve such cyclic queries is likely to be considerable, requiring no only optimization but also scheduling.

Query optimization is undoubtedly important for LBS, yet it is likely that spatial queries will continue to be long to execute, more so when the time for communication of the initial query, establishing user's location, and receipt of the response need to be transacted across a mobile network to and from an LBS provider. One way of reducing the time cost of a query is to reduce the amount of data to be searched by performing a 'soft clip' pruning of relevant data sets in anticipation of queries. For this to be effective there would need to be a means of anticipating what might be the geographical area of interest to a user. Brimicombe and Li (2006) have researched such a concept through their mobile space-time envelopes, which were introduced in Section 7.7 in relation to contextualizing the user. Their use in relation to the spatial query is discussed here.

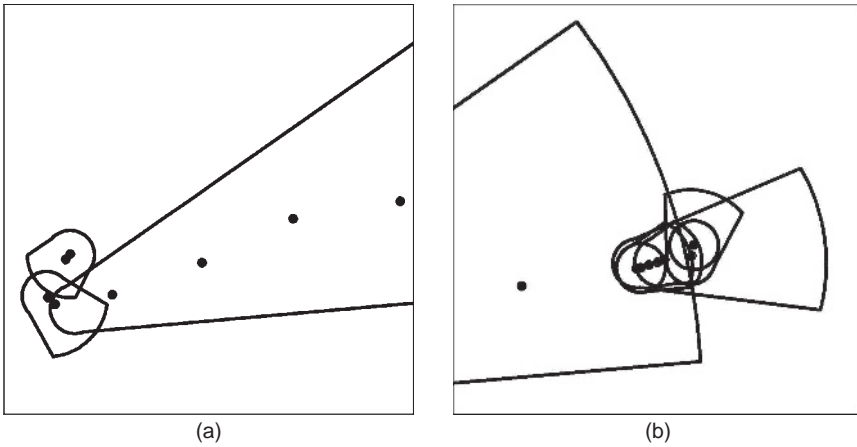
The geometric components of mobile space-time envelopes are given in Figure 8.23; these are the core, extension and limit. The core is a circular zone of radius  $d$  around the location of a user and is the default area of interest when the user is stationary or moving at low speed (e.g. walking pace). Once the speed of movement exceeds a threshold speed, the extension region is added to the core. This extension region is a rectangle formed by the core area moving forward in the direction of travel but splayed out by  $15^\circ$  on either side. The extent



**Figure 8.23** Geometric components of mobile space-time envelopes (based on Brimicombe and Li, 2006).

of the extension area is controlled by a limit calculated as the distance that a user would travel (at current velocity) in  $t$  minutes. Thus the faster a service user travels, the larger the mobile space-time envelope becomes; it is analogous to the beam of a search light travelling with and stretching out in front of the user. When a user is stationary or travelling slowly, they are at the centre of the envelope and may be willing to travel in any direction to obtain a nearby service (e.g. find a pizza restaurant). At speed, a user is less inclined to turn around and needs more advanced warning of available services and directions to find them, hence the elongation of the envelope well in front of the user. The envelope functions by soft-clipping the geographical area of likely interest (such as the filtering approach discussed above) and helps define which leaves of an R-tree (Figure 8.11) would be pertinent. In this way, when a query occurs (or a temporally valid query is refreshed for a new location) the amount of data to be searched has already been reduced.

A real example is provided in Figure 8.24 in which  $d = 1$  km and  $t = 15$  mins; each point is a GPS reading of location taken every 3 minutes and only selective envelopes are shown. The journey begins in Figure 8.24a where a service user is driving along small country roads to the nearest town. Because of the low speed, the envelopes are only slightly extended. Once leaving this town on the main road, the car speeds up to about 80 kph and the envelope extends ahead of the user, sufficiently far in Figure 8.24b to intersect with road works and traffic



**Figure 8.24** Mobile space-time envelopes: (a) at the beginning of the journey; (b) in the traffic queue at road works (see text for explanation; adapted from Brimicombe and Li, 2006).

queues at the approach to the next town. At this stage a user could be advised through a traffic alert and perhaps an alternative route could be found (though in this case there is no rational alternative that is likely to be quicker than going past the road works). The effect of the slow traffic at the road works on the envelopes is evident in Figure 8.24b. After going through the road works, the user speeds up (again extending the envelope) turns left (north) into the town and stops at a petrol station (where the GPS readings end for this example). The anticipatory, geographical soft clip to the mobile space-time envelopes are designed to speed database responses to likely queries (for the discernable context) by limiting the size of data that might need to be searched exhaustively.



# Chapter 9

## Communication in Location-Based Services

### 9.1 Introduction

---

An important aspect of LBS is to provide the information requested by mobile users in an effective and inherently useful way. In Chapter 7, the need for LBS to be context-aware as part of maximizing the utility of services was discussed; in Chapter 8, the importance of the spatial query to extracting relevant data and information was examined. In this chapter, the focus is on the communication between a service provider and users, particularly on the modes of communication by which information are delivered to users. These modes of communication also concern the way users communicate with service providers. Issues described in this chapter are to some extent interlinked with the topics discussed in the previous two chapters. Modes of communication in LBS can range from spoken language, written text, gesture, different types of map, to three-dimensional representation and virtual reality. These are discussed in Section 9.2 and Section 9.3.

Some LBS applications, such as navigation systems, provide a number of different communication modes by which users can access location-based information and services. For example, in-car navigation systems aim to assist people in their activities by giving them alternative ways to use information both through maps and voice. Navigation instructions can be delivered to user mobile devices with spoken words and maps showing the designated route. All along, an

important consideration of these LBS applications must be that pertinent and timely information is delivered and presented to users in an efficient and actionable way in order to meet needs. Therefore, the question arises as to what information is likely to be relevant in any given situation (also discussed in Chapter 7) and how it should best be communicated to users. Such communication can be further understood from the aspect of users communicating with devices to request information based on their needs. The communication between users and devices concerns the interaction between users, devices and environments where users are situated.

The interaction that takes place between users and LBS will strongly rely on the available device interface and technology specification (e.g. screen size of device, device processing capability, network data transfer rate). Beyond the technological interface design issues, it is important to consider the mode of communication in LBS as providing the link and interaction between users and service provider. LBS should be able to adapt to a user's interaction and to tailor the content and presentation of the information accordingly. Furthermore, particular aspects that relate to LBS users in mobile situations need to be considered. All of these will have an influence on the modes of communication employed in LBS, and are discussed in Section 9.4 and Section 9.5. Preconceived notions of appropriate approaches based on experience of fixed stations (e.g. desktop PC) cannot be assumed to work effectively in this new environment. This chapter poses some important questions and in providing answers draws upon new research being carried out by the second author, Dr Li, on LBS preferences and the interaction between the individual and mobile devices in urban wayfinding tasks; this is discussed in Section 9.5 and Section 9.6. Communication in LBS includes the presentation of information and delivery of services to users, the request for information by users and their interaction with mobile devices, and the dynamic contexts in which they are taking place. These pose considerable challenges in the design of LBS applications.

## **9.2 Modes of Communication in LBS**

---

With advanced development in mobile telecommunication networks and mobile devices, the delivery of data and information services via LBS can be achieved through a number of different modes of



communication to users. In the meantime, users can interact and communicate with mobile devices using a similar range of communication modes. Modes of communication concern both the presentation to users and interaction with users. The range of modes can include voice, text, gesture, graphics, image, video as well as 3-D models and virtual reality. In this section, a number of modes of communication employed in LBS applications are discussed, some of which are in common use whilst others are being developed and are yet to be widely applied. In terms of the development of technologies, new possibilities are arising all the time due to: the efficiency of processors, size of memory, wireless networking, sensors, optoelectronics and biomaterials. New modes of communication will emerge and will no doubt be incorporated into LBS applications.

The auditory mode of communication is a basic mode in mobile telecommunications, with spoken words and dialogues as a primary service. It is also a basic mode in LBS applications, often being referred to as voice/speech mode or voice user interface. Most in-car navigation systems provide voice instructions as one of their main services. The auditory mode of communication can be speech dialogues between people or just the playing of a recorded commentary. Where there is interaction and communication between users and systems, such as a user query dealt with by databases without human intervention, it usually needs the support of speech servers and functionality of a Voice User Interface (VUI). VoiceXML is often used in designing and building applications with a voice mode of communication. Text can also be output in an audio mode using a speech synthesizer. For example, there is a 'Read Out Loud' function in Adobe Acrobat that will allow a document to be listened to through an in-built speech synthesizer. Audio mode can be used either as output (presenting information) or as an input mode. Functions such as speech recognition are needed when audio is used as an input mode. Speech recognition can be achieved either by the speech recognition function residing at the server side, with the mobile device as a remote client using the function via Java-based voice streaming, or the mobile device can have a speech recognition function embedded within it and the resulting data passed to a server.

One main advantage of audio mode is in allowing information to be communicated without impinging on a user's visual attention, which can be particularly appreciated when users are engaged in other tasks at the same time. However, if the content provided by audio is lengthy and contains substantial amounts of information, users may

not be able to capture all the needed information due to limits in capacity for immediate recall and it can result in cognitive overload. Functions of 'playback' and 'pause' on audio devices may alleviate the problem to a certain degree, but by deploying them users may lack the consistency in acquiring the knowledge from the information being provided. Some users may feel a lack of privacy in listening to personalized information in public places. Audio can be played through earphones, which users might not always feel comfortable wearing. Furthermore, the content of services and types of applications can have an impact on the suitability of using the audio mode of communication. Some location-based notification and alert services can have greater impact when using an audio mode of communication, though this need not be limited to voice. There are also possibilities of using audio mode to connect to one's physical surrounding environment with sounds associated with it, or even using music to keep people on track. Volume, pitch and synthesized sounds can be used as ways of communicating location and track information.

Text as written words and symbols is another commonly used mode of communication. The Short Message Service (SMS), as a text mode, is one of the most popular services in mobile telecommunications. Text mode is also frequently used in LBS as a convenient way to deliver and present location-related information. Examples include delivering information about nearby available services to user mobile devices by texting, providing written navigation instructions to assist users in real-time, as well as 'push' types of services for reminders, warnings and advertising when users are within range, and status texting used for group communication in buddy-type systems. Furthermore, text mode is a common method for users to input requests for services. Text can be input by keypad or through a screen interface on the device. The text mode of communication can be more personal and private. In using text mode, the choice of words and the meaning of words perceived by users are issues that need to be considered. This is further discussed in Section 9.4. Again, applications and users can influence the suitability of text mode. For example, in applications introducing tourist attractions, users might find an audio commentary about sites more useful than a long text description displayed on their mobile devices. Alternatively, for applications providing wayfinding assistance, if users need wayfinding instructions for a long or complex journey in an unfamiliar environment, text mode (along with other modes such as maps) could provide them with a better understanding of the route.

There are a number of modes that are particularly used for input. One is the tapping of a stylus, mostly used in PDAs. It provides a way to select items/links and to input requirements through an on-screen keypad. Gesture is another mode of input. One way of implementing gesture is to use a stylus to point to an object on a display screen. Gesture can also be achieved using a mobile device (i.e. a mobile phone or a PDA) by pointing the device directly to an object in the real environment. To do this the device needs to have an integrated electronic compass or other orientation devices incorporated into it. Currently, gestures of this type are mostly limited to 'point-link' gestures for querying landmarks and 'line-link' gestures for querying street or other linear features (Wasinger *et al.*, 2003).

A haptic interface can also be used as a mode of communication. It often takes the form of a motorized computer-controlled device held in the hand. Information is transmitted through haptic sensing of the user's hand, which is by use of muscular sensors. Its main use to date has been for augmentation of vision, with the possibility of substituting haptics for other sensory modality (Golledge *et al.*, 2006). The haptic mode can be used to make virtual objects touchable, and to assist in identifying location and manipulating sensed features. This mode can be combined with other modes to provide different ways to input and access on-screen information. It can particularly benefit people with impaired vision or movement.

Graphics and images are another mode of communication frequently used in LBS. They can support location information as a means of further describing a location or an object. It can take the form of graphics, symbols or icons, an image of an object, or an image of a sign used to identify a location (e.g. shop name or road sign). Images directly related to certain locations can be satellite images, aerial photographs of certain parts of urban areas and rural landscapes, and digital photographs of objects, buildings and views of surrounding environments. As a mode of communication in LBS, it is often used in conjunction with other modes. In addition, video or a set of moving or animated images can be used to present activities related to certain locations.

Maps are one of the most commonly used modes to communicate information in LBS. They are a fundamental way of presenting spatial information. Most information and services in LBS are related to location. Maps are an efficient means of portraying large amounts of location-related information, and also as a basis for integrating other sources of information. The most commonly used examples in LBS

applications can be: points of interests and their associated attributes; wayfinding routes generated according to user requirements and displayed on a map; a track or current location of a mobile user displayed on a map. A number of issues in using maps as a mode of communication both in principle and in LBS are the focus of Section 9.3.

3-D representation on mobile devices can be achieved in a number of ways. One way is as a three-dimensional model of an environment. In order to provide accurate and realistic information for applications, cartographic data are essential in creating such 3-D models. A basic technique of doing this is to extrude building outlines according to building height. The building facades can then be rendered to a lesser or greater extent to achieve an appropriate level of realism. Another representation is to combine 3-D objects with a perspective plan of an environment. This is particularly common in navigation applications. 3-D landmark objects are displayed on a perspective map that can be viewed at different scales. These landmarks are often recognizable buildings, monuments, signs and symbols, which can be created as realistic 3-D object models rendered with either graphic images or photo-realistic images.

Virtual reality (VR) is also a 3-D representation of an environment, which usually is completed with rendered buildings. Virtual Reality or Virtual Environment (VE), in a literal sense, provides three-dimensional representations of computer generated objects projected to two-dimensional displays, within which people view and interact (Slater *et al.*, 2002). One common characteristic of VR is that it is not just the flat screen which people look at, but is a three-dimensional visual world in which people feel a degree of immersion (Lathrop, 1999). In general, VR can be interfaced with users employing either a more immersive way such as being in a CAVE (Cave Automatic Virtual Environment), wearing a head mounted display, or by a less immersive means such as using a large screen or just a device screen. VR can be used on mobile devices with interaction keypad, stylus or joystick. 3-D models aim to represent the equivalent of the real world for users. Therefore, any inaccuracies in the 3-D models, either topographically or in the appearance of the built environment, could have an even more misleading effect on user's understanding than a conventional two-dimensional map.

Although 3-D models may be rendered with photo-realistic images to give them a more realistic appearance, some studies show that 3-D models without photo-realistic rendering demonstrate advantages over those with photo-realistic rendering for small screen

mobile devices in terms of their usefulness (Plesa and Cartwright, 2007). Furthermore, designers need to consider whether a 2-D map representation would give users a better understanding of the configuration (layout) of an environment than a 3-D model or virtual reality representation. This is further discussed in Section 9.5.1. From a technological perspective, providing a 3-D representation related to a user's location on the move results in the need to handle a larger volume of spatial data in real-time. This leads to greater demands on both mobile networks and devices, such as the need for mobile networks with higher data transfer rates and mobile device with higher processing capability and higher resolution screens. A good example of a 3-D urban representation is given in Figure 9.1.

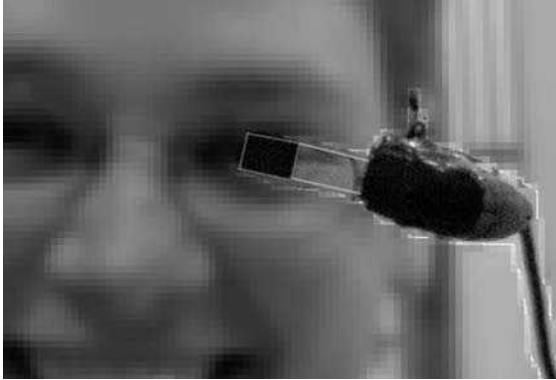
Augmented Reality (AR) brings the real world and information-based virtual worlds into one presentation of information while users are within the real environment that is represented in the augmented reality. It is the mixing of realities on the reality–virtuality continuum, with the real physical world on one end and synthetic virtual reality at the other (Milgram and Kishino, 1994; Milgram *et al.*, 1994). There are generally two ways to bring these two together (Holweg and Kretschmer, 2006). One is commonly known as see-through augmented reality, in which users can view the virtual constructs with the real world (as a background) in the form of a semi-transparent display. Such semi-transparent display equipment, often worn as a headset (Figure 9.2), is necessary for this approach. The other approach is to capture the real world as a video stream to be used as background in an AR scene with the virtual constructs projected onto it. Such AR can be used on mobile devices such as PDAs.

AR in itself can be viewed as a type of LBS, due to its ability to bring together the surrounding reality of a user's position and other relevant virtual information. The virtual information is usually keyed into the location of an AR scene. For instance, it can be spatially referenced text descriptions, graphic symbols or even avatars related to the location of the real world scene (Mountain and Liarokapis, 2007). AR visualization is often limited to only small areas. It is more complex than 3-D or VR. AR can only be of benefit to users if the user location and viewing direction can be determined within an acceptable degree of accuracy. Where, say, text information can be viewed in front of a relevant location, a lower accuracy of position fix may be acceptable. If, on the other hand, some virtual display needs to overlap with a corresponding object in the real world, then higher positional accuracy is required.





Figure 9.1 3-D view of Wuxi City, China (from <http://kp.Wuxi.Cn> viewed on 3 April 2008).



**Figure 9.2** An example of AR equipment worn in a similar way to spectacles. (Photograph by the authors).

Multimodal communication is employed in many applications to different extents. Most commonly used multimodal communication modes can be: auditory mode with text mode; text and/or auditory mode with images/graphics; spoken and gesture input combined with graphics and speech output; voice with maps; text with maps; 2-D maps with 3-D objects and 3-D models. Multimodal communication is adopted in LBS applications principally to enhance the usability of LBS. Taking in-car navigation systems as an example, their services are usually delivered in multimodal mode as maps for the immediate area covered with navigation instructions presented as audio mode (i.e. spoken instructions or dialogues). The route to be followed can also be displayed with graphic symbols (e.g. arrows, cross-hairs for current location) overlaid on the base map together with prominent landmarks. Location-based real-time tourist information guides are usually provided as a multimodal model, often being a combination of text, images, audio and maps (Bornträger *et al.*, 2003). Other multimodal examples are:

- 2-D and 3-D representation of the environment (Kray *et al.*, 2003);
- 2-D/3-D graphics combined with synthesized speech with user input as either speech or gesture in a pedestrian navigation and exploration system (Wasinger, 2003);
- 2-D maps combined with 3-D objects for area exploration on a mobile phone (Rist *et al.*, 2004). This experimentally includes: a 3-D camera flight along a user selected path; a

3-D 360° round view from the user's position; a virtual elevator ride that changes from 3-D view at ground level to a 3-D birds-eye view, which then fades to a schematic view.

Some LBS can deliver information and services through the wireless mobile Internet (Section 2.5.2). Web-based interfaces on mobile devices generally present content to a certain degree as multimodal communication. Furthermore, in LBS applications, spatial data are often integrated with other dynamic data such as real-time information on traffic and transport, weather and local information, information related to tourist attractions and shops, and available nearby services. It also needs to be borne in mind that different modes of communication are used to a certain extent to communicate different information content. This is particularly so from the user point of view. The content of the same navigation information communicated as text, graphics or maps will be understood by users to different degrees from the information offered by the system.

### 9.3 Maps in LBS

---

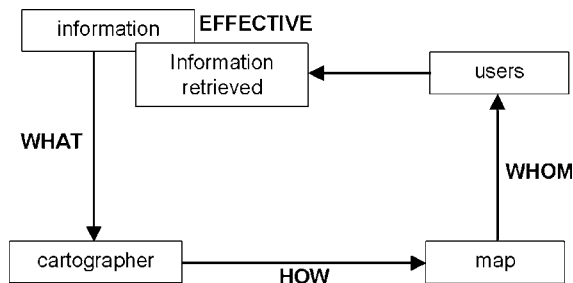
In LBS, most information and services are spatially related and therefore maps play a crucial role in communicating information. Maps have been accessible on the Web for a good many years and are used increasingly on mobile devices. The most commonly used examples of maps in LBS applications are: 'where am I now?', display of nearby services and instructions on how to find and get to certain destinations. It has been suggested that maps were the first multimedia products because they generally comprise text, diagrams, graphics symbols and geographic facts (Cartwright, 2008). Maps of geographical information show a certain degree of selection, classification and generalization. This is compared to satellite images or aerial photographs, which show all the visible details. Additional attributes, which are not necessarily visible from a satellite image or aerial photograph, can also be added to a map. Digital maps are the products of GIS (Chapter 3) and interactive maps are now a feature of Web services. Often in LBS applications, the main GIS-based processes are carried out at the server side due to the performance limitations of mobile devices.



### 9.3.1 Making Maps for Information Communication

The principle of making maps is closely related to the basic communication process in cartography (Figure 9.3): ‘How do I say what to whom, and is it effective?’ (Kraak, 2001). This concept is important for cartographic design, which aims to make maps that are understood effectively by users. The *what* aspect concerns the information content presented on a map. Such content needs to be analysed and determined, and the correct symbols for representing information on the map need to be decided accordingly. The aspect of *how* is the means by which the information is represented as a map. In achieving the creation of a good map, the basic concepts of cartographic theory developed by Bertin (1967) can be regarded as key guidelines (see below). Another aspect is how a user reads the map, which leads to *whom* – the user. Users should always be considered in the process of creating maps, with either an emphasis towards a particular group or more towards individual users. The *effective* aspect demonstrates the amount of information extracted and understood by users. The information represented through a map and the information retrieved by a user will rarely match exactly, which can be viewed as the different levels of effectiveness.

In deciding an appropriate symbology for a map, it is central to analyse the characteristics of the information that is to be visualized and understood (Kraak, 2001). A map contains point symbols, line symbols, polygon symbols and text which label features. The appearance of map symbols can vary in their shape, size and colour. Although the spatial location of an object is determined on a map, the symbol of the object with its variations can represent the object with different



**Figure 9.3** Basic communication process in cartography (adapted from Kraak and Brown, 2001).

characteristics. Six categories are used as the visual variables in cartography (Bertin, 1967); these are: size, value, texture, colour, orientation and form. Each of these variables can provide symbols with different meanings. For example, a line symbol with different sizes or colours can show different road types (e.g. motorway or minor road). How users perceive and understand maps can vary considerably according to the way in which information is represented on a map. The scale of a map is important for users to understand the real world environment that a map represents. Scale is thus an important consideration in GIS and cartography (Section 3.4.1). Furthermore, to let users understand a map effectively, extra descriptive information about a map is required. Such information can be a legend explaining the symbols used, a bar showing the map scale, a north arrow indicating the orientation. In conventional paper maps, notes on map projection and information about the sources, dates of survey and date of compilation are usually included. For a map displayed on a mobile device, such information is usually not displayed with the map due to the limitations of screen size (Figure 4.6). For the most part these have had to make way for copyright statements but should be available for access through user interaction with the map.

### 9.3.2 Digital Maps Used On-Screen

With the development of technologies such as PCs, the Internet, GIS and mobile devices, maps can be made and used in new and more varied ways. Maps can be designed to be used on-screen for a range of applications on the Web and for a range of devices, including mobile devices. Map data are widely used and stored digitally. Although there are limitations in using maps on-screen depending on device type, maps can be created that are more attractive and effective, often having interactive capability and often combined with other modes of communication, such as having multimedia capability in delivering information.

When maps are used as a mode of communication in LBS, most of the elements illustrated in Figure 9.3 need to be understood with more emphasis on mobility, user focus and small device limitations, even though the concepts discussed in Section 9.3.1 continue to be relevant. The boundary between *cartographer* as a map creator and *whom* as a map user has become increasingly blurred as users can decide what information they are interested in and can select or ignore

certain options in creating their maps. Due to the mobility of users, users might frequently require different map content and different ways of visualizing maps to meet their needs in real-time. It also needs to be recognized that there are a wide range of map users. Some might rarely use maps, others may do so professionally. The former may view map representations of spatial information as just another mode of communication (e.g. text, graphics). Few if any established cartographic principles might be applied in their process of selecting map elements and visualizing information on a map. This is similar to the concerns that have arisen with Web-delivered maps. Geographic information delivered as maps over the Web is viewed as part of the popular new media instead of scientific data, and therefore the delivery of cartographic artefacts over the Web needs to take into account that many users are likely to be inexperienced users (Cartwright, 2008).

Another element in the communication process of cartography, which is not emphasized in the basic process shown in Figure 9.3, is the user's request for information and their ability through the system to define and express their needs. Such user requirements are particularly important in LBS. In LBS, user needs can often change due to their geographical mobility, the tasks being carried out and the dynamic situations encountered in real-time. When the information retrieved from a map doesn't match users' current needs, users normally interact with their mobile devices to request further or different information. The element of user requirements is crucial in the communication process when maps are used in LBS. This brings us to the interactivity of on-screen maps that is discussed next.

Interactivity is one of the main characteristics of on-screen maps. It is particularly important for maps used in LBS because the interactivity offers the ability to display more information on small screen devices and to deliver larger volumes of data quickly to users on the move. It is more important for users in mobile situations to have relevant information delivered to meet their requirements than to be overloaded by unnecessary information. There is more emphasis on actual user interaction for maps displayed on mobile devices, even though there are some view-only maps in use. The interactivity is often used to connect different map contents and information contained in the current visualized contents to access extra information. It can also be used to switch between 2-D maps, 3-D models and virtual reality. Maps can be overlaid with aerial photographs, and combined with other modes of communication through user interaction. The commonly used interactive features for maps include map scrolling

and zooming (in and out), links, clickable features, mouse over and objects displayed on top of a map as pop-up frames.

The symbols used for maps on-screen can be interactive; and different map information content can be represented via such symbols. In general, for small screen display, the size and the resolution of map symbols should be considered both for legibility and for interactive functionality (e.g. symbols can be easily clickable to access other content). On-screen maps can also provide the possibility for animated symbols and symbols combined with other communication modes such as text and sound while users interact with them. More details on the design of map symbols for on-screen maps are provided by Kraak and Brown (2001). Although such interactivity has been studied mainly from a Web perspective, most of the features are relevant for on-screen maps and mobile devices in general. Furthermore, compared to traditional paper maps, map information represented on a screen can be considered as being different from two perspectives: visual hierarchy of content and the scale of map (van den Worm, 2001):

- Map content can be classified as primary, secondary or supportive. Primary level concerns the main theme of the map, some of which can be requested by users through interaction with the map. Secondary level usually refers to base map information (such as topography or road network) together with additional information on the main theme. The supportive level often relates to marginal information that could be retrieved by users for display if required. The supportive content can contain information such as legends and supplementary data to the main theme of a map.
- The scale of maps represented on-screen is, in principle, used more flexibly than paper maps because of the functionality for zoom. However, the content of a map continues to depend on the scale used. There are three general approaches to scale change through zoom. One is to zoom in or out of a map with its content unchanged. The images can be enlarged when zooming in, either with clearer and sharper features (for vector-based images) or with enlarged pixels (for raster-based images). Another approach to zooming in or out is deploying a set of maps at different scales, which are selected based on the level of zoom. The third approach is dynamic zooming, by which the served map content is related to the scale of a map.

In simple terms, zooming-in results in a map with more detailed content and *visa versa* for zooming-out (Figure 3.9). Many maps used for on-screen display, particularly on devices with small screens, show no scale information (Figure 4.6). A scale bar should always be shown if understanding the map information is closely related to the scale of representation. This is important as many LBS applications with maps as the mode of communication often provide users with map zooming functionality and use differently scaled maps across applications (Dillemath *et al.*, 2007). Furthermore, the way in which a scale bar is used on a small screen also needs to be considered because the scale of a map representation in mobile situations can be changed frequently.

Another characteristic of digital maps is that they can be presented and interacted with in a multidimensional way (Kraak, 2001). When spatial data are represented in three dimensions, some features might be hidden when displayed on a 2-D paper map or on a flat surface. However, a digital 3-D map can be used interactively by users in a 3-D space, and information can be queried and visualized in a multidimensional way. Furthermore, cartographic animations can either represent changes in spatial data (such as spatial patterns) over time or the same spatial data set with different representations. Interactive ability is usually needed in such visualizations. The 3-D representation, virtual reality and augmented reality discussed in Section 9.2 can be viewed as multidimensional cartography.

### 9.3.3 Map Generalization

From a technological perspective in LBS, the limited screen size of mobile devices and data transfer speeds over a mobile telecommunications network further emphasize the need for map generalization to control the amount of information on a map. Even though some of these limitations could be reduced with the development of technologies that increase the speed of data transfer, the information on a map delivered to a mobile device is currently limited by technology. Moreover, from a user perspective, one main focus of LBS is to produce more pertinent information that is easily communicated to users on the move in real-time, without producing cognitive overload. Therefore, it is important to understand map generalization when employing maps as a mode of communication in LBS.

‘Generalization is one of the foundations of human inquiry. Without generalization, the observations we make about the world around us remain isolated in space and time. Unless we are willing and able to generalize these layers of experience, we cannot expect to learn from them.’ (Cromley & Cromley, in McMaster and Shea, 1992 p. iii) Map generalization has been one of the most significant components in cartography and digital cartography, as well as in GIS. Geographical reality can be generalized as a map using multiple representations, such as at various scales and different levels of resolution. The motivations for map generalization can be listed as follows (Müller, 1991; Weibel and Dutton, 1999):

- **Economic requirements:** all the data collected are filtered and selected (within tolerable and controllable accuracy limits) during the data capture process due to financial and technological considerations.
- **Data robustness requirements:** generalization can be used to filter out errors, reduce unnecessary detail and consolidate the trends to achieve more consistent spatial databases, such as a smoothing operation for curves and surfaces that can reduce individual observation errors.
- **Multipurpose requirements:** generalization is motivated to provide multiple views of geographical data at various scales and levels of resolution as natural and anthropomorphic features of landscapes are scale-dependent.
- **Display and communication requirements:** spatial data need to be generalized for visual communication and as a means to best absorb and understand large amounts of information.
- **Primary spatial database requirements:** for any intended application, to build a data model that reflects the real world with appropriate content and level of resolution.

From the paper map point of view, map generalization can be considered as a process of compiling a map from a larger scale source map (reality) with regard to map context and graphic presentation (Morehouse, 1995). To reduce the scale and to meet a particular map purpose, map content and graphical representation are decided by cartographers. From a digital cartographic perspective, there are three elements in the objectives of generalization: theoretical, application specific and computational (McMaser and Shea, 1992). The theoretical element helps counteract the undesirable consequences of scale reduction, focusing on reducing complexity, maintaining spatial

and attribute accuracy, maintaining aesthetic quality and a logical hierarchy, and consistently applying generalization rules. The application-specific element concerns map purpose and intended audience, appropriate scale and retention of clarity. The computational element relates to the balance of the relationship between the sampling interval of data, data complexity and storage availability. From a GIS perspective, the emphasis of generalization is not only on the map production discussed above, but also on integration, analysis and subsequent visualization of information.

Map generalization, on the one hand, has a more traditional cartographic focus, which is the visualization of abstractions of geographic reality. On the other hand, it is a desire for geographical information at multiple levels of scale and data resolution. These two are often brought together, signifying that both the simplification of cartographic features in the graphic output and the changes in underlying data are parts of map generalization. These then are the two aspects of the generalization process: cartographic generalization and model generalization. Cartographic generalization, well known in the generalization research area, is usually considered as a map compilation process for resolving legibility problems. Numbers of algorithms have been developed in cartographic controlled generalization; there are spatial transformations and attribute transformations (McMaser and Shea, 1992):

- Spatial transformations centre on geographical and topological data, which include simplification, smoothing, aggregation, amalgamation; merging, collapse, refinement, exaggeration, enhancement and displacement. Further details can be found in McMaser and Shea (1992).
- Attribute transformations emphasize statistical characteristics of features with only the spatial changes that are necessary to describe the change in attribute information. Attribute transformation includes classification and symbolization.

Model generalization focuses more on geographical meaning in a generalization concept and on modelling spatial data at different levels of abstraction, rather than being limited only to cartographic features. Model generalization is aimed particularly at digital maps (Weibel and Dutton, 1999). In such generalization processes, data abstraction is the reduction of both spatial resolution and semantic resolution, whether the purpose is for data analysis or cartographic representation (Müller *et al.*, 1995). Ideally, a cartographic data model linked with multiple



representations at different scales should have cartographic meaning and associated geographic meaning related to the changes occurring through generalization (Buttenfield, 1995).

One approach suggested for model generalization is to employ object-oriented data modelling in generalization (Müller *et al.*, 1995). The object-oriented approach can be used for multiple representations from a single spatial data set in order to facilitate the model generalization (Li, 1995). In this approach, the geographical meaning underlying cartographic features can be investigated and presented when creating the class hierarchies for cartographic features. The object-oriented inheritance can then be used to describe different levels of abstraction of reality. Therefore, the data model created for generalization can delineate different levels of abstraction of a phenomenon for both cartographic features and their geographical meanings. When the scale changes from large scale to a smaller scale (e.g. from 1 : 1250 to 1 : 5000), certain features can be eliminated as required by the map specification. This can be achieved by selecting those classes at a higher level of abstraction in the class hierarchy of an object-oriented model. Another potential approach is to use agent technology for map generalization. In an agent-based approach, agents can be assigned to manage the generalization process with a local perspective in a geographic region and then communicate with other agents with a broader perspective at a more regional scale (Duchêne, 2003). This approach aims to ensure consistency in map generalization across the mapped space.

Approaches used for implementing computer-assisted map generalization can be grouped into: batch approach; interactive approach; and knowledge-based approach. In a batch approach, the map generalization of a data set is achieved through using algorithms, rules or constraints from an input data set processed in a computer without human involvement. For example, the Douglas–Peucker algorithm (Douglas and Peucker, 1973) is one of the most widely used line simplification algorithms; it is included in many GIS software as generalization functionality. A line data set can be generalized automatically using the function after setting up certain parameters. In a batch solution, a procedural process with certain rules and constraints can be applied for the generalization of different objects in a map. This approach is often based on mathematical procedures with emphasis on geometric issues. The interactive approach involves user interaction during the generalization process. Users can interact through an interface to select intended objects and desired generalization tools for high-level tasks during the generalization process. The knowledge-based



approach can be regarded as an extension of the interactive approach, combining artificial intelligence techniques with the batch and the interactive approach. One of the challenges is formalization of procedural knowledge in order to activate generalization methods (Mackaness, 2008).

### 9.4 Issues Around Modes of Communication in LBS

---

LBS provide users the mobility of accessing and using information and services through small, wireless, portable devices. LBS, however, pose a number of particular challenges in the modes used for communication between service provider and users as compared with similar applications using a fixed line connection. Some of these challenges have already been touched on in the previous two sections. Here, the main points will be recapped before these issues are fully addressed.

The characteristics of mobile devices used in LBS have an important influence on the interaction and communication between users, devices and environments. Firstly, mobile handheld devices are necessarily small in size and light in weight. Therefore, the processing capability and resources of such devices are generally less in comparison to the computers at a fixed location. This is related to the device's processor, memory and data storing capability, all of which can have effects on the mode of communication used in LBS.

Secondly, the bandwidth provided by mobile telecommunications networks used in LBS is constrained compared to fixed line connections. For mobile devices the transfer speed of data streaming can mean a long wait for large volumes of data. When dealing with 3-D models or virtual reality on a mobile device, there would be a large amount of data that needed to be transferred. Hence the transmission time could easily be too long to be acceptable. Furthermore, as discussed in Chapter 8, at the server side spatial queries are typically long queries. The 2G mobile wireless networks, such as GSM, can appear very slow with its 10 Kbps data transmission rate (Peng and Tsou, 2003), and even GPRS as a 2.5G network only has a data transfer speed of 115 Kbps. A fully implemented 3G network can achieve 2 Mbps data transmission rate, with 4G envisaging a data transmission rate of up to 100 Mbps, and even to 1 Gbps (Section 2.3.1), which would then be sufficient for most currently envisaged LBS applications.

Thirdly, mobile devices have limited display capability and small multifunction keypads, which have an effect on the modes of communication for LBS. The size of screen in mobile devices is small with relatively low resolution. The development of technologies will continually improve mobile devices; however, the nature of the small mobile device provides user interfaces that are different from a PC or laptop. There is also the need to consider the variation in screen resolution and quality in different makes and models of mobile device. For example, the colour of maps can be seen differently depending on the quality of device used, and an overview of information such maps could be lacking in smaller screen devices with lower resolution. In addition, keypad and small cursor can cause difficulty for some users. Importantly, limited screen size also puts constraints on interactivity in communication modes. Extra information such as label placement for images/objects can cause overlap of labels or too much overlap with images. Also, it may not be possible to view information supplied (say as a map) and have space left on the screen to allow users to initiate further queries, pan, zoom and so on without scrolling the information out of view.

Lastly, the operation of handheld mobile devices depends on the energy provided by batteries. Because a mobile device should ideally be lightweight and palm-size, the battery size and thus its energy supply is very limited. Battery weight is usually proportional to the energy supplied. Users of course can keep a backup battery for use, but energy consumption remains a concern even though there has been continuous development in battery technology (Tsalgatidou and Pitoura, 2001). All these issues can have effects on the mode of communication employed in LBS to present information and on how users can interact with mobile devices.

Most of the above issues are from a technological perspective. However, the mobility of LBS brings in additional issues around the nature and characteristics of mobile users. In employing a mode of communication in LBS, it is necessary to consider user's requests, retrieval and understanding of information delivered in real-time. It is also essential to understand how spatial information can be communicated effectively in situations that are mobile and dynamic. The level of effectiveness of communication is related to the representation of information, the types of queries posed by users and to the amount of information retrieved and understood by users. Furthermore, differences in individual spatial ability can further contribute to the effectiveness of communication.

In considering human factors around the mode of communication in LBS, one aspect is the semantics of communication between individuals and a database with or without the intermediary of a human. In LBS navigation applications, instructions are often given to users through maps, images, the spoken word and text. It is envisaged that various landmarks, points of interest (POIs) and key features of neighbourhood environments might be provided as spatial cues via a mobile device. On the other hand, users' ability to formulate queries and answer spatial questions may be limited either by a lack of awareness or by confusion over vocabulary and map reading. This strikes at the traditional core principles of GI-Science such as data modelling, data handling, generalization and visualization. Can objects be described in terms of neighbouring features in exactly the same way that users might describe them after looking at a map? The precise relation of cognition to language remains controversial (Mark, 1999). Can a common semantic for spatial descriptive terms between systems and users be arrived at when they are receiving information through handheld mobile devices? A number of cases reported in the press of people heading to wrong destinations (sometimes because of typing errors), or taking a wrong turn into the opposite direction of a motorway are caused to some degree by mis-specification in requesting information or a misunderstanding of the instructions given via their in-car navigation systems.

Another aspect concerns different levels of cognitive load associated with the mode of communication adopted. Cognitive overload can be more apparent in using information in real-time and in mobile situations in which there are often more distractions for users than it is in a more stable environment such as in an office or at home. This could increase the possibility of imprecise, ambiguous or incomplete understanding of information communicated. Furthermore, in the situation of users carrying out multitasks, which often occurs while users are on the move, the limits to users' cognitive load can have a considerable influence on how users interact with devices and understand information retrieved. The types of tasks for which users seek information could influence the mode of communication used. Thus, users may desire different modes of communication for different LBS applications.

Personal preferences for modes of communication need to be considered, too. Individuals do have their preferred way to access and interact with information and services. Such preferences can depend on user backgrounds, knowledge, physical condition or even just habit. For spatial information delivered via LBS, an individual's spatial ability could also reflect on their preference in choosing modes of

communication. Users' preferences may change when they encounter different situations, particularly in real-time on the move, even though there may have a strong personal default set of preferences. This is demonstrated through a case study in Section 9.5.

The level of user interaction can also influence modes of communication in LBS. It exists for any mode of communication adopted. The level of user interaction can be reflected in the degree of automation offered by mobile devices in terms of their modes of communication. At one extreme end of the automation scale, information and services can be delivered to users fully automated without any functionality for user interaction. This can be set up according to user personal preferences and user current location to reduce user cognitive overload when using LBS. However, it could create a passive way of communicating with users and result in problems, particularly in unpredicted circumstances. At the other end of the scale, a high level of user interaction may be provided. Although this offers the potential for very active communication between users and their mobile devices, it can overload users and increase their cognitive burden. The dynamics nature of automation level in LBS was discussed in Section 7.6 and Section 7.7.

Regarding the range of possibilities for modes of communication, to what extent can they be deployed to improve communication in LBS applications to the benefit of users? Some modes of communication could be out-of-date in the future, and new ones could be emerging with the development of technologies. But issues in LBS concerning how modes of communication can be better suited to the interaction between users and devices, and in the effective delivery of pertinent information and services to users remains an important area requiring further research.

## 9.5 Learning from Spatial Information

---

Providing users with effective and efficient modes of communication to access and use information is closely related to how people can better grasp and understand the information presented. LBS mainly deal with spatial information. Understanding the way in which individuals acquire spatial information, learn from it and develop spatial knowledge can provide insight into how modes of communication can be more effectively deployed in LBS.

### 9.5.1 Acquiring Spatial Knowledge

A study of the nature of spatial knowledge and how spatial knowledge is acquired, developed and used can give insights into how people behave and navigate while they interact with the environment. It can also provide a better understanding of how spatial information delivered through LBS assist users in performing spatial tasks. Spatial knowledge allows individuals to develop various mental schemata of an environment. Such knowledge is transferable between people through different means, such as instructions in text or voice, diagrams and maps. People's spatial knowledge structures are generally viewed as providing the basis for interpreting places in the environment. Spatial knowledge structures are a subset of an individual's knowledge of the environment. A knowledge structure, from the perspective of information processing, is also viewed as a set of symbolic structures representing certain aspects of an individual and the individual's environment (Golledge and Stimson, 1987).

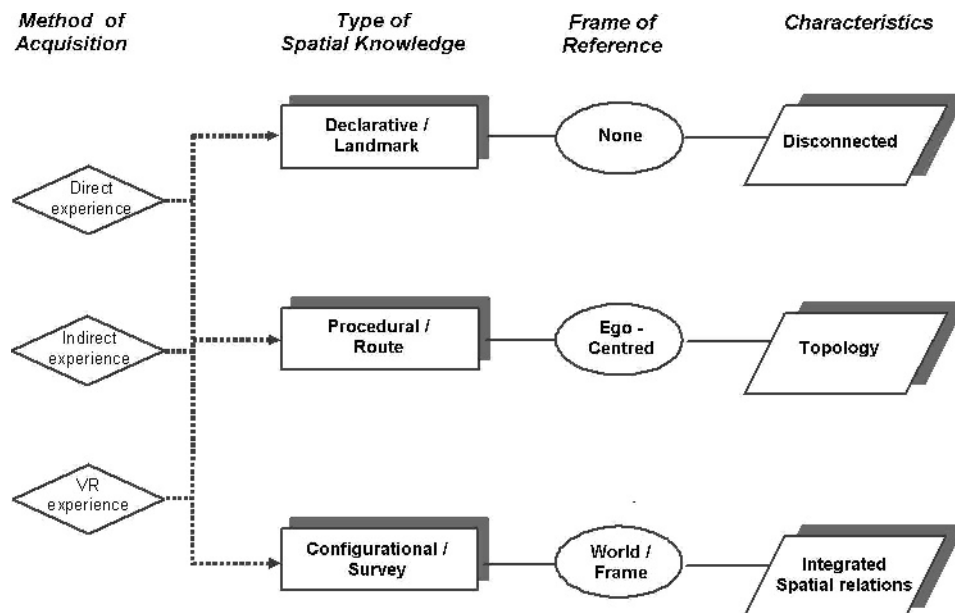
Spatial knowledge can be acquired in the form of three basic components as declarative, procedural and relational/configurational (Shemyakin, 1962; Siegel and White, 1975; Kuipers, 1978; Thorndyke and Hayes-Roth, 1982; Stern and Leiser, 1988; Golledge and Stimson, 1997).

- **Declarative knowledge:** commonly known as landmark knowledge. This type of knowledge refers to those objects and/or places with meaning or significance attached to them.
- **Procedural knowledge:** also known as route knowledge. This typically refers to knowledge about movements and mostly consists of procedural descriptions with landmarks and path elements. It concerns an understanding of the process of how to travel or find one's way from one locality to another.
- **Configurational knowledge:** sometimes referred to as survey knowledge, relational or metric knowledge. This generally refers to the integrated knowledge of the layout of a space and the inter-relationship of the elements within it, which enables people to traverse that space in complex configurations of paths and nodes within some external frame of reference. Configurational knowledge is considered to comprise not only visual and geometric, perceptive and descriptive information, but also spatial relations between objects or places. Such knowledge can allow people to facilitate the

integration of spatial hierarchies and the understanding of spatial phenomena, such as linkage and connectivity between landmarks, routes and regions.

People acquire and develop their spatial knowledge through various experiences and processes; these may include recognizing and understanding characteristics of objects, localities and inter-relationships between elements. Such experiences and processes can take place either through 'direct' or 'indirect' experience. 'Direct' experience is usually viewed as the experience gained through activities in a real environment, which can refer to active learning modes in which people experience an environment via perceptual focusing, head and body movement; for example, learning or exploring routes in a spatial environment. 'Indirect' and 'conceptual' experience relates to that gained through assimilating simplified and symbolized representations rather than from exposure to real environments. 'Indirect' experience can be referred to as a passive learning mode, mostly not involving direct contact with the environment. Such experience can be map study, verbal instructions, through newspapers, the Internet and mobile devices. It is commonly accepted that the 'indirect' experiences, particular through studying map type information, provide people with a better understanding of spatial relations and configuration of the environment, whilst 'direct' experiences enable people to have better route knowledge. An illustrative summary of spatial knowledge is given in Figure 9.4.

Spatial knowledge can also be acquired through VR environments that are structured so as to simulate real environments. As discussed in Section 9.2, VR and 3-D model presentation can be used as modes of communication in LBS. VR experience, as a means of environmental exposure, has a number of shared common characteristics with 'direct' experience, despite subtle differences. For example, spatial knowledge acquired through 'direct' and VR experiences are both shown to be orientation-free when compared with map learning experience (Tlauka and Wilson, 1996). Studies also show that people who acquire spatial knowledge in virtual reality often have similar capabilities to those who acquire it via 'direct' experience, such that they can demonstrate extensive and accurate route knowledge, but have less well developed survey knowledge (Witmer *et al.*, 1996; Wilson, 1997; Ruddle *et al.*, 1997). Therefore, it is possible that VR as a mode of communication in LBS can simulate real world experiences; however, it might not be able to improve user knowledge of the



**Figure 9.4** An overview of the nature of spatial knowledge, its frames of reference and salient characteristics.



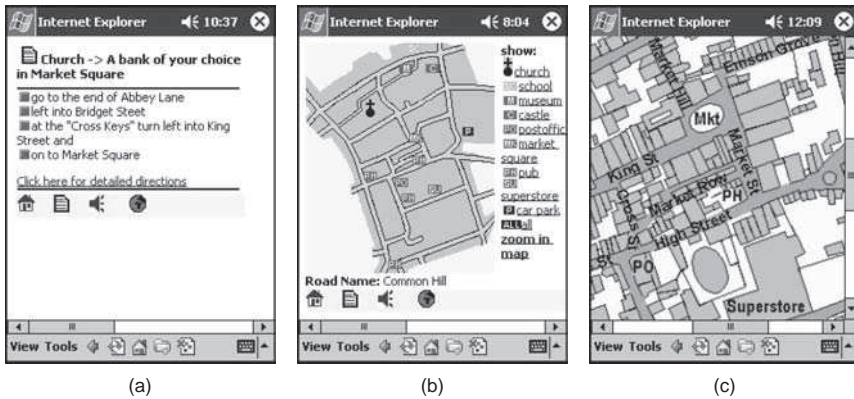
spatial relations in an environment. It needs to be noted that the differences amongst various VR and 3-D modes used are likely to provide different levels of realism and active involvement with the environment. Also, birds-eye type of 3-D representations could have a similar effect on spatial learning as using maps.

### **9.5.2 A Study of User Preference for Different Modes of Communication**

In this case study, user preference for modes of communication was studied through a pedestrian wayfinding scenario in immersive VR. Task-based wayfinding activities were set up in two different urban areas; and a set of tasks that involved ‘walking’ to successive destinations needed to be completed. The urban areas were completely unfamiliar to all those who undertook the tasks. The design for each successive destination in the wayfinding tasks was intended to entail different levels of complexity of the environment in terms of length of route, numbers of turnings and the type and number of choice points passed. A LBS wayfinding assistance application was simulated on a PDA. The users (the participants in this study) had free choice to choose the mode of communication and type of information they preferred, with the aim of exploring issues about preferences in modes of communication during actual tasks. Users were allowed to choose any information available from the PDA depending on their preference and needs in assisting their wayfinding tasks. There were no restrictions upon what type of information could be accessed through the PDA, in which mode of communication it was conveyed, or how often they accessed the information.

Although information can be communicated to users in a wide range of modes, in this study the wayfinding assistance was provided to users as text mode of route instructions, voice mode of route instructions, schematic map mode with overall street layout and landmarks, and detailed map mode with zoomed-in partial areas (Figure 9.5). Urban areas, where the wayfinding tasks took place, were virtual reality urban models. These VR models were built based on topographical maps of real urban areas with photorealistic texture from buildings in the same areas. The VR models were implemented in a fully immersive CAVE setting as a part of the test environment created (see Li and Longley, 2006 for more details). The complete user interaction with the LBS wayfinding assistance and their



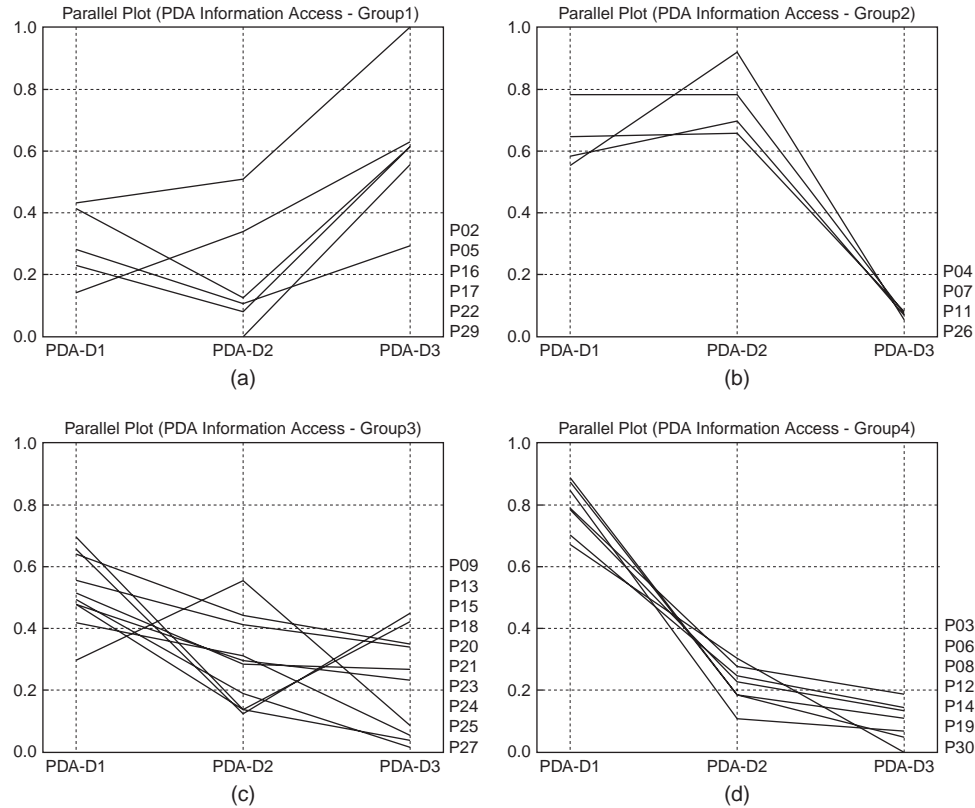


**Figure 9.5** Examples of the wayfinding assist information: (a) text mode for route instructions; (b) schematic map mode with overview street layout and landmarks; (c) detailed map mode with zoomed-in partial area (reproduced from Li and Longley, 2006).

behaviour during wayfinding task completion were observed and recorded in this test environment. This enabled the study of user preferences in real-time and during tasks in mobile situations.

User preference in mode of communication was identified in the study based on the concepts of spatial knowledge (Section 9.5.1). The frequency of access and the time spent on consulting and studying the information presented were used for analysis. The preference for voiced route instructions was not included in the final result due to the small volume of use in these wayfinding experiments. This low level of voice preference was explained by user feedback that indicated that the voice mode of instruction through a mobile device without earphones was the least desirable for pedestrian wayfinding. The feedback also showed that capturing and remembering all the needed information was not easy particularly when instructions in voice mode become lengthy (Section 9.2).

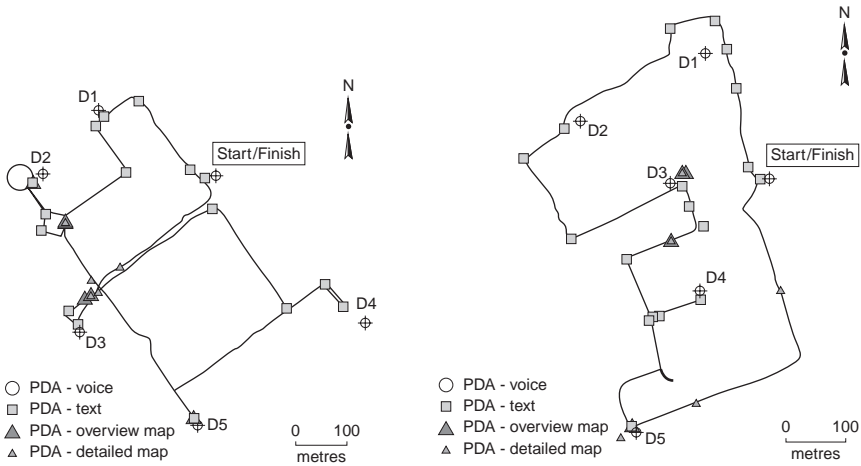
Amongst all users, there were clearly different preferences in the choice of mode of communication for wayfinding assistance. Four groups are identified, named IN-G1 to IN-G4, whose characteristics are illustrated using parallel plots in Figure 9.6. On the x-axis of the four parallel plots in Figure 9.6, PDA-D1 refers to schematic map mode with overview street layout and landmarks, whilst PDA-D2 refers to detailed map mode with zoom-in on partial areas; PDA-D3 refers to text mode of route information (i.e. text mode of route



**Figure 9.6** Parallel plot diagrams demonstrating four user groups with different patterns of preferences for the mode of communication: (a) text preference; (b) schematic and detailed map preference; (c) no strong preference; (d) schematic map preference (reproduced from Li, 2005).

instructions). As illustrated in these parallel plots, the users in the Group IN-G1 have much higher scores on using wayfinding assistance presented in the text mode of route information (PDA-D3 in Figure 9.6a) compared to the other two modes of communication within this group. In addition, their scores in the text mode of route information are also higher than any of the other use groups shown on Figures 9.6b, 9.6c and 9.6d. The Group IN-G1 users show a pattern of preference for the text mode of route information. The users in both groups IN-G2 and IN-G4 have low scores on PDA-D3, the text mode of route information (shown on Figures 9.6b and 9.6c respectively). However, users in Group IN-G4 have higher scores on schematic map mode (PDA-D1), compared with other modes within its group (Figure 9.6d) and in comparison with other groups. These users have much lower scores in both the text mode of route information (PDA-D3) and the detailed map mode (PDA-D2). On the other hand, the users in Group IN-G2 have high scores on both map modes. This might reflect the ways in which these users either use both map modes for their wayfinding tasks, or access the schematic map mode first then switch to the detailed map mode. Finally, the users in Group IN-G3 (Figure 9.6c) do have slightly higher scores in the schematic map mode (PDA-D1) and slightly lower scores in the text mode of route information (PDA-D3). Thus, the pattern in this group does not show any particular strong preference. These preferences are shown by users during real-time wayfinding activities and using their mobile devices for assistance. Although only limited modes of communication (but most commonly used) were studied here, it has demonstrated that users have their preferences on the way information is communicated in assisting their spatial tasks. However, dominant preferences for a particular mode of communication does not preclude using other modes in mobile situations, which is further explored in this case study and is discussed below.

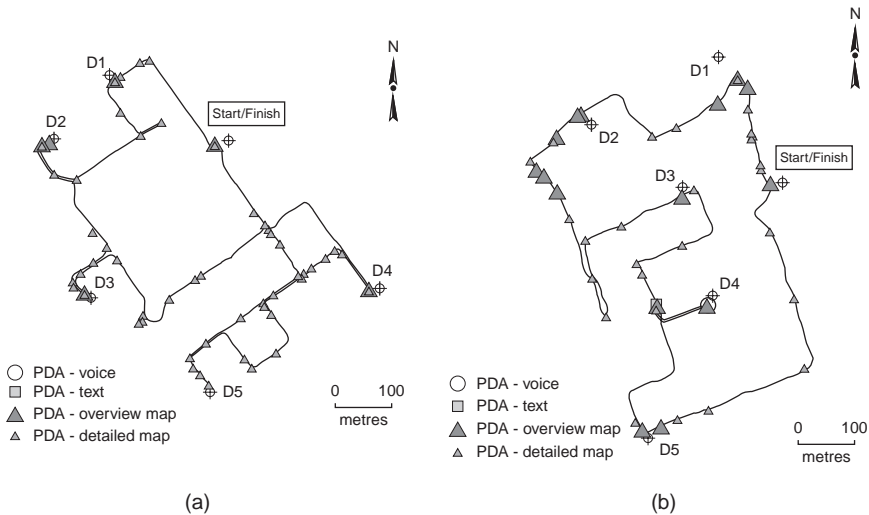
The user preference in the mode of communication can be influenced by the mobile situations encountered by users. In this study, the effect on user preference by different wayfinding situations along the routes was explored. Only two individual users, User1 and User2, from the main study are taken as examples here. User1 in both urban areas (Figure 9.7) showed a noticeable preference for the text mode of route information and is from Group IN-G1 (Figure 9.6a). However, this dominant pattern started to change when the user perceived that the coming tasks were more complex. At locations D2 and D3 in urban area U1 (Figure 9.7a), a switch between text mode of



**Figure 9.7** Modes of communication used in accessing wayfinding assistance by user1; (a) in urban area U1; (b) in urban area U2 (reproduced from Li, 2005).

route information and map mode is evident. The schematic map mode and detailed map mode were both used to assist in understanding the routes to be taken despite text route instructions being available. Another situation where this strategy of switching mode was used is where the user encountered challenging situations (e.g. the roundabout along the route from D2 to next destination D3 in urban area U1 shown in Figure 9.7a). Both map modes were brought into use. The text mode of route information was resumed as the main source once the difficulty had been resolved using map mode. A similar strategy is observable in the other urban area U2 (Figure 9.7b). The text mode of route information dominates the use of wayfinding assistance but at location D3 both map modes were used to assist the planning of the next task to the destination D4. After resuming the use of the text mode of route information, a further challenge arose partway between locations D3 and D4 necessitating the use again of map mode (Figure 9.7b).

The use of different modes of communication in wayfinding assistance for User2 is shown on Figure 9.8. User2 is from Group IN-G2 (Figure 9.6b), which is the map mode dominant user group with a particular preference for detailed map mode. As shown in Figure 9.8a, in urban area U1, there is a clear preference for the use of the detailed map mode. The use of the schematic map with overview is restricted to most starting points of each of the wayfinding tasks



**Figure 9.8** Modes of communication used in accessing wayfinding assistance by user2; (a) in urban area U1; (b) in urban area U2 (reproduced from Li, 2005).

(starting points D1 to D4). The user returns to consult the detailed map mode along the routes. This would indicate that schematic maps with overview are being used to gain knowledge on the configuration of the area when planning the wayfinding tasks and route choices. It also suggests that the detailed map mode (showing only part of the whole area, though they can be scrolled) does not provide users with sufficient knowledge of the entire spatial layout. In the urban area U2 (Figure 9.8b), User2 uses a similar strategy.

The study presented here on user preference for different modes of communication was carried out in the context of pedestrian wayfinding and based on the principles of spatial knowledge acquisition. The study provides a certain degree of understanding in the use of modes of communication in LBS and their dynamics.

## 9.6 Multimodal and Context-Aware Modes of Communication

The core challenge around modes of communication in LBS is to provide users with effective and efficient ways to access and use pertinent information and services in mobile situations and in real-time.

In LBS such information and services have to be communicated to users with the aim of assisting their spatial tasks and should be relevant to their location and activities in mobile situations. When considering mode of communication in LBS, it is necessary to take account of the dynamics in locations, users and technologies arising from the mobility of users, as well as the particular LBS application. Therefore, the appropriate mode of communication for users in LBS should be multi-modal and informed by context-awareness. Attention to both could have the potential to make modes of communication more effective and increase the utility of the information delivered. Issues discussed in Section 9.4 remain important areas of research. Moreover, the types of LBS applications aimed at also play an important role – in other words there is no ‘one-size-fits-all’ solution.

There have been studies of context-adapted cartographic visualization. It can be implemented by choosing appropriate scales automatically, deciding appropriate symbolization by algorithms and switching to text output when graphics are found to be inadequate for some devices (Gertner, 2004). Adequate forms of representation can be derived from parameters related to the type of device in use and to different users in different situations (Reichenbacher and Töllner, 2003). However, this context-adaptive approach poses challenges both to technology designs and in measuring and defining the factors that need to be adapted, particularly with regard to users in mobile situations. Context-adapted visualization is currently limited to adapting communication to user profiles that influence the presentation (e.g. font size, colour scheme and map elements), which can be implemented either by using a pre-established user list of preferences or by relying on input by users themselves at the time of access (Gertner, 2004). A third (and yet to be tried) approach would be to use agent technologies to learn user preferences and build up a profile over time based on user behaviour, situational contexts and expressed preferences. Situational factors, such as current light condition (day/night) and user travel speed, can also be incorporated into context-awareness visualization. For example, the in-car navigation system by TomTom<sup>®</sup> can change its map scale according to a car’s speed. The mobile space–time envelope approach discussed in Section 7.6 adapts the ‘soft clip’ of information content that might be presented to a user according to velocity and direction. Furthermore, as discussed in Section 9.5, the choice of mode of communication can change in real-time when carrying out tasks in different locations.

To consider the mode of communication with context-awareness with regard to locations, users and devices, the context-awareness

concept discussed in Chapter 7 and illustrated in Figure 7.2 can be applied. Environment, users and technology act as the three main strands of context in this conceptualization. These three strands are dynamic due to the mobility of the user and nature of LBS applications being used. Services provided by LBS should be able to respond to the user's interaction by adapting the content (Chapter 7) as well as the mode of communication by which information is presented and communicated. With the dynamic context awareness concept, mobility in LBS can be viewed as: locations where LBS are generally used; mobile users who are using LBS information and services; and mobility effects on the heterogeneous technologies which constitute the LBS infrastructure (Section 7.5). In other words, presentation and interaction methods in the mode of communication can be influenced by different environments, different users and by the various devices and technologies used in its delivery. Furthermore, the types of LBS applications aimed at, the various situations (related either to the environment or to users) and the activities being carried out can also have an effect on the adequacy of the mode of communication used in LBS. The most relevant factors from these contexts should be considered in providing users with an effective mode of communication. It is also essential to reflect on the user's understanding of the information presented. The discussion on spatial knowledge in Section 9.5 has implications for this.

Multimodal communication with context-awareness (MMC + CA), in general, enhances usability and provides flexibility in interaction. Some of the different forms of multimodal communication were discussed in Section 9.2. MMC + CA allows users and systems to choose single or combined modalities according to the user's preferences, the situation encountered, the tasks being carried out and other social and physical contexts. MMC + CA can also provide alternatives both in presentation and interaction where there is a need to reduce the effects caused by the limitations of mobile devices, such as the size of display screen and the speed of data transfer. MMC + CA offers a more open and flexible approach compared to automated solutions with predefined assumptions. It also allows more interaction for information flow between service provider and users, which could reduce misunderstanding and ambiguity. Multimodal communication should be considered carefully with regard to the express need for context-awareness in LBS.





# Chapter 10

## The Business of Location-Based Services

### 10.1 Introduction

---

In the foregoing chapters of this book, a number of technical areas that are fundamental to LBS have been looked at. Bringing these all together into applications that can be marketed to generate a worthwhile return on investment is another matter. This chapter, the last one in the book, therefore focuses on a number of social, legal and business issues that are going to be important when bringing applications to the market. LBS are a new emerging market in which there are few tried and trusted business models and applications. The industry, for the most part, is still strong on hype, short on delivery and customer uptake. This is due in part to the complexity of the value chains involved in LBS, from positioning systems to selling suitable mobile devices – much more so than for many other markets. To explore these, the next section considers LBS as an emerging sector within the (world) economy and how its pattern of development may reflect the emergence of other ICT-based sectors. As discussed in Chapter 2, there has been considerable technological development and convergence that has made LBS possible as a heterogeneous technology. LBS are thus in a position to inherit many of the business models that have emerged since the mid 1990s for e-Commerce (e-Business) but, as will be discussed, the mobility of the consumer requires special consideration. From this sector overview, the emergence of specific products and how to build a sustainable market

share are then considered. This is not intended to be a primer on business and marketing *per se*, and those readers raring to launch an application are recommended some salutary wisdom from Arbor and Bjerke (1997), Foxall (1997), Agar (2004) and Williams *et al.* (2005). Finally, there are three sections on key areas that can be broadly classified as standardization issues, legal issues and social issues.

## 10.2 Emerging Sectors

---

Doing business with technologies has been around for millennia. Doing business on the Internet (e-Commerce) has only been around since the mid 1990s. Conducting transactions from mobile devices (mobile e-Commerce or MeC) is by comparison still very much at the starting blocks. The largest MeC sector to date has been the purchase and download of ring tones; but this is not to trivialize MeC. Just as e-Commerce has drawn on traditional business models and evolved new ones, so MeC has drawn on e-Commerce business models and, in view of some of its unique aspects, will need to evolve new ones. LBS providers will thus have a range of existing models to draw upon, including growing insights into MeC. It is, therefore, worthwhile looking at some key features of these models as they might relate to LBS.

### 10.2.1 Internet-Based Business Models

A business model is understood here as being ‘a logical architecture for product, service and information flows, including a description of the involved business actors and their roles, as well as sources of revenue’ (Tsalgatidou and Pitoura, 2001 p. 225). Business transacted over the Internet has had a huge impact on post-industrialized economies. From the perspective of a consumer, there are broadly two models that are directly recognized: those services which we access for free and those for which we need to pay. So for example we, the public, can use Google, Mapquest (Figure 3.1) or ThompsonLocal.com (Figure 4.5) all for free, whilst if we wish to acquire a product say from Amazon or through E-Bay we are expected to pay for the purchase. Services that are free to the consumer still need to generate revenue, if not a profit (or contribute to profit elsewhere within the same business), and any

user of the Internet will recognize revenue generation, for example through banner advertising. Underlying this initial crude classification of Internet services are a number of business models:

- **Hosting.** Individuals and small organizations usually don't wish to go through the complication (and cost) of setting up and managing their own Internet servers and dedicated lines in order to get access to e-mail and have Web pages available over the Internet. Instead, they are willing to subscribe to a company for the use of their servers and high speed connections to route e-mail to/from their local PC, to host their Web sites and provide domain names. These companies can also provide broadband connections to their users in competition with telephone companies. This type of business has, since the late 1980s, mushroomed into an industry in its own right.
- **Application Service Provider (ASP).** This is where a company holds or obtains licenses for applications software, and rents or leases access to those applications out to its client users. This can be as complex as facilitating e-Commerce for smaller organizations by providing access to databases, PHP (PHP Hypertext Processor) and Web activity analysis tools (e.g. WebTrends), to something as simple as some Java code that will enhance some aspect of functionality on a mobile phone. The type of ASP e-Commerce services just cited are usually offered in conjunction with hosting as integrated packages.
- **E-shop.** Probably the best known example of an e-shop is Amazon, which started by selling books over the Internet but then expanded into, CDs, DVDs, electronic goods and so on. It is now a global business. The concept of shopping on-line for just about any product that can be found in a high street shop or agency has become commonplace. The purchases are sold and delivered by post or courier against a credit/debit card payment. Whilst e-shops promote convenience (in the comfort of your own home 24 hours a day, delivered to your door) coupled with easy text-based searching, and have certainly drawn customers away from the high street, they have not yet re-created the social experience of shopping. Indeed many bookshops have enhanced the social experience of browsing and buying books by installing or

leasing space for coffee shops, providing easy chairs and so on (as well as having parallel e-shopping) as a means of staying in business.

- **E-auction.** The example of an e-auction house *par excellence* is E-Bay. The business model centres on automating over the Internet the traditional approach to bidding. Sellers can also submit descriptions of their goods via the Internet. Some e-auction sites support the payment and delivery process. Sites like E-Bay, however, have developed a trust system whereby customers can vote on the quality and reliability of services from suppliers; the risks nevertheless remain with the buyer.
- **E-market place.** This is similar to the e-auction model, except that goods are offered for sale at fixed prices by suppliers (or even for free as long as the buyer collects). It is more common here for the e-market provider to handle the payment transactions on behalf of the suppliers. A trust system of suppliers can also be implemented. Both e-auction and e-market place make revenue through commission and/or from advertising placed on their sites.
- **Information brokerage.** With the exponential rise in information available over the Internet, it is not surprising that a wide range of information services should emerge. Dominant among these are on-line search engines. The market leader is currently Google, but it was by no means the first in the market. Yahoo, AltaVista and AskJeeves were early entrants. This segment relies heavily on advertising revenues and fees for preference listing in order to generate profits. These companies harness the free accessibility to information on the Internet through their Web crawlers in order to populate their search engine databases (Figure 5.1) in an automated way and at very little cost. This is archetypal post-industrialism – using information both as raw material and as product (Section 1.2). Another segment is the trade directory, such as on-line yellow pages, where businesses pay subscription fees and for enhanced visibility in order to attract potential customers.
- **Portals.** This is where a number of complementary services, offered more often than not by different companies, are brought together through a uniform interface that adds value to the services. On the one hand this can minimize the cost of Internet visibility to the separate companies and

increase their sales, and on the other the customer can make multiple purchases that are optimally configured for them. Typical amongst this segment are the on-line travel agents (e.g. Expedia and Octopus) that allow customers to integrate their purchases of flights, hotels and hire cars (all from different providers) through the one interface and payment system.

- **Value-chain integration.** This model brings together a number of parties that collaborate in a supply chain to produce a final good. The product of one party becomes the raw material for another. Each party in the chain performs a sequence of activities, either in parallel to other activities or on partially finished products from 'upstream' parties, that are passed 'just-in-time' to the next party in the value chain. This requires the parties to share information pro-actively, such as being able to query each other's databases remotely so as to know when orders have been placed 'downstream' or that partially finished products are being shipped from upstream. The obverse of this model is *outsourcing*, where one party no longer wishes to produce an entire good itself but wishes to focus on particular aspects of production and instead sets up a supply chain producing aspects or components of the finished good to its specification. This is typical of how the automotive industry is now organized, where the car maker only produces some key components but where, nevertheless, all components need to be available at an assembly line without hold-ups for un-delivered components or the need to stockpile components.
- **Trust services.** These are services that ensure safe transaction can occur across the Internet. They offer an interface with the banking system for, say, the processing of credit card payments as well as encryption.
- **Collaborative and participative content.** This has been a very substantial growth area in the new millennium. It encompasses on-line social networks (e.g. Facebook, Second Life, MySpace), collaborative gaming (e.g. EverQuest), repositories of on-line content (e.g. YouTube and Flickr) and collaboratively produced content (e.g. Wikipedia). Some social networking and gaming sites have been so successful that they have established their own internal economies with the sale and purchase of virtual goods and services. There are now on-line brokers through whom virtual money can be exchanged into real money and vice versa.

There will, of course, be hybrid or diversified business models that have a combination of the above. A good example of this is Amazon, which has evolved to incorporate e-shop, e-market place and e-auction. LBS providers thus have a good menu of tried and tested models to draw on. It is not difficult to see LBS providers forming hybrids of ASP, value-chain integration (particularly in bringing together all the components of LBS architecture in Figure 4.8 into a seamless service), information brokerage and portals.

### 10.2.2 Implications of Using Mobile Devices

All the above Internet-based business models have developed around fixed terminals, or at least PCs and notebooks on stable networks. The use of mobile devices brings about a number of new challenges that need to be addressed within MeC, some of which have already been discussed in previous chapters. These challenges are derived both from the device itself and from using wireless telecommunication networks.

Clearly mobile devices need to be small and light in order to make them portable in an every day, all the time sense. Since the early 1990s mobile phones have shrunk from being brick size to a size that will easily fit into the palm of your hand. This has inevitably led to small screens and although LCD technology has vastly improved, still presents problems, for example in the display of legible maps without the over-use of zoom and pan. Keypads are also small and need to be multifunctional. A mobile phone typically has only 14–16 keys and a tiny cursor control in order to access all the diverse functionality. Performing transactions from such a device requires careful design of the interface in order to make the process intuitive without cognitive overload, and should certainly be mindful of the fact that unless a speaker or Bluetooth earpiece is being used, the screen cannot be viewed or additional keys pressed whilst the device is placed against the ear to hear spoken instructions.

Again, whilst chip technologies continue to obey Moore's Law by inexorably increasing in power and reducing in size, the memory and processing resources of a mobile device are very much lower than that of a PC or notebook computer. Applications such as LBS can't rely on the mobile device having lots of memory to store large extents of mapping or sufficient processing power for, say, map generalization. The finite energy resource of the battery is probably the biggest limitation facing mobile devices. Whilst energy used during standby is

generally very small, batteries can quickly become depleted if continually receiving data or if the device has to continually transmit its location. In general, battery consumption is higher when sending data than receiving it. Mobile devices are generally not designed for multitasking. It is thus quite common that if one is, say, writing an SMS text message and a voice call comes in to the device, the text editor will be switched off in order to handle the in-coming call. This can have potentially serious implications for using mobile devices for transactions. Finally, mobile devices are more easily lost or damaged than fixed PC links, leading to loss of important transaction data and any user profiles stored on them.

Implications also arise from the nature of wireless telecommunication networks. Firstly, a mobile device may not always be communicating with the network. Interruptions may arise when entering tunnels, on underground trains or in remote areas. Interruptions through loss of connection can occur when hand over is being made from one cell to another, more so when the user is travelling at speed. Such interruptions at a critical time in the transaction are likely to lead to considerable uncertainty as to whether the transaction has taken place and may even mean that the transaction has to start afresh. Wireless communications are currently much more error prone than fixed link Internet connections. This means that MeC transactions require more redundancy in the data coding, which has implications for data transmission rates and battery consumption. When roaming, signal strength and available bandwidth are bound to vary, so that the design of LBS architectures cannot be based around the performance characteristics of any one single network and may even need to take the prevailing network conditions as context in framing a response to a user.

### 10.2.3 Mobile Device-Based Business Models

Many of the business models outlined in Section 10.2.1 are being explored in relation to MeC. Clearly banner and other pop-up advertising have limited potential for revenue generation on small screens. Instead, many services are offered at an enhanced (even premium!) call or text rate that is shared between the service provider and the network operator. Network operator's business models, given the current saturation of the mobile phone market in the United States, Japan and Europe for example, are strongly geared towards customer



retention as well as expanding their market share. In the United Kingdom, for example, this has meant that many of the operators give the mobile phones away for free as part of signing up for minimum period contacts (typically 12–18 months).

One business model adapted for MeC that has considerable potential for LBS is a publisher/subscribe model (Chen *et al.*, 2003; Leung *et al.*, 2003; Brimicombe and Li, 2006). Within this model is a framework in which publishers produce information and subscribers consume it. The service broker between publishers and subscribers is the LBS provider, with whom subscribers maintain a subscription for services. Service brokers track their subscribers through location-aware mobile devices (e.g. mobile phone, in-car navigation system). What is further required is an event broker or trigger mechanism that can automatically establish that the area of interest of a publisher overlaps with that of a subscriber and thus sets in place an exchange of information. In Section 7.7 and Section 8.7.4 the concept of mobile space–time envelopes was discussed. These can act as event brokers within a publish/subscribe framework for LBS, that is they can be used by the LBS provider as a trigger to geographically ‘soft clip’ objects that fall within a subscription domain in anticipation of queries, and as a trigger for pushed alerts from publishers as subscribers come within range. The envelopes themselves are geographical areas of changing shape and size that surround a subscriber and act as the event broker for that individual.

The use of mobile space–time envelopes in this way can be stated algebraically (e.g. Leung *et al.*, 2003). Define  $M = \{m_1, m_2, \dots, m_k\}$  as a set of multidimensional metadata spaces that categorize the information content of a publish/subscribe system. Each metadata space is defined as a tuple  $m = (D_m, V_m)$ , where  $D_m = \{d_1, d_2, \dots, d_n\}$  is the set of data or subject dimensions (e.g. coordinates, time, attributes) that exist within that space, and where  $V_m$  is the set of values for those dimensions. Define  $dom(d)$  as the domain of values for each dimension  $d$  and the domain of values of a metadata space can be taken as  $V_m = \prod dom(d)$ , the Cartesian product of the dimension domains. Within each metadata space  $m$ ,  $r = (C_r, V_m^r)$  is a data region where  $C_r = \{c_1, c_2, \dots, c_n\}$  is a set of constraints and  $V_m^r$  is the set of values, the domain, of region  $r$  with respect to the metadata space  $m$ . A region is valid if  $V_m^r \neq \emptyset$ . Publishers and subscribers may have a common interest in a region (e.g. traffic news, the weather forecast, the price(s) of real ale at the George & Dragon pub); publishers to provide, subscribers to consume. Note that spaces and regions may not necessarily be spatially defined.



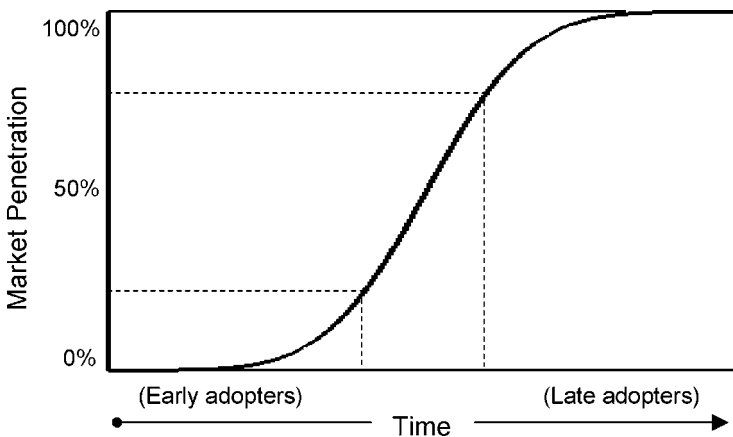
If a region defined by a publisher is denoted as  $p = (C_p, V_m^p)$ , that is including conditions of sale, use and copyright as well as data content; and that of a subscriber as  $s = (C_s, V_m^s)$ , that is including service levels, preferences and conditions of notification, then if  $V_m^p \cap V_m^s \neq \emptyset$  it would indicate that the subscriber would derive utility in accessing the publisher's information. For the event to be brokered spatially a subscriber needs to be within range with respect to speed and direction. If, for example, only Euclidean distance is used (say, a 10 km threshold from the George & Dragon pub), then a notification would have little utility, and would likely be a nuisance, if the trajectory of a subscriber is outwards towards the threshold (moving away from the George & Dragon) rather than inwards from the threshold (moving towards the George & Dragon). So, if a mobile space-time envelope is denoted as  $e = (C_e, V_m^e)$ , notification can occur if  $V_m^p \cap (V_m^s \cup V_m^e) \neq \emptyset$ , that is the union  $V_m^s \cup V_m^e$  provides the subscription region with a defined spatial envelope that brokers the range trigger for a notification. The shape of the mobile space-time envelope, modified according to the trajectory of the individual, provides additional geographical intelligence and helps assure the level of utility of a notification. This is one important way in which business models can take advantage of the mobility of potential customers.

## 10.2.4 Adoption and Hype Curves

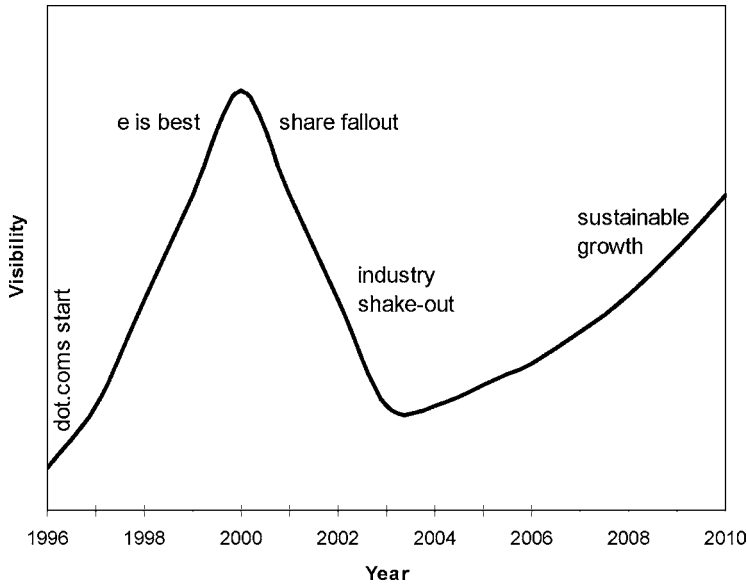
The creation of new market sectors, their growth and eventual maturation to market saturation or even ubiquity, follows a well-trodden path that can be modelled. A new market sector is typically started by a single supplier, the company that made the invention or which holds the patent, though if several innovators launch into a new market simultaneously there may be a small number of suppliers. Numbers of sales will initially be low, typically bought by customers who can be classed as early adopters – those willing to take risks on new products. Because production is small and sales low, the product is usually expensive. As the number of users gradually grows, as the product is marketed, so other companies – imitators – see the potential in the market sector and enter with similar, often cheaper products. Sales start to accelerate; production increases to meet demand and there is increasing price competition as more producers enter the market. From the consumer side, there may start to be peer pressure to own what is rapidly becoming a 'must-have'. Eventually, though, sales start

to grow at a diminishing rate as the market starts to reach saturation with purchases typically by late adopters. This is a typical scenario for new technologies from the telephone to the MP3 player. When USB flash drives (data sticks, cyberkeys) were first introduced in about 2000 they were a novelty, only a few companies produced them, they cost £1.00 per Mb of storage and not many PC users owned one. By 2008 they have become almost ubiquitous amongst PC users, there are many suppliers and can be purchased for about £0.02 per Mb. The model that describes this pattern of adoption is the *logistic curve* (Figure 10.1). This is an S-shaped curve reflecting the slow initial take-off, almost exponential accelerating usage and then final slow-down as maximum possible market penetration is approached. It is quite expected that adoption of LBS technologies will follow this model (Shiode *et al.*, 2004).

Another phenomenon that can also be observed with the entry into the market of a new sector is the hype curve. Often exaggerated claims are made about completely new products and services to ‘talk up the market’, attract investors and so on. This can create a bubble of over enthusiasm which bursts once it is realized what the true potential and growth of the market is likely to be. This is usually followed by a shake-out of producers (bankruptcy, mergers and acquisitions) leading to a more rational longer term growth. This has happened a number of times with new technologies, most recently in the dot.com boom as illustrated in Figure 10.2. The dot.com was all about e-Commerce and how ‘e was best’ no matter what the nature of the



**Figure 10.1** Logistic curve modelling a typical new market sector.



**Figure 10.2** Hype curve for the dot.Com bubble.

business or its consumers. Huge amounts of venture capital were invested and share values on the Nasdaq soared to dizzy heights. The crash in 2000 was spectacular but has led to more prudent and sustainable growth in the sector. LBS have had their own share of hype and it is have hoped that they can avoid the excesses of the dot.coms.

## 10.3 Emerging Products

Once a particular sector (e.g. LBS) is recognized as being capable of generating substantial revenues, one or more suppliers are going to enter the market with specific products. Each supplier is going to try and maximize its market penetration for the particular market segment that its product targets. The way suppliers tend to operate is to increase cumulative sales by introducing successive generations of products and by tailoring their product range to different segments. To enter a market segment, a supplier must at least have a *selling proposition*. For example, for a SatNav it may be as simple as ‘with SatNav X you will never get lost again’. But most SatNav suppliers could make the same claim for their own product. A supplier may therefore have to go a step further by having a *unique selling proposition* – a claim that is unique or at least

novel to their product. This may take the form: ‘with SatNav X and our own network of traffic sensors, you will never get stuck in traffic again’. Even this claim (hyped as it is!) may not be sufficient to entice the public to buy in. What is then required is a *value-added selling proposition*, which builds onto the unique selling proposition some sort of value-added feature for the buyer over competing products. Without too much stretch of the imagination just such a proposition could be made for SatNav X: ‘with SatNav X and our own network of traffic sensors, you need never get stuck in traffic, be late for meetings or even be late for a date ever again’.

Of course, in the world of marketing many value-added selling propositions are frequently more subtle and certainly snappier; examples that come to mind are ‘Guinness is good for you’ (United Kingdom), ‘Du beau, du bon, Dubonnet’ (France), Coca Cola’s ‘The real thing’ and American Express’ ‘Don’t leave home without it’. Sometimes, of course, consumers can work it out for themselves without recourse to slogans: the value-added proposition of USB flash drives hardly needs to be stated for PC users (the target segment) to recognize them as being (currently) the universally most convenient and cost-effective way of carrying data around.

And what will be the value-added selling proposition for LBS? No doubt there will a different one for each product type as they come on offer, but for LBS in general there should be two broad elements. The first is that there will need to be a proven case that LBS products are better than carrying paper (e.g. conventional maps, print-outs from PC-accessed Internet services). The second is that the results to queries will need to impress or afford an element of surprise to the user, that is provide a response for which the user’s perceived utility exceeds the cost of the service within the time taken. For a snappy slogan one might use those comforting words on many street plans: ‘You are here,’ coupled of course with: ‘You’ll soon be there.’

Two other important tactics to be used with emerging products are to have successive generations of products each with their own life cycle and to have different versions of a product suited to different market segments. Having successive generations of products, each an improvement on the last, allows the supplier the opportunity to ramp up cumulative sales. If each product follows the logistic curve in Figure 10.1 over its life cycle, then there is a point in time when the growth in sales slows – at the inflection when the curve starts to flatten out. At this point about 70% to 80% of those consumers likely to buy into the product have already done so. If at this point the next generation of the product

is introduced it opens up the opportunity of resale to those who have already bought in and to market the product more widely. This next generation product also has a logistic curve but piggy-backed onto the first generation product and should result in higher cumulative sales. A clear example is the generational development of Microsoft's operating system from the original command line DOS through to Vista.

Market segmentation recognizes that not everybody wants the same thing and that within a market sector it is possible to divide up potential consumers into a number of different groups (segments) such that the wants within a group are broadly similar and can be profiled. Segmentation by geography is known as *geodemographics*, and the reader wishing an in-depth treatment of this topic is referred to Harris *et al.* (2005). That a particular market sector can be segmented has been long recognized and is now a fundamental aspect of product positioning and marketing. Targeting the exploitation of each segment with different versions of a product is a way of increasing sales. Thus, as SatNavs have grown in popularity in the United Kingdom as a mass consumer product since 2005, so product differentiation has increased largely on price (reflecting the range of functionality, data content, voice options, screen size, etc.) to make it a 'must have' across all socio-economic segments of vehicle drivers. The purchasing behaviour of each segment tends to be analysed carefully since the existing behaviour of a segment is often the best predictor of future behaviour (Foxall, 1997).

One important issue that needs to be borne in mind when designing new and innovative ICT applications, is the effect of social appropriation and shaping of new technologies (Williams *et al.*, 2005). It is often the case that the eventual use and utility of new technologies is far removed from what the original developers had in mind. A point in case is the mobile phone. Developed exclusively with business users in mind, young nonbusiness users have largely appropriated the technology as an independent means of communication and as a status symbol. These users have steered producers, through market forces, into adaptations of both design and functionality, and have created a diverse industry (including ring tones, wallpaper, screen savers and mobile gaming) unimaginable by the initial developers (Agar, 2004). Although LBS are currently viewed as a niche application, it is likely that they will only become generically ubiquitous through a process of social appropriation that will act to mould them to mass consumer demand. What LBS will look like and what services might be offered at that stage is almost impossible to second guess now, though the underlying fundamentals of handling and visualizing spatial data will most probably stay the same.

## 10.4 Standardization Issues

---

LBS are heterogeneous technologies spanning a number of cognate disciplines (e.g. GIScience, electronics, computer science, manufacturing) and require the interoperable use of existing systems as well as introducing new ones. On the one hand, for interoperability to occur and on the other to have reliably engineered services, there must be standards and adherence to standards. A standard is a formally agreed document which establishes a norm or the means of adherence to a norm. Thus a standard can be a technical specification (how to produce something), a method of testing (how to determine the property or performance of some substance or manufactured object) or a procedure (how to operate some equipment or installation). Standards can be legally mandated by governments where formal regulation is deemed necessary, but often they are developed and agreed upon voluntarily by an industry in order to overcome a technical barrier, some aspect of incompatibility or lack of interoperability that is likely to hold back that industry in some way. A standard may simply arise through *de facto* practice that becomes accepted as the norm. It has often been said that the problem with standards is that there are so many of them!

Many countries have their own national standards organization. In the United Kingdom there is the British Standards Institution; a relevant example of a standard being BS 7666: Spatial Datasets for Geographical Referencing. This was an early standard applied to the activities of GIS, first issued in 1993 and revised in 2000. Adherence can be purely voluntary, but if for example a list of postal addresses is said to conform to BS 7666 then its structure will have been standardized, all the elements of an address will be present and its content can be clearly understood and used with confidence. Other standards are incorporated into legislation such as in the support of building codes, signage on highways and so on where safety becomes an important issue. British Standards have some caveats to be noted which reflect the fact that standards are technical in nature:

It has been assumed in the drafting of this British Standard that the execution of its provisions is entrusted to appropriately qualified and experienced people.

A British Standard does not purport to include all the necessary provisions of a contract. Users of British Standards are responsible for their correct application.

**Compliance with a British Standard does not itself confer immunity from legal obligations (original emphasis; BSI, 2000).**

The British Standards Institution also represents the United Kingdom at an international level in Europe (CEN standards) and at the International Organization for Standardization (ISO). Here there is both a need to participate in drafting international standards and in ensuring national harmonization. An example of a standard which was first agreed internationally and then brought into the United Kingdom as a British Standard was ISO 19109 (BS ISO 19109), Geographic Information – Rules for Application Schema, which certainly has relevance to LBS. Indeed, ISO Technical Committee 211 (ISO TC211) specifically focuses on developing standards, issued as the ISO 19100 series, relating to the geo-information area. Another key standard relevant to LBS is ISO 19115, Geographic Information – Metadata, the US equivalent being FGDC-STD-001-1998, Content Standard for Digital Geospatial Metadata. Such standards may be adapted locally for implementation within particular sectors. Thus ISO 19115 formed the basis within the United Kingdom for the development of an e-Government Metadata Standard (e-Government Unit, 2006) as part of the UK e-Government Interoperability Framework, which also incorporates aspects of ISO 15836, Dublin Core Metadata Element Set, for expressing metadata in HTML. The ‘Dublin Core’ (named after a workshop in Dublin, Ohio in 1995) is a fifteen element generic vocabulary for use in metadata descriptions and conforms to notions of best practice for the Semantic Web (<http://dublincore.org>).

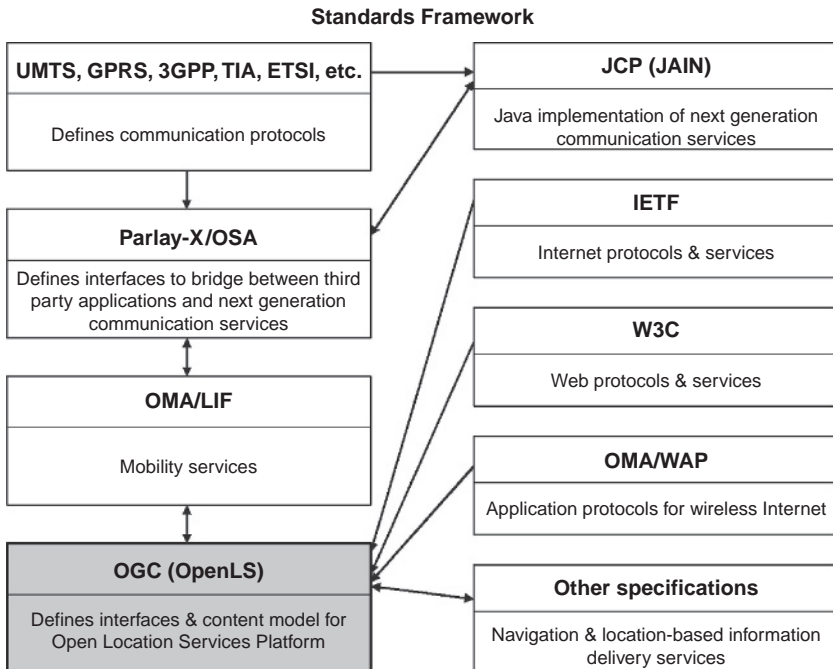
Two examples of an industry coming together to create its own standards in the context of LBS are the GSM Association and the Open Geospatial Consortium (formerly the Open GIS Consortium). The GSM Association has promulgated a nonbinding standard (permanent reference document) for LBS (GSM Association, 2003). The objectives of this standard were to brief the industry at large of the nature of LBS, to develop guidelines for interoperability between operators and to provide operational guidelines ahead of more formal standards being introduced. It covers definitions, system architecture, principles of privacy, positioning accuracy and the principles of charging where multiple operators are involved (due to roaming).

In Section 3.3.2 the concept of open systems was introduced and how the main players within the GIS software industry had come together in 1994 to form the Open GIS Consortium (OGC) (<http://www.opengeospatial.org>). OGC works closely with ISO TC211 and



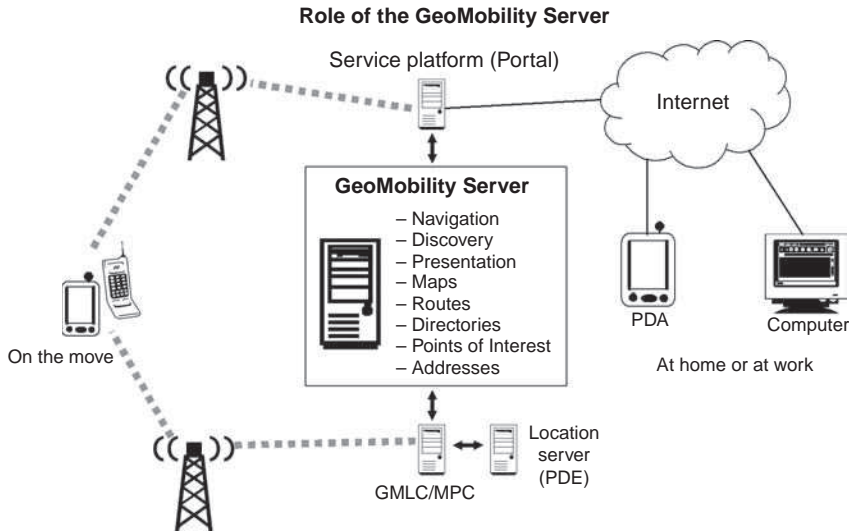
## Location-Based Services and Geo-Information Engineering

the US Federal Geographic Data Committee (FGDC). OGC's OpenLS (OGC, 2004) is a specification for LBS middleware (middleware is the term given to reusable software components that support the rapid development of distributed services and applications). As can be expected in a heterogeneous technological environment, OpenLS has to sit conformably with a range of other standards as illustrated in Figure 10.3. The core feature of OpenLS is a GeoMobility Server (GMS) which works in conjunction with an LBS provider's portal to act as middleware to manage the geographical content of queries and to assist in fixing the location of the user. The role of the GMS in an LBS application is illustrated in Figure 10.4. In Chapter 4, in discussing LBS architecture, a differentiation was made between the LBS provider and the positioning services provider, whose key activity is to fix location of users and targets as a service to LBS providers. In Figure 10.4 the GMS acts as the LBS provider's middleware in accessing the gateway of the positioning services provider (denoted as GMLC – Gateway Mobile Location Centre) in order to obtain a user's location. The GMS is also



**Figure 10.3** Standards framework for OpenLS (based on OGC, 2004: <http://www.Openspatial.Org>, all rights reserved <http://www.openspatial.org/legal/>).





**Figure 10.4** The role of the GeoMobility server in OpenLS (based on OGC, 2004 <http://www.OpenSpatial.Org>, all rights reserved <http://www.openspatial.org/legal/>).

envisaged as carrying out a number of core services with regard to solving LBS queries. This includes finding relevant points of interest, determining travel routes and creating maps.

Finally, *de facto* standards emerge from the popular and widespread use of a particular technology, which only later is formalized through a standard. An example of this is SQL (Chapter 8). Others are Bluetooth (Chapter 2) and Java (for Web applications) and its variant J2ME for mobile applications, which also has a Location API. Such is the technological heterogeneity of LBS and the need for interoperability and seamless communication, that any potential service provider will need to be conversant with and work to a very wide range of standards, though the availability of compliant middleware is likely to ease the task.

## 10.5 Legal Issues

LBS providers will enter the market in order to make a return on their investment; otherwise the business is unlikely to be sustainable. To this end, any provider will need within its business model to protect its rights, ensure the noninfringement of the rights of others and to limit its

liabilities and other risks. In other words, since every legally constituted business is registered within a jurisdiction it will need to be cognisant of the legal regimes of the country where it is registered and all those countries where it chooses to operate. This is a very large and specialist topic; here the relevant issues of patent, copyright and liability will be considered. For a much fuller treatment, the reader is referred to Cho (1998, 2005).

### 10.5.1 Patents

Patents are the oldest form of protection offered to inventors to safeguard their intellectual property rights. They typically apply to new inventions (e.g. objects, machines, designs) or to new industrial processes (e.g. method of making something). They do not apply to computer programmes *per se*. A patent typically lasts for 16 years and provides a patentee with a temporary monopoly on their invention – time enough to exploit the invention (or authorize someone else to exploit it) and make a financial return on the invention without competition. A patent applies only to the jurisdiction (country) where a patent is taken out. Getting protection worldwide for an invention is both a complex and costly exercise. A patent can only be granted by disclosure of the specification of the invention, which must not previously have been publicly disclosed. Once disclosed, the patent issuing authority (in the relevant jurisdiction) will have the validity of the invention checked and, once a patent is granted, the specification can be accessed by the public but cannot be infringed for the period that it is protected.<sup>1</sup> The validity of an invention for which protection is sought generally rests on the following criteria (Cho, 1998):

- an invention must display a degree of novelty, there having been an inventive step on the part of the patentee;
- the invention must not be obvious with regard to what is already known;
- there can be no prior claim to the invention;
- the invention must be useful in as much that it can perform as claimed in the specification and is replicable (i.e. someone

---

<sup>1</sup> Albert Einstein's first employment (1902–1909) was undertaking just such work as a clerk in the Swiss Federal Patent Office in Bern. This period coincided with his formulation of his special theory of relativity – the most creative period in his life (Miller, 1989).

using the specification can produce the object or make the process work);

- there are no ‘internal objections’ that might arise due to lack of clarity or succinctness, insufficient description or presence of ambiguity in the specification.

Whilst the granting of a temporary monopoly may not appear to be in consumers’ interests (after all, many countries have legislation to prevent monopolistic practices) and may result in consumers paying artificially high prices for the product, there is a certain logical rationale to patents that has allowed them to remain on the statutes. These are (Ricketson, 1984):

- that each person may have rights in their ideas and that these ideas, being intellectual property, should be afforded some protection from being taken by others;
- that in creating an invention, the inventor is providing a service to society and should therefore be able to seek compensation for the time and effort involved;
- since industrial innovation is desirable within a society, the profit expectation afforded by a patent provides an incentive for investment towards an invention;
- since it is desirable within a society that inventions are not kept secret, thereby inhibiting the pace of development, the granting of a temporary monopoly is a reward for disclosing the specification;
- that the publishing of specifications as part of the patenting process is invaluable information to researchers.

A key patent that affects LBS is US Patent No. 6,240,360 issued on the 29 May 2001 (<http://patft.uspto.gov>). The patentee is Sean Phelan, who trades as Multimap.com (Figure 4.6; also <http://www.multimap.com>); the title of the patent is ‘Computer system for identifying local resources’. The abstract to the patent states:

A map of the area of a client computer is requested from a map server. Information relating to a place of interest is requested from an information server by the client computer. The information is superimposed or overlaid on a map image at a position on the map image corresponding to the location of the place of interest on the map. The information (or ‘overlay’) server may contain details of, for example, hotels, restaurants, shops or the like, associated with the geographical coordinates of each location. The map server contains map data, including coordinate data representing

the spatial coordinates of at least one point on the area represented by the map. *[numbers referring to a diagram have been omitted.]*

Basically, this covers the process of requesting from a client computer (which could also be construed as a mobile device) both maps and other information (points of interest) for a specified area which can be viewed as superimposed on each other. The map data are stored on one server (the map server) whilst the points of interest are stored on another server (the information server). This patent appears to monopolize the process of Internet mapping (Cho, 2005) and any similar process employed in delivering LBS. Any other system of delivering maps and points of interest (such as in Figure 3.1 and Figure 3.2) either over the Internet or to mobile devices must either carefully ensure that the patent is not infringed or must pay (negotiated) royalties to the patentee. The existence of this patent will serve either to prevent the development of products and services based on this fundamental process of Internet mapping (until the patent expires) or will spur on others to devise alternative processes. Radcliffe (2003) is of the opinion that the patent does not meet the validity criteria in that the process is not a novel invention in that it merely automates an existing manual process and that there was prior art pre-dating the filing of the patent request. However, since the patent has been granted it can only be revoked by a US court ruling should anyone wish to mount a challenge. As of March 2008 there were 66 414 US patents returned for the Boolean selection: location AND based AND services.

### 10.5.2 Copyright

Copyright provides for protection against unauthorized copying of material (including electronic data) by third parties. It is a key pillar in the protection of rights in intellectual property and is likely to impinge on most aspects of LBS. A typical copyright statement is as follows (Brimicombe, 2003):

© 2003 Allan Brimicombe

All rights reserved. No part of this book may be reprinted or reproduced or utilized in any form or by any electronic, mechanical or other means, now known or hereafter invented, including photocopying and recording, or in any information or retrieval system, without permission in writing from the publishers.

Copyright confers protection for the author for their lifetime and, in the European Union, for 70 years after their death (in other jurisdictions this may be 50 years). Where an 'author' is not identified then the protection is for 70 (or 50) years after its creation. Copyright infers protection for the expression of an idea rather than the idea itself and is immediate from the moment of expression in a material form without formality. Acceptable media of expression for copyright protection to arise is wide, including handwritten or drawn documents, printing on paper or other material and digital encoding on electronic media. Works can be literary, musical, artistic or indeed computer programs and database content, provided they are original. The term 'original' does not necessarily mean that the work shows novelty or innovativeness but rather that the expression of the work has been produced by the skill and labour of the originator. This is why database content and their representation as maps and other forms of visualization can be protected under copyright; hence the importance of copyright compliance in LBS.

Copyrighted material can be used by third parties on payment of royalties to the copyright owner, or indeed the owner can sell or otherwise transfer the right to others. There is, however, scope for 'fair dealing' under copyright legislation. There can be copying of a small proportion of a work (typically 10% of a work) without permission for research or study, review and news reporting. There can also be the reproduction or adaptation of computer code necessary for the development of an interoperable product. Purchasers can also back-up software (i.e. making an electronic copy for security). Software and data vendors have gone to considerable trouble to protect their products from piracy and unauthorized use in order to safeguard their ownership and income streams. From the use of key codes through dongles to on-line registration and activation, suppliers have sought increasingly tighter ways of enforcing copyright.

One way in which copyright is enforced for data is through digital watermarks. These have been particularly effective for map and imagery data. For vector map data, slight changes to certain point, line or polygon features can be introduced in an audited way in order to provide proof of ownership if this should become necessary in court. For images, a watermark can be achieved by slightly altering the value (colour) of some pixels so as to create a unique and auditable pattern that can be read back even if the image is copied, say, from the Internet. Some providers, such as the Ordnance Survey in the United Kingdom, are anxious to rigorously protect their copyright and hence

their revenues (Ordnance Survey sued the Automobile Association for \$30 million for breach of copyright (Longley *et al.*, 2005)). Others, however, are happy to see much freer access to and use of their data (e.g. the United States model) as a form of public good that can stimulate the economy. There still needs to be a statement of copyright associated with these products to prevent others from copyrighting the data. Thus in the area of open source software there is the GNU General Public License, which combines the usual copyright © with the statement ‘everyone is permitted to copy and distribute verbatim copies . . .’ (from Cho, 2005).

In LBS, it is understood that a provider will need to bring together a range of data sets in order to offer a credible service. Licenses to use these data sets will need to be contracted for and royalties paid to the owners/guardians. These costs will of course be passed on to the consumer or offset to a greater or lesser extent by, say, advertising. A provider will also have to ensure that its use of data conforms to the terms of its licenses when providing a service to its customers. Nevertheless, by bringing together diverse data sets in a novel way or by creating new, derived information through analysis of the data sets, then LBS providers can claim their own copyright for their unique products as long as in doing so they give due recognition to the copyright of the base data used. This is an important informational characteristic of GIS in that in bringing together already copyrighted data, it is possible to generate new information that can carry its own copyright. Copyright in this way becomes multilayered and complex to administer.

### 10.5.3 Liability

An LBS provider would need to consider its liabilities when offering information services that direct people to locations, and in doing so a provider would have a legal duty of care. Aspects of this have already been touched on in Section 7.5 when considering a user’s context. The issue of who carries responsibility for the production, analysis, dissemination and use of geographical information has been debated for quite some time in the world of GIS (e.g. Epstein and Roitman, 1987; Aronoff, 1989; Cho, 1998) and for a number of years there was even a journal entitled *GIS and the Law*. Ultimate users of geographical information may only have a cursory knowledge (if at all) of how geographical information is produced. They are often unaware of the

uncertainty in data and maps (Section 5.4.1, Section 5.4.2) and may place undue heavy reliance on geographical information in making both personal and business decisions. When things go wrong and injury or damage to individuals and property occurs, plaintiffs and their lawyers will want to place the responsibility and obtain compensation. So the issue of liability for information is clearly important. A very readable explanation of the principles and how they affect professional liability can be found in Wickins (1989); of course, any reader wishing to set up a GIS-based or LBS business should seek professional legal advice.

Liability arises where an individual fails to perform their duties or responsibilities (Cho, 1998). There are two major areas of liability that need to be considered under the law: contract and tort (negligence). Liability in contract is based on an agreement that is legally enforceable between parties. Under common law such a contract may be oral or written or a mixture of both; there may be two or more parties involved. A contract becomes binding when: there is an offer that is accepted, where in consideration of promises given something of value (payment, rights, benefit) is returned, and that all parties have reached consensus on the subject, content and provisions of the contract. Given the technological heterogeneity of LBS, it is unlikely that any single provider could deliver all aspects of LBS entirely on their own. There would need, for example, to be contracts for the purchase and servicing of equipment, contracts with network providers and contracts with data providers. Regardless of the contractual complexity in setting up the service, there we would need to be some form of contract between provider and user or provider and those wishing to advertise in order to generate an income stream. In general, only the parties to a contract are legally bound by it and can use it as a defence. Other persons or organizations cannot use the contract to gain rights even though they may be affected by it in some way; this is known as the principle of privity of contract. This leads to a fundamental distinction between law and morality. Thus, if someone sets up a contract to sell a business to someone else, but then does not deliver on the contract, the buyer can take the seller to court. If, on the other hand, a seller successfully contracts to sell a business and immediately sets up a rival one in the same area, it may seem morally wrong but the buyer does not have grounds to sue for breach of contract.

The other major area of liability is tort or negligence. This is a legal duty of care not to injure other persons or their interests in property. The principle of privity does not apply to tort – whoever is



hurt by some action arising through lack of due care has the right to seek redress. Contract and tort are quite separate branches of the law. The duty of care is partly based on reasonable foreseeability. In other words, whilst all due care can be taken in creating a safe product, the producer cannot be expected to necessarily foresee all the ways its product may be used outside its original intended use. However, in an increasingly litigious society most products now have to come with statements of proper usage, warnings, let-out clauses and caveats. So anyone switching on a SatNav will see a warning telling them it is dangerous to look at or otherwise interact with the device whilst driving. (See also the clauses stated in Section 10.4 above in relation to British Standards.) For a case of negligence it is necessary to prove that the producer had a legal duty of care, was in breach of that duty and that there was consequential damage. Liability can arise not just through careless acts but through omissions (not saying or doing something which then results in harm) or indeed through careless statements which a reasonable person may rely on as information.

This has implications for the topics discussed in Chapter 9 where it needs to be reasonably foreseen how a reasonable person might use, say, a map delivered by LBS for wayfinding purposes. An important case in this regard is cited in Epstein and Roitman (1987) and concerns a plane crash on the approach to an airport (*Aetna Casualty and Surety vs Jeppeson & Co.*, 1981). Jeppeson & Co. published the instrument approach charts used by the pilots and were found to provide accurate information about the airplane approach to the airport. However, there were two graphic displays of the information from different perspectives but the scale of the graphics were different by a factor of five, and were different from other charts in the publication. The differing scales led to a mistaken reliance by the pilots on the position of certain features shown in relation to the airport. The publishers of the chart were found to be at fault since '(the) professional must be able to rely on the accuracy of this information if he or she is to benefit from the mechanization or special compilation of the data . . .'; but the pilots shared the fault as '(the) professional . . . will be expected to use his or her professional judgment in evaluating the information and not rely blindly on what he or she is told by the computer or the charts' (quoted in Epstein and Roitman, 1987 p. 366). In many LBS applications there will certainly be a 'special compilation of the data', as the response is expected to be tailored to location and context, but users may not be professional users of such



information and may not be relied upon to exercise high levels of ‘judgment in evaluating the information’.

There is, in general, no legal right to privacy, partly because the notion of privacy is highly subjective. Traditionally there has been self-regulation, say by the press, but such are the commercial gains to be made from intrusive revelations into the private lives of the rich and famous that so-called privacy laws have been introduced in a number of jurisdictions. What has been more strongly recognized in the age of databases and digital data, particularly in the European Union, is individuals’ rights to keep their personal information private (i.e. not openly accessible to the public). This has become more important with regard to identity theft for the purposes of fraud. The EU Data Protection Directive (95/46/EC), which is enforced through member states’ legislation such as the UK 1998 Data Protection Act, mandates the safeguarding of personal information. The Act not only regulates the storage and updating of personal data (which can be both factual and expressions of opinion) but also the processing, sharing and eventual disposal of that data, as well as the rights of individuals to know what data are being held about them. It is a criminal offence to unlawfully obtain, sell or otherwise disclose personal data.

An LBS provider is likely to store conventional personal data about users (address, banking details and so on) as well as tracking, preference and transactions data which would be construed as personal data as they pertain to an individual. Even if such data is only held short term, the provider has a duty safeguard such data (e.g. from hackers) and must make sure that in disposing of the data (say, wiping from the disks) that they could still not be accessed by others (i.e. a simple file delete may not be relied upon to physically remove the data). For LBS the location and activities of individuals can be tracked and recorded in an unprecedented way, and unless users are confident that such data will be protected from theft, inadvertent disclosure or misuse they are unlikely to become subscribers. Protection of personal data is thus not only a legal obligation but a pivotal component of the business model.

## 10.6 Social Issues

---

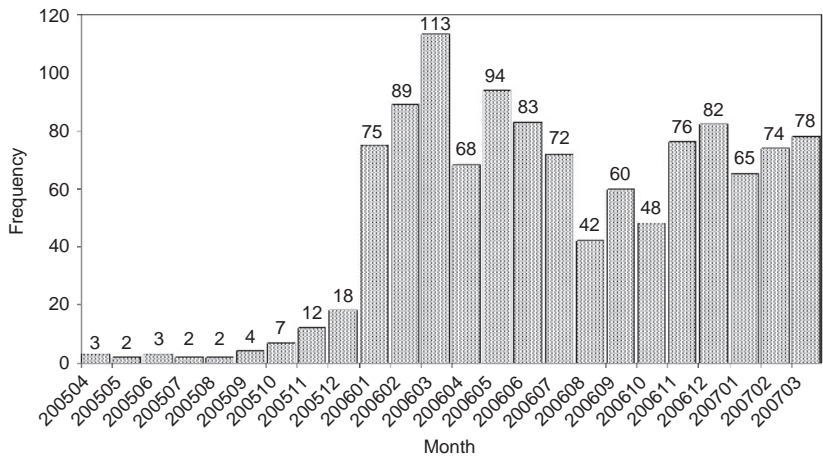
In Section 10.3 the effect of social appropriation on the shaping of new technologies was discussed briefly and how it can be the case that the

eventual use and utility of new technologies may be far removed from what the original developers had in mind. This arises from the interaction between society and a new technology. There are also sometimes unintended effects of new technologies, even new behaviours, some of which can indeed be deviant. Who, for example, in the early days of computing could have foreseen the need for a computer and network security industry in the face of today's unremitting onslaught of viruses? From an LBS perspective there are already viruses that attack mobile devices and can be transmitted via SMS (e.g. Mibir.A), Bluetooth (e.g. Cabir) or in file-sharing applications for, say, mobile games (e.g. Frontal.A Trojan). Mobile devices are easily dropped, left on restaurant tables and otherwise lost or stolen. This means loss of service and whilst not the fault of the LBS provider shows a certain level of vulnerability of mobile device based services in society at large. A concern for LBS users might be the amount of personal data that might be accessible from a lost or stolen mobile device if it contains profiles and preferences that can be edited locally (on the mobile device) as part of the contextual information uploaded by an LBS provider. Even if the mobile device stays in one's possession, it is still technically feasible to tap mobile phones and other wireless transmissions or even to hack into a service provider's servers in order to snoop on someone's itinerary and movements. Users of computers and mobile devices these days have a heightened awareness of the need to protect personal information and their data.

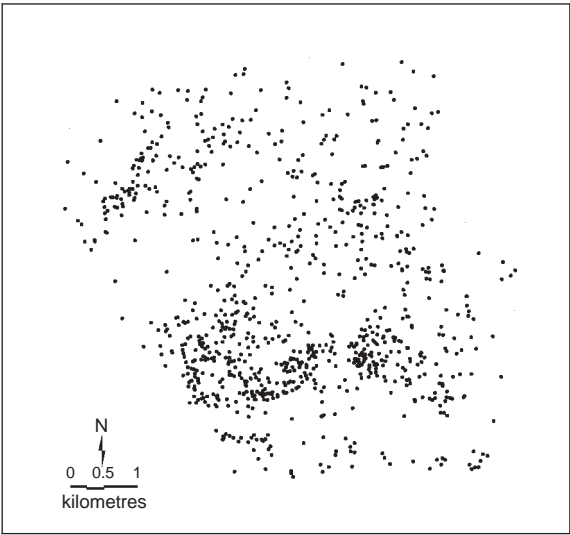
One social aberration that has had to be legislated against in many countries is the use of mobile devices whilst driving! This has implications for how LBS are accessed whilst driving; they would certainly need to be hands-free and require minimal visual contact with the mobile device. In a similar vein, public service advertising has been warning of the dangers of crossing roads whilst distracted by the use of a mobile device, no doubt prompted by rising death and injury of pedestrians in such circumstances.

A prime example of how producers and providers need to consider social implications when they engineer their products is the rise in thefts of in-car navigation systems (SatNavs) in tandem with their rise in popularity. In late 2005 the status of SatNavs moved from being a 'nice-to-have' consumer good to being a 'must-have'. In the United Kingdom this first became noticeable in the 2005 Christmas shopping season. The SatNav is probably the first geo-engineered product to have such widespread appeal. Monthly SatNav theft for one area in London over a two-year period from April 2005 to March

2007 is charted in Figure 10.5a. For this area, SatNav theft accounted for 13% of incidents of theft from motor vehicles. In 2007 some 11 600 portable SatNavs were stolen from vehicles in London – on average one every 45 minutes! Such has been the extent of the problem right

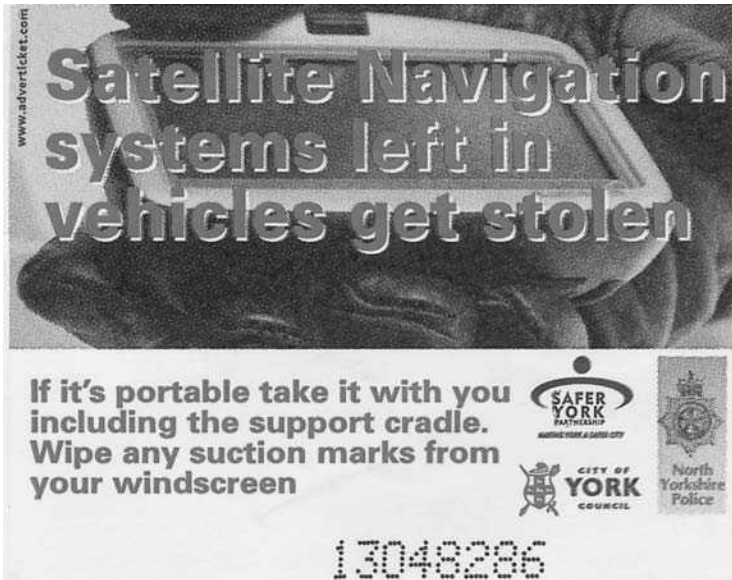


(a)



(b)

**Figure 10.5** SatNav theft in an area of London: (a) monthly thefts from April 2005 to March 2007; (b) distribution of thefts.



**Figure 10.6** Warning to motorists about SatNav theft on the rear side of a pay-and-display parking ticket.

across the United Kingdom that police forces have been using all means at their disposal to educate owners into not leaving portable SatNavs in their cars – Figure 10.6 shows a police warning on the back of a pay-and-display parking ticket from York in the north of England. For the incidents of SatNav theft mapped in Figure 10.5b for one area in London, the victims of these thefts are not just those that live locally, some  $\frac{2}{3}$  live outside the area as shown in Figure 10.7. For the Euclidean distance between victim's home and location of SatNav theft, the median is just over 11 kilometres. A quarter of these distances are below 260 m indicating thefts from local inhabitants whilst another quarter travelled more than 75 km from home to the crime scene, no doubt using their SatNav to navigate them there! The time of day of theft shown in Figure 10.8 compares weekdays with weekend (the peak at 1 a.m. is an artefact of crime reporting when there is uncertainty in actual time during the night and should be ignored). The weekday frequencies are clearly higher than for weekends if only because it compares five days with two; but the main interest is the peak time for the thefts. During weekdays this rises steeply to a lunchtime peak whereas at weekends thefts rise steadily



Figure 10.7 Distribution of victim home addresses for SatNav thefts in Figure 10.5b.

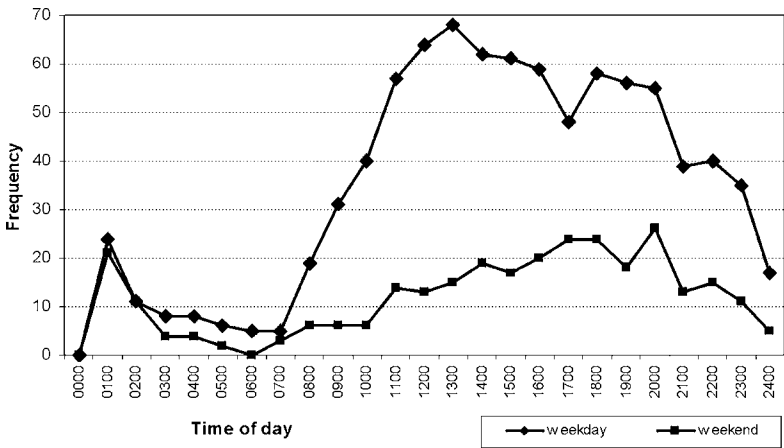


Figure 10.8 Different in time of day for weekday thefts of SatNav and those at weekend.

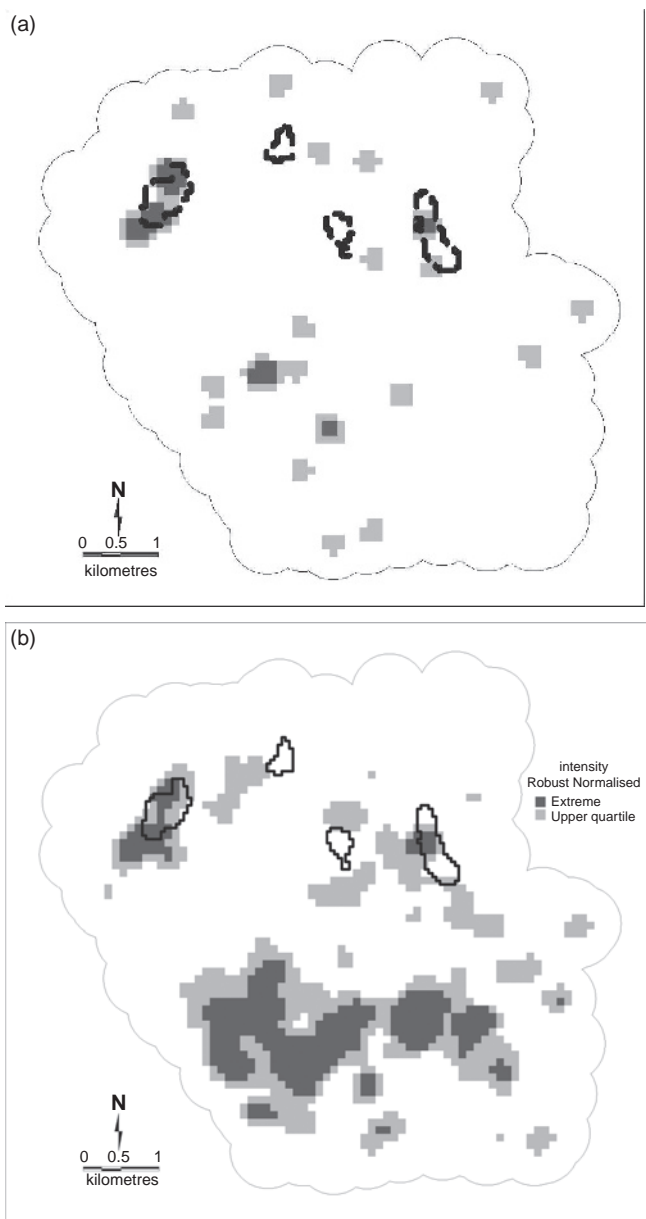
during the day to an evening peak. For this area of London, the geographical distribution of the early thefts up to Christmas 2005 (Figure 10.9a) shows the main concentration of thefts (hot spots) around the commercial centres. However, once SatNavs became a 'must have' the largest concentrations of theft moved to residential areas (Figure 10.9b) away from the usual centres of crime.

Could the vulnerability of SatNav to theft been foreseen? Indeed it could have! The same problem first arose with mobile phones from the late 1990s onwards when they became a mass consumer product. Theft of mobile phones, for example, reached almost epidemic proportions in London where users were advised not to openly display and use mobile phones. The solution was to have each device identifiable through an International Mobile Equipment Identity (IMEI) number made up of an eight digit Type Approval Code (TAC), a six digit serial number (SNR) and a check digit (CD).<sup>2</sup> It can be displayed on a mobile phone by dialling \*#06#. The IMEI has a dual purpose. Firstly, it allows a provider to refuse service to and block a mobile phone that presents an IMEI during handshake that is known to be stolen, effectively disabling a stolen mobile phone. Secondly, it provides a means, such as through <http://www.immobilize.com>, of recovering stolen property. This process of rendering a mobile phone useless to a petty thief (but not entirely to the ardent hacker) has substantially reduced the amount of mobile phone theft.

SatNav manufacturers produce some models capable of wireless communication for real-time updates of traffic conditions. One security approach then would be to extend wireless communication to all models so that not only could a SatNav be passworded, but it could also have an IMEI that requires wireless registration each time it is switched on (as with a mobile phone). SatNavs could also be provided with fingerprint recognition, as are many laptop PCs and most PDAs. An alternative security approach would be to use Secure Digital (SD) card keys that need to be slotted-in before a SatNav can be activated. This key could be created by the owner on-line using a PC to access the manufacturer's Web portal. By registering the IMEI of a newly purchased SatNav, the SD could be coded to activate the SatNav. When the SatNav is not in use the SD card can be removed, rendering the SatNav unusable without it. It is clear from this case

---

<sup>2</sup> From 1 January 2003, superseding an earlier coding scheme.



**Figure 10.9** Hot spot mapping: (a) cumulative to end November 2005; (b) cumulative to end March 2007.

study of SatNavs, that in geo-information engineering the possible social implications (and risks) of a product need to be well thought through at the design stage.

### 10.7 Conclusions

---

This is the end of the book. LBS at first sight seem conceptually simple from the definition. They are, in a complex heterogeneous technology and ground-breaking as GI Engineering. In a way, each chapter has been a piece in a jigsaw puzzle which when put together provides an in-depth view and contextualization of this new and exciting area. At the same time, remove a piece – ignore the topics set out in any one chapter – and it is likely that an application of LBS will not turn out successful. We have stuck largely to the principles that underpin LBS rather than looking at the details of specific applications. New applications, devices and middleware will be entering the market and evolve – they will easily be found through the Web. With the knowledge and understanding now gained, readers will be able to assess these critically. LBS are still in the early stages of their journey to becoming ubiquitous and firmly embedded in society. What we have tried to stress is the richness of the cross-disciplinary domain from the perspectives of pure and applied research through to engineering, design and setting up in business. LBS have a value-added selling proposition for students, academics, professionals and entrepreneurs alike:

– You are here...you'll soon be there –



# Acronyms

AA	Automobile Association
ACL	Asynchronous Connectionless
A-FLT	Advanced Forward Link Trilateration
A-GPS	Assisted-GPS
IrDA	Infrared Data Association
AMPS	Advanced Mobile Phone System
AOA	Angle of Arrival
AP	Access Point
API	aerial photographic interpretation
API	Application Program Interface
AR	Augmented Reality
AS	Anti-Spoofing
ASP	Application Service Provider
ATM	Automatic Teller Machine
BFS	Breadth-First Search
bps	bits per second
BS	Base Station
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CAVE	Cave Automatic Virtual Environment
CD	check digit
CDMA	Code Division Multiple Access
cdmaOne	Interim Standard 95 (also known as IS-95)
Cell-ID	Cell Identification
CGI	Cell Global Identity, also known as Cell-ID
COO	Cell of Origin, also known as Cell-ID
CORS	Continuously Operating Reference Station
CPU	Central Processing Unit
D-AMPS	Digital American Mobile Phone Service
DARPA	The Defense Advanced Research Projects Agency

## Acronyms

DDL	Data Definition Language
DFS	Depth-First Search
DGPS	Differential Global Positioning System
DIME	Dual Independent Map Encoding
DML	Data Modification Language
DNS	Domain Name System
DSSS	Direct Sequence Spread-Spectrum
DVD	Digital Video Disc
E	easting
ECEF	Earth-Centered-Earth-Fixed
EDGE	Enhanced Data Rates for GSM Evolution
E-FLT	Enhanced Forward Link Trilateration
EGNOS	European Geostationary navigation Overlay Service
E-OTD	Enhanced-Observed Time Difference
E-TACS	Extended-TACS
ETSI	European Telecommunications Standards Institute
FAA	Federal Aviation Administration
FCC	Federal Communications Commission (USA)
FDMA	Frequency Division Multiple Access
FGDC	Federal Geographic Data Committee (USA)
FHSS	Frequency-Hopping Spread Spectrum
FM	Frequency Modulation
FRNs	Fixed Relay Networks
FTP	File Transfer Protocol
GBF	Geographical Base File
GDOP	Geometric Dilution of Precision
GIS	Geographical Information Systems
GIScience	Geo-Information Science
GML	Geography Markup Language
GMLC	Gateway Mobile Location Centre
GMS	GeoMobility Server
GMSC	Gateway Mobile Switching Centre
GMSK	Gaussian Minimum Shift Keying
GNSS	Global Navigation Satellite Systems
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile communications
HCI	Human-Computer Interaction
HDOP	Horizontal Dilution of Precision
HIPERLAN	High Performance Radio Local Area Network
HLRs	Home Location Registers
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
I/O	Input-Output (read-write from storage media)

ICT	Information and Communication Technologies
iDen	Integrated Dispatch Enhanced Network
IEEE	Institute of Electrical and Electronic Engineers
IGS	International GPS Service
IMEI	International Mobile Equipment Identity
IMTS	Improved Mobile Telephone Service
INS	Inertial Navigation Systems
IP	Internet Protocol
IR	infrared
IS-136	Interim Standard 136
IS-95	Interim Standard 95 (also know as cdmaOne)
ISM	Industrial Scientific and Medical
ISP	Internet Service Provider
ITN	Integrated Transport Network
ITU	International Telecommunication Union
J-TACS	Japanese-TACS
LAAS	Local Area Augmentation System
LADGPS	Local Area DGPS
LAN	Local Area Network
LBS	Location-Based Services
LiDAR	Light Detection and Ranging
LMUs	Location Measurement Units
MAUP	Modifiable Areal unit Problem
MBR	Minimum Bounding Rectangle
MBWA	Mobile Broadband Wireless Access
MeC	Mobile e-commerce
MIMO	Multiple-Input-Multiple-Output
MMS	Multimedia Message Services
MS	Mobile station (usually known as mobile phones)
MSSC	Mobile Service Switching Centre
N	nothing
NICTs	New Information and Communication Technologies
NMT	Nordisk Mobile Telefoni (Nordic Mobile Telephony in English)
NREN	National Research and Education Network
NSF	National Science Foundation
NSFNET	National Science Foundation Network
OFDM	Orthogonal Frequency Division Multiplexing
OGC	Open GIS Consortium (now Open Geospatial Consortium)
OO	Object-Oriented
OSS	Operations and Support System
OWL	Web Ontology Language
PANs	Personal Area Networks

## Acronyms

PC	Personal Computer
PCC	Percentage Correctly Classified
PCS	Personal communication Service
PDA	Personal Digital Assistant
PDC	Pacific Digital Cellular
PDOP	Position Dilution of Precision
PHP	PHP: Hypertext Processor
PIN	Personal Identification Number
POI	Points of Interest
PPS	Precise Positioning Service
PRN	Pseudo-Random Noise
PSAP	Public Services Answering Point
PSK	Phase Shift Keying
PSP	Positioning Services Provider
PSTN	Public Switched Telephone Network
QA	Quality Assurance
RA	Relational Algebra
RAC	Royal Automobile Club
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RDOP	Relative Dilution of Precision
RFID	Radio Frequency Identification
RMSE	Root Mean Standard Error
SA	Selective Availability
SCO	Synchronous Connection-Oriented
SDI	Spatial Data Infrastructure
SDSS	Spatial Decision Support Systems
SIM	Subscriber Identity Module
SMS	Short Message Service
SPS	Standard Positioning Service
SQL	Structured Query Language
TA	Timing Advanced
TAC	Type Approval Code
TACS	Total-Access Communications System
TCP	Transmission Control Protocol
TDOA	Time Difference Of Arrival
TD-SCDMA	Time Division-Synchronous Code Division Multiple Access
TDMA	Time Division Multiple Access
TIA	Telecommunications Industry Association
TOA	Time of Arrival
TTF	Time To First Fix
UHF	Ultra High Frequency

UMTS	Universal Mobile Telecommunication System (also known as W-CDMA)
URI	Uniform Resource Identifier
URL	Uniform Resource Locators
USDC	US Digital Cellular
UTRAN	UMTS Terrestrial Radio Access Network
UWB	Ultra Wide Band
VDOP	Vertical Dilution of Precision
VE	Virtual Environment
VHF	Very High Frequency
VLR	Visitor Location Registration
VoIP	Voice over Internet Protocol
VR	Virtual Reality
VSF	Variable-Spreading-Factor
VUI	Voice User Interface
WAAS	Wide Area Augmentation System
WADGPS	Wide Area DGPS
WAN	Wide Area Network
WAP	Wireless Application Protocol
W-CDMA	Wideband CDMA (also known as UMTS)
WGS84	World Geodetic System 1984
WiDen	Wideband Integrated Dispatch Enhanced Network
WiFi	Wireless Fidelity
WiMax	Worldwide Interoperability for Microwave Access
WISP	Wireless Internet Service Provider
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network
WML	Wireless Markup Language
WTP	WAP Transaction Protocol
WTLS	Wireless Transport Layer Security
WWW	World Wide Web
XML	Extensible Markup Language

# References

- Abowd, G.D., Atkeson, C.G., Hong, J. *et al.* (1997) Cyberguide: a mobile context-aware tour guide. *Wireless Networks*, **3**, 5, 421–33.
- Agar, J. (2004) *Constant Touch: A Global History of the Mobile Phone*. Icon Books, Duxford.
- Albrecht, J. (2005) A new age for geosimulation. *Transactions in GIS*, **9**, 451–54.
- Allen, J.F. (1984) Towards a general theory of action and time. *Artificial Intelligence*, **23**, 123–54.
- American Society of Civil Engineers (1983) *Map Uses, Scales and Accuracies for Engineering and Associated Purposes*, ASCE, New York.
- American Society of Photogrammetry and Remote Sensing (1985) Accuracy specifications for large scale maps. *Photogrammetric Engineering and Remote Sensing*, **51**, 195–99.
- Antoniou, G. and van Harmelen, F. (2004) *A Semantic Web Primer*, The MIT Press, Cambridge, MA.
- Arbor, I. and Bjerke, B. (1997) *Methodology for Creating Business Knowledge*, Sage, Thousand Oaks, CA.
- Aronoff, S. (1989) *Geographic Information Systems: A Management Perspective*, WDL Publications, Ottawa.
- Atkinson, P. and Tate, N. (2000) Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer*, **52**, 607–23.
- Baldi P., Frasconi, P. and Smyth, P. (2003) *Modelling the Internet and the Web: Probabilistic Methods and Algorithms*, John Wiley & Sons, Ltd, Chichester.
- Barkhuus, L. and Dey, A. (2003) Is context-aware computing taking control away from the user? Three levels of interactivity examined. Proceedings 5th International Symposium on Ubiquitous Computing, Seattle, 149–56.
- Batty, M. (1990) Invisible cities. *Environment and Planning B*, **17**, 127–30.
- Batty, M. (2001) Contradictions and conceptions of the digital city. *Environment and Planning B*, **28**, 479–80.

- Bedford, M. (2004) Are you ready for the ride? *GEO: Connexion UK*, **3** (1), 50–51.
- Bell, D. (1980) The social framework of the information society. In *The Microelectronics Revolution* (ed. T. Forester), Blackwell, Oxford, 500–49.
- Benenson, I. and Torrens, P.M. (2004) Geosimulation: object-based modelling of urban phenomena. *Computers Environment & Urban Systems*, **28**, 1–8.
- Bennett, F., Richardson, T. and Harter, A. (1994) Teleporting – making applications mobile. Proceedings of IEEE workshop on mobile computing systems and applications, Santa Cruz, California, 82–4.
- Berners-Lee, T. (1999) *Weaving the Web*. Harper San Francisco, San Francisco.
- Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. *The Scientific American*. Available at <http://www.sciam.com/article.cfm?id=the-semantic-web>.
- Bertin, J. (1967) *Sémiologie Graphique*, Mouton, Den Haag.
- Bickmore, T.W. and Schilit, B.N. (1997) Digestor: device-independent access to the World Wide Web. *Computer Networks and ISDN Systems*, **29**, 1075–82.
- Bornträger, C., Cheverst, K., Davies, N. *et al.* (2003) Experiments with multi-modal interfaces in a context-aware city guide. *Mobile HCI 2003*, Springer-Verlag, Berlin, 116–30.
- Braun, P. (2003) *Primer on Wireless GIS*, The Urban and Regional Information Systems Association, Park Ridge, IL.
- Brenner, C. and Haala, N. (2001) Automated reconstruction of 3D city models. In *3d Synthetic Environment Reconstruction* (ed. M. Abdelguerfi), Kluwer, Boston, MA.
- Brimicombe, A.J. (2003) *GIS, Environmental Modelling and Engineering*, Taylor & Francis, London.
- Brimicombe, A.J. (2006) A dual approach to cluster discovery in point event data sets. *Computers, Environment and Urban Systems*, **31**, 4–18.
- Brimicombe, A.J. (2008) Location-based services and GIS. In *Handbook of Geographical Information Science* (eds J.P. Wilson and A.S. Fotheringham), Blackwell, Oxford, 581–95.
- Brimicombe, A.J. and Li, Y. (2006) Mobile space-time envelopes for location-based services. *Transactions in GIS*, **10**, 5–23.
- Brimicombe, A.J. and Yeung, D. (1995) An object oriented approach to spatially inexact socio-cultural data. Proceedings 4th International Conference on Computers in Urban Planning & Urban Management, Melbourne, Australia, 2, 519–30.
- Brown, A. and Feringa, W. (2003) *Colour Basics for GIS Users*, Prentice Hall, Harlow, UK.
- Bruber, T. (1993) A translation approach to portable ontology specification. *Knowledge Acquisition*, **5**, 199–220.

## References

- BSI (2000) *BS 7666-1:2000 Spatial Data-Sets for Geographical Referencing Part 1: Specification for a Street Gazetteer*, British Standards Institution, Chiswick.
- Buehler, K. and McKee, L. (1998) *The Open GIS Guide: Introduction to Interoperable Geoprocessing and OpenGIS Specification*, Open GIS Consortium Inc., Wayland, MA.
- Bureau of Budget (1947) *National Map Accuracy Standards*, US Government Printing Office, Washington, D.C.
- Burrough, P.A. (2000) Whither GIS (as systems and as science)? *Computers, Environment & Urban Systems*, **24**, 1–3.
- Burrough, P.A. and Frank, A. (1996) *Geographic Objects With Indeterminate Boundaries*, Taylor & Francis, London.
- Burrough, P.A. and McDonnell, R.A. (1998) *Principles of Geographical Information Systems*, Oxford University Press, Oxford.
- Buttenfield, B.P. (1995) *Object-Oriented Map Generalization: Modelling and Cartographic Considerations*. In *GIS and Generalisation: Methodological and Practical Issues* (eds J.-C. Muller, J.-P. Lagrange and R. Weibel), Taylor & Francis, London, 91–105.
- Carter, H. (1981) *The Study of Urban Geography*, Edward Arnold, London.
- Cartwright, W.E. (2008) Mapping in a digital age. In *The Handbook of Geographic Information Science* (eds J.P. Wilson and A.S. Fotheringham), Blackwell, Malden, MA, 199–221.
- Cartwright, W.E., Gartner, G. and Peterson, M.P. (2007) *Multimedia Cartography*, Springer-Verlag, Berlin.
- Castells, M. (1989) *The Informational City*, Blackwell, Oxford.
- Castells, M. (1996) *The Information Age: Economy, Society and Culture. Volume I: the Rise of the Network Society*, Blackwell, Oxford.
- Castells, M. (1997) An introduction to the information age. *City*, **7**, 6–16.
- Castells, M. (1998) *The Information Age: Economy, Society and Culture. Volume III: End of Millennium*, Blackwell, Oxford.
- Castells, M. (2001) The information city, the new economy, and the network society. In *People, Cities and the New Information Economy* (eds A. Kasvio *et al.*), Palmenia, Helsinki, 22–37.
- Castells, M., Fernandez-Ardevol, M., Qiu, J.L. and Sey, A. (2006) *Mobile Communication and Society – a Global Perspective*, MIT Press, Cambridge, MA.
- Chan, T.O. and Williamson, I.P. (1999) The different identities of GIS and GIS diffusion. *International Journal of Geographical Information Science*, **13**, 267–81.
- Chen, G. and Kotz, D. (2000) A survey of context-aware mobile computing research. *Dartmouth Computer Science Technical Report TR2000-381*, Dartmouth College, Hanover, NH.
- Chen, P. (1976) The entity-relationship model: towards a unified view of data. *ACM Transactions on Database Systems*, **1**, 265–77.



- Chen, X., Chen, Y. and Rao, F. (2003) An efficient spatial publish/subscribe system for intelligent location-based services. *Proceedings 2nd International Workshop on Distributed Event-Based Systems*, San Diego, CA, 1–6.
- Cheverst, K., Davies, N., Mitchell, K. *et al.* (2000) Developing a context-aware electronic tourist guide: some issues and experiences. *CHI 2000 Conference Proceedings*, 17–24.
- Cho, G. (1998) *Geographic Information Systems and the Law: Mapping the Legal Frontiers*, John Wiley & Sons, Ltd, Chichester.
- Cho, G. (2005) *Geographic Information Science: Mastering the Legal Issues*, John Wiley & Sons, Ltd, Chichester.
- Chrisman, N. (1997) *Exploring Geographic Information Systems*, John Wiley & Sons, Inc., New York.
- Cliff, A.D. and Ord, J.K. (1981) *Spatial Processes: Models and Applications*, Pion, London.
- Coad, P. and Yourdan, E. (1991) *Object-Oriented Analysis*, Prentice-Hall International, Englewood Cliffs, NJ.
- Cofta, P. (2007) Confidence, trust and identity. *BT Technology Journal*, **25** (2), 173–78.
- Coleman, D.J. and McLaughlin, J. (1998) Defining global geospatial data infrastructure (GGDI): components, stakeholders and interfaces. *Geomatica*, **52**, 129–43.
- Comer, J.C. and Wikle, T.A. (2008) Worldwide diffusion of cellular telephone, 1995–2005. *The Professional Geographer*, **60**, 252–69.
- Couclelis, H. (2004) The construction of the digital city. *Environment and Planning B*, **31**, 5–19.
- Curry, M.R. (1995) GIS and the inevitability of ethical inconsistency. In *Ground Truth* (ed. J. Pickles), The Guilford Press, New York, 68–87.
- Daconta, M.C., Obrst, L.J. and Smith, K.T. (2003) *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*, John Wiley & Sons, Inc., Indianapolis, IN.
- Date, C.J. (1990) *An Introduction to Database Systems*, Volume 1, Addison-Wesley, Reading, MA.
- de Smith, M.J., Goodchild, M.F. and Longley, P.A. (2007) *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*, Matador, Leicester.
- Dey, A.K. (2001) Understanding and using context. *Personal and Ubiquitous Computing Journal*, **5**, 4–7.
- Dijkstra, E.W. (1959) A note on two problems of connexion with graphs. *Numerical Mathematics*, **1**, 269–71.
- Dillemuth, J., Coldsberry, K. and Clarke, K.C. (2007) Choosing the scale and extent of maps for navigation with mobile computing systems. *Journal of Location Based Services*, **1**, 46–61.
- Dix, A. and Abowd, G. (1996) Modelling status and event behaviour of interactive systems. *Software Engineering Journal*, **11**, 334–46.

## References

- Dix, A., Rodden, T., Davies, N. *et al.* (2000) Exploiting space and location as a design framework for interactive mobile systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, **7**, 285–321.
- Dogan, M. and Pahre, R. (1990) *Creative Marginality: Innovation at the Intersections of Social Sciences*, Westview Press, Boulder, CO.
- Douglas, D.H. and Peucker, T.K. (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, **10**, 112–22.
- Dourish, P. What we talk about when we talk about context. *Personal and Ubiquitous Computing Journal*, **8**, 19–30.
- Drummond, J., Joao, E. and Billen, R. (2007) Current and future trends in dynamic and mobile GIS. In *Dynamic and Mobile GIS: Investigating Changes in Space and Time* (eds J. Drummond, R. Billen, E. Joao and D. Forrest), CRC Press, Boca Raton, FL, 289–300.
- Duchêne, C. (2003) Automated map generalisation using communicating agents. Proceedings of the Twenty-first International Cartographic Conference, Durban, RSA, 160–69.
- Duckham, M., Goodchild, M. and Worboys, M. (2003) *Foundations of Geographical Information Science*, Taylor & Francis, London.
- Dueker, K. and Kjerne, D. (1989) *Multipurpose Cadastre: Terms and Definitions*. American Society for Photogrammetry and Remote Sensing and American Congress on Surveying and Mapping, Falls Church, Virginia.
- Dyson, E., Gilder, G., Keyworth, G. and Toffler, A. (1996) Cyberspace and the American dream. *The Information Society*, **12**, 295–308.
- Egenhofer, M.J. and Herring, J.R. (1990) A mathematical framework for the definition of topological relationships. Proceedings 4th International Symposium on Spatial Data Handling, Zurich, Switzerland, 803–13.
- e-Government Unit (2006) E-Government Metadata Standard Version 3.1, Cabinet Office, London.
- Ekpenyong, F., Palmer-Brown, D. and Brimicombe, A.J. (2009) Updating of road network databases: spatio-temporal trajectory grouping using snap-drift neural network. Proceedings 10th international conference on engineering applications of neural networks, Thessalonica, Greece, 237–46.
- Epstein, E.F. and Roitman, H. (1987) Liability for Information. Reprinted in *Introductory Readings in Geographic Information Systems* (eds D.J. Peuquet and D.F. Marble, 1990), Taylor & Francis, London.
- ESRI (2000) *What Are Location Services? the GIS Perspective*, ESRI, Redlands, CA.
- FCC (2005) Wireless 911 Services. Federal Communications Commission, Washington D.C., [www.fcc.gov/cgb/consumerfacts/wireless911srv.html](http://www.fcc.gov/cgb/consumerfacts/wireless911srv.html) (viewed 17 Aug 2006).
- Fels, S., Sumi, Y., Etani, T. *et al.* (1998) Progress of C-MAP: a context-aware mobile assistant. Proceedings of AAAI Spring Symposium on Intelligent Environments, Palo Alto, CA, 60–7.

- Ferber, J. (2005) Concepts et méthodologies multi-agents. In *Modélisation Et Simulation Multi-Agents* (eds F. Amblard and D. Phan), Lavoisier, Paris, 23–48.
- FGDC (1997) Content Standard for Digital Geospatial Metadata, Federal Geographic Data Committee, Washington, DC.
- Fickas, S., Kortuem, G. and Segall, Z. (1997) Software issues in wearable computing. Proceedings of the CHI workshop on research issues in wearable computers, ACM Press, New York.
- Flewelling, D.M. and Egenhofer, M.J. (1999) Using digital spatial archives effectively. *International Journal of Geographical Information Science*, **13**, 1–8.
- Fogli, D., Pittarello, F., Celentano, A. and Mussio, P. (2003) Context-aware interaction in a mobile environment. *Mobile HCI 2003*, Springer-Verlag, Berlin, 116–30.
- Fox, R.W. (1984) The world's urban explosion. *National Geographic*, **166**, 179–85.
- Foxall, G. (1997) *Marketing Psychology: The Paradigm in the Wings*. Macmillan, Basingstoke.
- Garland, D. (2000) The culture of high crime societies. *British Journal of Criminology*, **40**, 346–75.
- Gartner, G. (2004) Location-based mobile pedestrian navigation services – the role of multimedia cartography. Proceedings of ICA UPIMap 2004, Tokyo, Japan.
- Giaglis, G.M., Pateli, A., Fouskas, K. *et al.* (2002) On the potential use of mobile positioning technologies in indoor environments. Proceedings of 15th Bled Electronic Commerce Conference – e-Reality: Constructing the e-Economy, Bled, Slovenia, 413–29.
- Golledge, R.G. and Stimson R.J. (1997) *Spatial Behavior: A Geographic Perspective*, The Guilford Press, New York.
- Golledge, R.G., Rice, M.T. and Jacobson, R.D. (2006) Multimodal interfaces for representing and accessing geospatial information. In *Frontiers of Geographic Information Technology* (eds S. Rana and J. Sharma), Springer, Berlin, 181–208.
- Goodchild, M. (1990) Spatial information science. Proceedings 4th International Symposium on Spatial Data Handling, Zurich, Switzerland, Vol 1, 3–12.
- Goodchild, M. (1992) Geographical information science. *International Journal of Geographical Information Systems*, **6**, 31–45.
- Goodchild, M. (2001) Metrics of scale in remote sensing and GIS. *International Journal of Applied Earth Observation and Geoinformation*, **3**, 114–20.
- Goodchild, M. and Gopal, S. (1989) *Accuracy of Spatial Databases*, Taylor & Francis, London.
- Goodchild, M., Yuan, M. and Cova, T. (2007) Towards a general theory of geographic representation in GIS. *International Journal of Geographic Information Science*, **21**, 239–60.

## References

- Graham, S. (1998) The end of geography or the explosion of place? Conceptualising space, place and information technology. *Progress in Human Geography*, **22**, 165–85.
- Graham, S. and Marvin, S. (1996) *Telecommunications and the City: Electronic Spaces, Urban Places*, Routledge, London.
- Graham, S. and Wood, D. (2003) Digitizing surveillance: categorization, space and inequality. *Critical Social Policy*, **23**, 227–48.
- Gralla, P. (2006) *How Wireless Works*, 2nd edn, Que Publishing, Indianapolis, IN.
- Grejner-Brzezinska, D. (2004) Positioning and tracing approaches and technologies. In *Telegeoinformatics- Location-Based Computing and Services* (eds H.A. Karimi and A. Hammad), CRC Press, Boca Raton, FL, 69–106.
- Grejner-Brzezinska, D.A., Li, R., Haala, N. and Toth, C. (2004) From mobile mapping to telegeoinformatics: paradigm shift in geospatial data acquisition, processing and management. *Photogrammetric Engineering & Remote Sensing*, **70**, 197–210.
- Gruber, T.R. (1995) Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, **43**, Issues 4–5, 907–28.
- GSM Association (2003) Location Based Services version 3.1.0. Permanent Reference Document SE.23, GSM Association, <http://www.gsmworld.com>.
- Guarino, N. and Giaretta, P. (1995) Ontologies and knowledge bases: towards a terminological clarification. In *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing* (eds N. Mars), IOS Press, Amsterdam, 25–32.
- Hall, P. (1999) The future of cities. *Computers, Environment and Urban Systems*, **23**, 173–85.
- Haque, A. (2003) Information technology, GIS and democratic values: ethical implications for IT professionals in public service. *Ethics and Information Technology*, **5**, 39–48.
- Harris, J.W. and Stocker, H. (1998) *Handbook of Mathematics and Computational Science*, Springer-Verlag, New York.
- Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS and Neighbourhood Targeting*, John Wiley & Sons, Ltd, Chichester.
- Hewlett-Packard (2000) Welcome to CoolTown. Video, Hewlett-Packard Company, Palo-Alto, CA.
- Heywood, I., Cornelius, S. and Carver, S. (2005) *An Introduction to Geographical Information Systems*, Pearson, Harlow.
- Hofman-Wellenhof, B., Collins, J. and Lichtenegger, H. (2001) *Global Positioning System (GPS): Theory and Practice*, 5th edn, Springer-Verlag.
- Holtgrewe, B.J. and Freeze, J.T. (2002) *MapPoint for Dummies*, John Wiley & Sons, Inc, Hoboken, NJ.

- Holweg, D. and Kretschmer, U. (2006) Augmented reality visualization of geospatial data. In *Frontiers of Geographic Information Technology* (eds S. Rana and J. Sharma), Springer, Berlin, 229–40.
- Huang, B., Yi, S. and Chan, W.T. (2004) Spatio-temporal information integration in XML. *Future Generation Computer Systems*, **20**, 1157–70.
- Hull, B. (2003) ICT and social exclusion: the role of libraries. *Telematics and Informatics*, **20**, 131–42.
- Jiang, B. and Yao, X. (2006) Location-based services and GIS in perspective. *Computers Environment and Urban Systems*, **30**, 712–25.
- Julia, L. and Bing, L. (1999) Travel MATE: A demonstration of SRI's multimodal augmented tutoring environment. Proceedings of the Second IEEE and ACM International Workshop on Augmented Reality (IWAR'99), San Francisco, CA.
- Kaasinen, E. (2002) User needs for location-aware mobile services. *Personal and Ubiquitous Computing*, Springer, London, 70–9.
- Kellerman, A. (2000) Phases in the rise of the information society. *Info*, **2**, 537–41.
- Kohiyama, K. (2005) A decade in the development of mobile communications in Japan (1993–2002). In *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life* (eds M. Ito, D. Okabe and M. Matsuda), MIT Press, Cambridge, MA, 61–76.
- Kraak, M.-J. (2001) Cartographic principles. In *Web Cartography: Developments and Prospects* (eds M.-J. Kraak and A. Brown), Taylor & Francis, London, 53–72.
- Kraak, M.-J. and Brown, A. (2001) *Web Cartography: Developments and Prospects*, Taylor & Francis, London.
- Kray, C., Laakso, K., Elting, C. and Coors, V. (2003) Presenting route instructions on mobile devices. Proceedings of IUI03: Intelligent User Interfaces, ACM Press, New York.
- Kreller, B., Carrega, D., Shankar, J. *et al.* (1998) A mobile aware city guide application. Proceedings of ACTS Mobile Communications Summit, Rhodes, Greece, 60–5.
- Kuipers, B. (1978) Modeling spatial knowledge. *Cognitive Science*, **2**, 129–53.
- Kumar, K. (1995) *From Post-Industrial to Post-Modern Society*, Blackwell, Oxford.
- Küpper, A. (2005) *Location-Based Services – Fundamentals and Operation*, John Wiley & Sons, Ltd, Chichester.
- Lachapelle, G., Falkenberg, W., Neufeldt, D. and Kielland, P. (1990) Marine DGPS using code and carrier in a multipath environment. *Lighthouse*, **41**, 33–7.
- Lam, N. and Quattrochi, D. (1992) On the issues of scale, resolution, and fractal analysis in the mapping sciences. *The Professional Geographer*, **44**, 88–98.

## References

- Lane, G. (2003) Urban Tapestries: Wireless Networking, Public Authoring and Social Knowledge. [http://www.probicus.org/urbantapestries/Unis\\_WW\\_paper.html](http://www.probicus.org/urbantapestries/Unis_WW_paper.html).
- Langran, G. (1992) *Time in Geographical Information Systems*, Taylor & Francis, London.
- Lathrop, O. (1999) Virtual Reality. [www.inf.ed.ac.uk/teaching/courses/cg/web/intro-graphics/vr.html](http://www.inf.ed.ac.uk/teaching/courses/cg/web/intro-graphics/vr.html) (viewed July 2005).
- Laurini, R., Servigne, S. and Tanzi, T. (2001) A primer on TeleGeoProcessing and TeleGeoMonitoring. *Computers, Environment and Urban Systems*, **25**, 248–65.
- Leadbeater, C. (1999) *Living on Thin Air: the New Economy*, Hodder and Stroughton, London.
- Leung, H.K.Y., Burcea, I. and Jacobsen, H.-A. (2003) Modeling location-based services with subject spaces. Proceedings of the 2003 Conference on Collaborative Research, Toronto, Canada, 171–81.
- Li, C. (1995) An Exploratory Study of Generalization of GIS Data Layers Using Object-Oriented Modelling, Unpublished MSc Thesis, University College London.
- Li, C. (2005) Pedestrian Wayfinding Using Mobile Devices: an investigation of spatial information transaction and interaction, Unpublished PhD Thesis, University College London.
- Li, C. (2006) User preference, information transactions and location-based services: a study of urban pedestrian wayfinding. *Computers, Environment and Urban Systems*, **30**, 726–40.
- Li, C. and Longley, P. (2006) A test environment for location-based services applications. *Transactions in GIS*, **10**, 43–61.
- Li, C. and Maguire, D. (2003) The handheld revolution: towards ubiquitous GIS. In *Advanced Spatial Analysis: The CASA Book of GIS* (eds P. Longley and M. Batty), ESRI Press, Redlands, CA, 193–210.
- Li, C. and Willis, K. (2006) Modelling Context Aware Interaction for Wayfinding using Mobile Devices. Proceedings of the MobileHCI 06, Espoo, 97–100.
- Lillywhite, J. (1991) Identifying available spatial metadata. In *Metadata in the Geosciences* (eds D. Medyckyj-Scott *et al.*), Group D Publications, Loughborough, 3–12.
- Little, A. (2006) Driving a changing world. *GEOconnexion International*, February issue, 56–7.
- London Research Centre (1999) The Capital Divided, [www.london-research.gov.uk/hs/hspov.htm](http://www.london-research.gov.uk/hs/hspov.htm) (viewed 6 March 2000).
- Long, S., Kooper, R., Abowd, G. and Atkeson, C. (1996) Rapid prototyping of mobile context-aware applications: the Cyberguide case study. Proceedings of the 2nd Annual International Conference on Mobile Computing and Networking, ACM Press, New York, 97–107.



- Longley, P.A., Brooks, S.M., McDonnell, R. and Macmillan, W.D. (1998) *Geocomputation: A Primer*, John Wiley & Sons, Ltd, Chichester.
- Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2005) *Geographical Information Systems and Science*, 2nd edn, John Wiley & Sons, Ltd, Chichester.
- Mackaness, W.A. (2008) Generalization of spatial databases. In *The Handbook of Geographic Information Science* (eds J.P. Wilson and A.S. Fotheringham) Blackwell, Malden, MA, 221–38.
- Macmillan, W. (1998) Epilogue. In *Geocomputation: A Primer* (eds P.A. Longley *et al.*), John Wiley & Sons, Ltd, Chichester, 257–64.
- Maguire, D.J. and Longley, P.A. (2005) The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems*, **29**, 3–14.
- Mark, D.M. (1999) Spatial representation: a cognitive view. In *Geographical Information Systems, Volume 1* (eds P.A. Longley *et al.*), John Wiley and Sons, Inc., New York, 81–89.
- Mark, D.M. (2003) Geographic information science: defining the field. In *Foundations of Geographical Information Science* (eds M. Duckham *et al.*), Taylor & Francis, London, 3–18.
- Markopoulos, P., Ruyter, B. and Mackay, W. (2007) Introduction to this special issue on awareness systems design. *Human-Computer Interaction*, **22**, 1–6.
- Marmasse, N. and Schmandt, C. (2000) Location-aware information delivery with ComMotion. Proceedings of Second International Symposium on Handheld and Ubiquitous Computing, HUC 2000, Bristol, UK, 157–71.
- Masuda, Y. (1990) *Managing in the Information Society: Releasing Synergy Japanese Style*, Blackwell, Oxford.
- Matsuda, M. (2005a) Discourses of keitai in Japan. In *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life* (eds M. Ito, D. Okabe and M. Matsuda), MIT Press, Cambridge, MA, 19–40.
- Matsuda, M. (2005b) Mobile communication and selective sociality. In *Personal, Portable, Pedestrian: Mobile Phones in Japanese Life* (eds M. Ito, D. Okabe and M. Matsuda), MIT Press, Cambridge, MA, 123–42.
- McHaffie, P. (2000) Surfaces: tacit knowledge, formal language, and metaphor at the Harvard Lab for Computer Graphics and Spatial Analysis. *International Journal of Geographical Information Science*, **14**, 755–773.
- McMaster, R.B. and Shea, K.S. (1992) *Generalization in Digital Cartography*, Association of American Geographers, Washington, DC.
- Medyckyj-Scott, D., Newman, I., Ruggles, C. and Walker, D. (1991) *Metadata in the Geosciences*, Group D Publications, Loughborough.
- Milgram, P. and Kishino, F. (1994) A taxonomy of mixed reality visual displays. *IEICE Transactions on Information Systems Series D*, **77**, 1321–29.

## References

- Milgram, P., Takemura, H., Utsumi, A. and Kishino, F. (1994) Augmented reality: a class of displays on the reality-virtuality continuum. *Telemanipulator and Telepresence Technologies*, **2351**, 282–92.
- Miller, A.I. (1989) Imagery and intuition in creative scientific thinking: Albert Einstein's invention of the special theory of relativity. In *Creative People at Work: Twelve Cognitive Case Studies* (eds D.B. Wallace and H.E. Gruber), Oxford University Press, New York, **171**, 187.
- Miller, H.J. (2005) A measurement theory for time geography. *Geographical Analysis*, **37**, 17–45.
- Miller, H.J. and Han, J. (2001) *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis, London.
- Morehouse, S. (1995) GIS-based map compilation and generalization. In *GIS and Generalisation: Methodological and Practical Issues* (eds J.-C. Muller, J.-P. Lagrange and R. Weibel), Taylor & Francis, London, 21–30.
- Morton, G. (1966) A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing. IBM Canada Ltd., unpublished report.
- Moss, M.L. and Townsend, A.M. (2000) How telecommunications systems are transforming urban spaces. In *Cities in the Telecommunications Age: The Fracturing of Geographies* (eds J.O. Wheeler *et al.*), Routledge, New York, 31–41.
- Mountain, D. and Liarokapis, F. (2007) Mixed reality (MR) interfaces for mobile information systems. *Aslib Proceedings: New Information Perspectives*, **59**, 422–36.
- Mountain, D. and Raper, J. (2001) Spatio-temporal representations of individual human movement for personalising Location Based Services. Proceedings GISRUUK 2001, University of Glamorgan, UK, 579–82.
- Müller, J.-C. (1991) Generalization of spatial databases. In *Geographical Information Systems: Principles and Applications*, Vol **1** (eds D.J. Maguire, M.F. Goodchild and D.W. Rhind), Longman, Harlow, 457–75.
- Müller, J.-C., Lagrange, J.-P. and Weibel, R. (1995) *GIS and Generalisation: Methodological and Practical Issues*, Taylor & Francis, London.
- Muller, N.J. (2000) *Bluetooth Demystified*, McGraw-Hill, New York.
- Muller, N.J. (2003) *Wireless A to Z*, McGraw-Hill, New York.
- National Research Council (1997) *The Future of Spatial Data and Society*, National Research Council, Washington, D.C.
- Negroponte, N. (1995) *Being Digital*, Knopf, New York.
- Nielson, J. (2000) *Designing Web Usability*, New Riders, Indianapolis, IN.
- Niles, S. and Hanson, S. (2003) A new era of accessibility? *URISA Journal*, **15** (APA1), 35–41.
- NTIA (2000) *Falling Through the Net: Towards Digital Inclusion*, National Telecommunications and Information Administration, Washington, D.C.
- OGC (2004) OpenGIS Location Services (OpenLS): Core Services. Open GIS Consortium, <http://www.geospatial.org>.



- O'Hare, G.M.P. and O'Grady, M.J. (2003) Gulliver's Genie: a multi-agent system for ubiquitous and intelligent content delivery. *Computer Communications*, **26**, 117–87.
- Onsrud, H.J. (1995) Identifying unethical conduct in the use of GIS. *Cartography and Geographical Information Systems*, **22**, 90–7.
- Openshaw, S. and Abrahart, R. (2000) *Geocomputation*, Taylor & Francis, London.
- Openshaw, S. and Taylor, P.J. (1981) The modifiable areal unit problem. In *Quantitative Geography* (eds N. Wrigley and R.J. Bennett), Routledge, Henley-on-Thames, 60–70.
- Oppermann, R. and Specht, M. (2000) A context-sensitive nomadic exhibition guide. Proceedings of Second International Symposium on Handheld and Ubiquitous Computing, HUC 2000, Bristol, UK, 127–42.
- Ordnance Survey (2007) TOPO-96 Capture Policy, Ordnance Survey, Southampton.
- Pandey, S., Harbor, J. and Engel, B. (2000) *Internet-Based Geographic Information Systems*, The Urban and Regional Information Systems Association, Park Ridge, IL.
- Peano, G. (1890) Sur une courbe qui remplit toute une aire plane. *Mathematische Annalen*, **36** (A), 157–60.
- Peng, Z.R. (1999) An assessment framework for the development of Internet GIS. *Environment and Planning B: Planning and Design*, **26**, 111–32.
- Peng, Z.R. and Tsou, M.H. (2003) *Internet GIS: Distributed Geographic Information Services for the Internet and Wireless Networks*, John Wiley & Sons, Inc., Hoboken, NJ.
- Peuquet, D.J. (1994) It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, **84**, 441–61.
- Peuquet, D.J. (1999) Time in GIS and geographical databases. In *Geographical Information Systems Volume 1: Principles and Technical Issues* (eds P.A. Longley *et al.*), John Wiley & Sons, Inc., New York, 91–103.
- Pfeifer, T., Magedanz, T. and Hubener, S. (1998) Mobile Guide – Location-aware applications from the lab to the market. Proceedings of the Fifth International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services (IDMS'98), Oslo, Norway, 8–11.
- Pickles, J. (1995) *Ground Truth*, Guilford Press, New York.
- Plesa, M.A. and Cartwright, W. (2007) An evaluation of the effectiveness of non-realistic 3D graphics for city maps on small-screen devices. In *Dynamic and Mobile GIS: Investing Changes in Space and Time* (eds J. Drummond, R. Billen, E. Joao and D. Forrest), CRC Press, Boca Raton, FL, 141–59.

## References

- Priestnall, G. and Polmear, G. (2006) A spatially-aware mobile test bed for exploring and enhancing spatial literacy skills. Proceedings GIRUK 2006, University of Nottingham, UK, 387–91.
- Prim, R.C. (1957) Shortest connection networks and some generalisations. *Bell Systems Technical Journal*, **36**, 1389–401.
- Radcliffe, J. (2003) Death of copyright – long live patents. 2003 Cambridge Conference, Ordnance Survey, Southampton, UK, paper 4D.2B.
- Rakkolainen, I. and Vainio, T. (2001) A 3D city info for mobile users. *Computers & Graphics*, **25**, 619–25.
- Raper, J., Gartner, G., Karimi, H. and Rizos, C. (2007) A critical evaluation of location based services and their potential. *Journal of Location Based Services*, **1**, 5–45.
- Rappaport, T.S. (2002) *Wireless Communications Principles and Practice*, 2nd edn, Prentice Hall PTR, Upper Saddle River, NJ.
- Raptis, D., Tselios, N. and Avouris, N. (2005) Context-based design of mobile applications for museums: a survey of existing practices. Proceeding of MobileHCI 05, Salzburg, Austria, 153–60.
- Reichenbacher, T. (2003) Adaptive methods for mobile cartography. Proceedings of the 21st International Cartographic Conference, Durban, RSA, 1311–21.
- Reichenbacher, T. and Töllner, D. (2003) Design of an adaptive mobile geovisualization service. In *LBS and TeleCartography* (ed. G. Gartner), Geowissenschaftliche Mitteilungen, 17–23.
- Reiner, R. (2000) Crime and control in Britain. *Sociology*, **34**, 71–94.
- Rheingold, H. (2002) *Smart Mobs: The Next Social Revolution*, Basic Books, Cambridge, MA.
- Rhind, D. (1988) A GIS research agenda. *International Journal of Geographic Information Systems*, **2**, 23–8.
- Ricketson, S. (1984) *The Law of Intellectual Property*, Law Book Company, Sydney.
- Rist, T., Baldes, S. and Brandmeier, P. (2004) Aligning information browsing and exploration methods with a spatial navigation aid for mobile city visitors. Proceedings of AVI'04, Gallipoli, Italy, 226–30.
- Rizos, C. (2002) Introducing the Global Positioning System. In *Manual of Geospatial Science and Technology* (eds J. Bossler, J. Jensen, R. McMaster and C. Rizos), Taylor & Francis, London and New York.
- Robins, K. and Webster, F. (1999) *Times of Technoculture*, Routledge, London.
- Ruddle, R.A., Payne, S.J. and Jones, D.M. (1997) Navigating buildings in 'desk-top' virtual environments: experimental investigation using extended navigational experience. *Journal of Experimental Psychology: Applied*, **3**, 143–59.
- Ruthven, M. (2004) *Fundamentalism: The Search for Meaning*, Oxford University Press, Oxford.

- Samet, H. (1984) The quadtree and related hierarchical data structures. *Computing Surveys*, **16**, 187–260.
- Sampei, S. (2002) Modulation and demodulation techniques for wireless communication systems. In *Wireless Communications in the 21st Century* (eds M. Shafi, S. Ogose and T. Hattori), IEEE Press, Piscataway, NJ, 217–38.
- Schilit, B.N. and Theimer, M.M. (1994) Disseminating active map information to mobile hosts. *IEEE Network*, **8** (5), 22–32.
- Schilit, B.N., Adams, N.L. and Want, R. (1994) Context-aware computing applications. Proceedings of the Workshop on Mobile Computing Systems and Applications, IEEE Society, Santa Cruz, CA.
- Schilit, B.N., Adams, N.L., Gold, R. *et al.* (1993) The PARCTAB mobile computing system. Workshop on Workstation Operating Systems, 34–9.
- Schmalstieg, D. and Reitmayr, G. (2007) The world as a user interface: augmented reality for ubiquitous computing. *Location Based Service and TeleCartography*, Springer, 369–91.
- Schmidt, A. and Van Laerhoven, K. (2001) How to build smart appliances? *IEEE Personal Communications*, **8** (4), 66–71.
- Schmidt, A., Geigl, M. and Gellersen, H.-W. (1999) There is more to context than location. *Computers & Graphics*, **23**, 893–901.
- Sengupta, R. and Sieber, R. (2007) Geospatial agents, agents everywhere . . . . *Transactions in GIS*, **11**, 483–506.
- Sharma, J. (2002) *Oracle Spatial*, Oracle Corporation, Redwood Shores, CA.
- Shekhar, S. and Chawla, S. (2003) *Spatial Databases: A Tour*, Prentice Hall, Upper Saddle River, NJ.
- Shemyakin, F.N. (1962) General problems of orientation in space and space representations. In *Psychological Sciences in the USSR*, Vol **1** (eds B.G. Anan'yev *et al.*) NTIS Report No. TT62–11083, Office of Technical Services, Washington, D.C., 184–225.
- Shiode, N., Li, C., Batty, M. *et al.* (2004) The impact and penetration of location-based services. In *Telegeoinformatics: Location-Based Computing and Services* (eds H.A. Karimi and A. Hammad), CRC Press, Boca Raton, FL, 349–66.
- Siegel, A.W. and White, S.H. (1975) The development of spatial representations of large-scale environments. In *Advances in Child Development and Behavior* (ed. H.W. Reese), Academic Press, New York, 9–55.
- Simmons, G.F. (1963) *Introduction to Topology and Modern Analysis*, McGraw-Hill, Singapore.
- Slater, M., Steed, A. and Chrysanthou, Y. (2002) *Computer Graphics and Virtual Environments: From Realism to Real-Time*, Pearson, Harlow.
- Smith, B. and Mark, D.M. (2003) Do mountains exist? Towards an ontology of landforms. *Environment and Planning B*, **30**, 411–27.
- Smith, C. and Meyer, J. (2004) *3G Wireless With WiMAX and Wi-Fi: 802.16 and 802.11*, McGraw-Hill, New York.

## References

- Smith, J., Kealy, A., Mackaness, W. and Williamson, I. (2004) Spatial data infrastructure requirements for location based journey planning. *Transactions in GIS*, **8**, 23–44.
- Stern, E. and Leiser, D. (1988) Levels of spatial knowledge and urban travel modeling. *Geographical Analysis*, **20**, 140–55.
- Stevens, S.S. (1946) On the theory of scales of measurement. *Science*, **103**, 677–80.
- Story, M. and Congalton, R.G. (1986) Accuracy assessment: a users perspective. *Photogrammetric Engineering & Remote Sensing*, **52**, 397–9.
- Studer, R., Ankolekar, A., Hitzler, P. and Sure, Y. (2006) A semantic future for AI. *IEEE Intelligent Systems*, **21** (4), 8–9.
- Sui, D. (2005) Will ubicomp make GIS invisible? *Computers, Environment and Urban Systems*, **29**, 361–7.
- Taferner, M. and Bonek, E. (2002) *Wireless Internet Access Over GSM and UMTS*, Springer, Berlin.
- Tait, M.G. (2005) Implementing geoportals: applications and distributed GIS. *Computers, Environment and Urban Systems*, **29**, 33–47.
- Thapa, K. and Case, S. (2003) An indoor positioning service for bluetooth ad hoc networks. Proceedings 36th Midwest Instruction and Computing Symposium, Mankota, MN.
- Thorndyke, P.W. and Hayes-Roth, B. (1982) Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology*, **14**, 560–89.
- Thrift, N. and French, S. (2002) The automatic production of space. *Transactions of the Institute of British Geographers NS*, **27**, 309–35.
- Tlauka, M. and Wilson, P.N (1996) Orientation-free representations from navigation through a computer-simulated environment. *Environment and Behavior*, **28**, 647–64.
- Tobler, W. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**, 234–40.
- Tomlinson, R.F. (1984) Geographic information systems – a new frontier. *The Operational Geographer*, **5**, 31–35.
- Toye, E., Sharp, R., Madhavapeddy, A. *et al.* (2006) Interacting with mobile service: an evaluation of camera-phones and visual tags. *Personal and Ubiquitous Computing*, **11**, 97–106.
- Tsalgatidou, A. and Pitoura, E. (2001) Business models and transactions in mobile electronic commerce: requirements and properties. *Computer Networks*, **37**, 221–36.
- Tsou, M.H. and Battenfield, B.P. (2002) A dynamic architecture for distributed geographic information services. *Transactions in GIS*, **6**, 355–81.
- Tsui, P.H.Y. and Brimicombe, A.J. (1997) Adaptive recursive tessellations (ART) for Geographical Information Systems. *International Journal of Geographical Information Science*, **11**, 247–63.

- Tufte, E.R. (1983) *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT.
- Uitermark, H.T., van Oosterom, P.J.M., Mars, N.J.I. and Molenaar, M. (2005) Ontology-based integration of topographic data sets. *International Journal of Applied Earth Observation and Geoinformation*, **7**, 97–106.
- Urry, J. (2000) Mobile sociology. *British Journal of Sociology*, **51**, 185–203.
- van den Worm, J. (2001) Web maps design in practice. In *Web Cartography: Developments and Prospects* (eds M.-J. Kraak and A. Brown), Taylor & Francis, London, 87–108.
- van Dijk, J. and Hacker, K. (2003) The digital divide as a complex and dynamic phenomenon. *The Information Society*, **19**, 315–26.
- van Oort, P. (2006) *Spatial Data Quality: From Description to Application*, Wageningen University, Wageningen.
- van Oosterom, P.J.M. (1993) *Reactive Data Structures for Geographic Information Systems*, Oxford University Press, Oxford.
- Visintainer, F. and Darin, M. (2006) Preliminary requirements and strategies for map feedback. FeedMap Report D2.1, available from [http://www.ertico.com/download/feedmap\\_documents/Deliverables/FM031v21-WP2-D2.1-PreliminaryRequirements.pdf](http://www.ertico.com/download/feedmap_documents/Deliverables/FM031v21-WP2-D2.1-PreliminaryRequirements.pdf).
- Visser, U., Stuckenschmidt, G. and Vögele, T. (2002) Ontologies for geographic information processing. *Computers & Geosciences*, **28**, 103–17.
- Wachowicz, M. (1999) *Object-Oriented Design for Temporal GIS*, Taylor & Francis, London.
- Wagtendonk, A.J. and de Jeu, A.M. (2007) Sensible field computing: evaluating the use of mobile GIS methods in scientific fieldwork. *Photogrammetric Engineering & Remote Sensing*, **73**, 651–62.
- Walker, M., Turnbull, R. and Sim, N. (2007) Future mobile devices – an overview of emerging device trends, and the impact on future converged services. *BT Technology Journal*, **25** (2), 120–5.
- Want, R., Hopper, A., Falcao, V. and Gibbons, J. (1992) The active badge location system. *ACM Transactions on Information Systems*, **19**, 91–102.
- Ward, A., Jones, A. and Hopper, A. (1997) A new location technique for the active office. *IEEE Personal Communications*, **4** (5), 42–7.
- Wasinger, R., Stahl, C. and Krüger, A. (2003) M3I in a pedestrian navigation & exploration systems. *Mobile HCI 2003*, Springer-Verlag, Berlin, 481–5.
- Webster, F. (2004) *The Information Society Reader*, Routledge, London.
- Weibel, R. and Dutton, G. (1999) Generalising spatial data and dealing with multiple representations. In *Geographical Information Systems: Principles and Technical Issues*, Vol 1 (eds P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind), John Wiley & Sons, Inc., New York, 125–55.
- Weiser, M. (1991) The computer for the twenty-first century. *Scientific American*, September, 94–110.

## References

- Weiser, M. (1993) Some computer science issues in ubiquitous computing. *Communications of the ACM*, **36** (7), 94–83.
- Wheeler, J.O., Aoyama, Y. and Warf, B. (2000) *Cities in the Telecommunications Age: The Fracturing of Geographies*, Routledge, New York.
- Wickins, R. (1989) *Professional Liability*, Hong Kong University Press, Hong Kong.
- Wiener, N. (1968) *The Human Use of Human Beings: Cybernetics and Society*, Sphere Books, London.
- Williams, D.H. (2005) The deadline for the E911 mandate approaches. where do things stand? Directions Magazine, [www.directionsmag.com/article.php?article\\_id=2014&tv=1](http://www.directionsmag.com/article.php?article_id=2014&tv=1) viewed (15 Aug 2006).
- Williams, R., Stewart, J. and Slack, R. (2005) *Social Learning in Technological Innovation: Experimenting With Information and Communication Technologies*, Edward Elgar, Cheltenham, UK.
- Wilson, P.N. (1997) Use of virtual reality computing in spatial learning research. In *A Handbook of Spatial Research Paradigms and Methodologies, Volume 1: Spatial Cognition in the Child and Adult* (eds Foreman and Gillet), Lawrence Erlbaum Association Inc., Hove, 181–206.
- Wilson, J.P. and Fotheringham, A.S. (2008) *The Handbook of Geographical Information Science*, Blackwell, Oxford.
- Witmer, B.G., Bailey, J.H., Knerr, B.W. and Parsons, K.C. (1996) Virtual spaces and real world places: transfer of route knowledge. *International Journal of Human-Computer Studies*, **45**, 413–28.
- Worboys, M.F. (1994) Object-oriented approaches to geo-referenced information. *International Journal of Geographical Information Systems*, **8**, 385–99.
- Worboys, M.F. and Duckham, M. (2006) Monitoring qualitative spatio-temporal change for geosensor networks. *International Journal of Geographical Information Science*, **20**, 1087–108.
- Wray, R. (2007) One year's digital output would fill 161bn iPods. *Guardian Weekly*, **176** (13), 19.
- Yuan, M. (2008) Adding time to geographic information systems databases. In *The Handbook of Geographic Information Science* (eds J.P. Wilson and A.S. Fotheringham), Blackwell, Oxford, 169–84.
- Zeimpekis, V., Giaglis, G.M. and Lekakos, G. (2003) A taxonomy of indoor and outdoor positioning techniques for mobile location services. *ACM SIGecom Exchange*, **3**, 4, 19–27.
- Zipf, A. and Jöst, M. (2006) Implementing adaptive mobile GIS services based on ontologies: examples from pedestrian navigation support. *Computers, Environment and Urban Systems*, **30**, 784–79.

# Index

- Absolute Time of Arrival 194–5
- Access
  - controls 18–19, 34, 342
  - data 166–8
    - see also* Query processing
  - network connectivity 185–6, 223–4
- Access Points (APs) 48, 49
- Accuracy
  - data 155–7, 159–61
  - and granularity 159–60
  - position finding 119–20, 134, 169–70
    - GPS 145, 174, 181–2
    - needs 170–1, 179
    - sources of error 176–8
    - ‘wander’ 187–9
- Active Badge 203, 213
- ActiveBat 204
- Actor-networks 11
- Adaptive radio interface 46
- Addressable features 142
- AddressPoint 142
- Advanced Forward Link Trilateration (A-FLT) 197
- Aerial photographic interpretation (API) 146
- Agent-based technologies 29, 78–9, 80
- Angle of Arrival (AOA) 192–3
- Anti-Spoofing (AS) 180
- Apple iPhone 55
- Application areas 132
  - see also* Mobile e-Commerce (MeC)
- Application Service Providers (ASP) 315
- ArcGIS 97
- Architecture, *see* Systems architecture
- Assisted-GPS (A-GPS) 205
- Attribute data 91–4
  - accuracy 160–1
  - data structures 250
- Auditory mode of communication 281–2, 305
- Augmented reality (AR) 285
- Back-office activities 13
- Base mapping data 142
- Base Station Controller (BSC) 33, 40
- Base Station Subsystem (BSS) 40
- Base Transceiver Stations (BTS) 33–4, 40
- Batteries 55–6, 298, 317–18
- BlackBerry 55
- Bluetooth 49–51, 58, 200–1
- British Standards 326–7
- BT (British Telecom) Fusion 55
- B-tree index 251
- Business location 11–12, 13
- Business models 136, 314–18
- Call centres 13
- Carrier phase ranging 181
- Cartographic data, *see* Geographical information; Maps
- CCTV data collection 151
- CDMA, *see* Code Division Multiple Access (CDMA)
- Cell Global Identity (CGI) 189–92
- Cell-ID/TA 36, 170, 189–92
- Cell of Origin (COO) 189–92



## Index

- Cellular networks, *see* Mobile wireless networks
- cHTML (compact HTML) 62
- City Guide 213
- City structure 13
- City targetted services 15, 213, 214
- C-MAP 213
- Code Division Multiple Access (CDMA) 39, 40
- CDMA2000 44–5, 197
- Codepoint 142
- Cognitive factors 106, 299, 300–4
- Collaborative networks 7, 12–13, 317
- Commodification of urban space 18–19
- ComMotion 213
- Communication modes 135–6, 280–8, 300
  - auditory 281–2, 305
  - interactivity, *see* Interactivity
  - multimodal 311
  - presentation, *see* Presentation
  - semantics 299
  - user preferences 304–9
  - visualization, *see* Visualization
- Compact HTML (cHTML) 62
- Compatibility issues 105, 221–4
  - see also* Interoperability
- Computer networks 22–3
  - see also* Mobile wireless networks
- Configurational knowledge 301
- Connectivity matrices 270–1
- Connectivity, network 185–6, 223–4
- Constraints 297–8
- Content 138
  - see also* Mobile e-Commerce (MeC)
- Context awareness
  - context-adaptive approach 310–11
  - defined 211, 212–13
  - dynamic aspects 215–17, 228–34
  - environment as context 216, 217–21, 310–11
  - influence 123, 134–5, 209–11
  - interactions 215–17, 228–34
  - scope 212
  - situation context 220–1
  - technology as context 216, 221–4, 297–8
  - user as context 123, 134–5, 216, 225–8
- Continuously Operating Reference Station (CORS) 185
- Contract liability 335–6
- Coordinate systems 82–3
- Copyright 332–4
- Corporate headquarters 13
- Crime 19, 338–43
- Cyberguide 213
- Data
  - ambiguity 157
  - attributes 91–4
    - accuracy 160–1
    - data structures 250
  - collection 144–5
    - from CCTV 151
    - ground survey 148–9
    - remote sensing 146–8
    - for SatNav 149–50, 151, 153
    - Web crawlers 151–2
    - from wireless networks and mobile devices 152
  - compatibility 105
  - currency 140, 152–5, 253–62
  - definition 137
  - errors 156
  - fitness-for-use 157
  - granularity 139–40
    - and geometric accuracy 159–60
    - and position finding 170–1
    - and update frequency 153
  - layers 91
  - metadata, *see* Metadata
  - modelling 88
  - precision 156
  - presentation, *see* Presentation
  - primitives 89–91
  - quality 155–9
  - reliability 157
  - representation 100, 235–6, 261–2
  - requirements 133, 140, 170–1
  - resolution 84–5, 156
  - temporality 253–62
  - types 82–3, 142–3
    - dynamic 143
    - interval data 83



- mobile 143
- nominal data 82
- ordinal data 83
- ratio data 83
- static 143
- uncertainty 101–2, 157–9
- updating 150, 152–5
- vagueness 156
- Databases
  - geographical information
    - systems 89, 97
  - indexing of 250–3
  - object oriented 242–3
  - relational, *see* Relational database management systems (RDBMS)
  - subscriber 34, 41
  - temporal dimension 260–2
  - see also* Data sets
- Data cards 55
- Data definition language (DDL) 265
- Data modification language (DML) 265
- Data protection 337
- Data representation 261–2
- Data sets 138–44
  - accessing 166–7
  - accuracy 155–7, 159–61
  - examples 141–2
  - infrastructure 166–8
  - licenses 334
  - standard 326
  - types 142–3
  - volumes 138, 140–1
- Data structures 105
  - attribute data 250
  - object oriented models 242–4
  - raster data 240–2
  - topological data 245–50
  - vector data 237–40
- Data transmission rates 297
- ‘Dead reckoning’ 205
- ‘Death of distance’ 10
- Declarative knowledge 301
- Delivery tracking 132
- Density clustering 11–12
- Differential GPS (DGPS) 182–5
- ‘Digital city’ 9–14
- ‘Digital divide’ 8
- Digital watermarks 333
- Dijkstra’s algorithm 274
- DIME (Dual Independent Map Encoding) 75
- Display, *see* Presentation
- Domaine names 23
- Domain Name System (DNS) 23
- Douglas–Peucker algorithm 296
- Drive-by surveys 149–50
- Driving directions
  - route planning 95–6
  - SatNavs, *see* In-vehicle navigation systems
  - web-based maps 70
- Dual Independent Map Encoding (DIME) 75
- Duty of care 335–6
- E112 Mandate 125–6
- E911 Mandate 124–5
- e-auctions 316
- e-commerce
  - business models 314–18
  - see also* Mobile e-Commerce (MeC)
- ‘Edge cities’ 13
- EDGE (Enhanced Data Rates for Global Evolution) 42–3
- E-FLT (Enhanced Forward Link Trilateration) 197
- EGNOS (European Geostationary Navigation Overlay Service) 185
- e-Government 7, 327
- e-Government Interoperability Framework 327
- 8 Phase Shift Keying (8 PSK) 42
- Electronic spaces 10
- e-mail 16–17, 22
- e-market place 316
- Emergency services 124–6
- Energy consumption 298, 317–18
- Enhanced Cell-ID 191–2
- Enhanced Data Rates for Global Evolution (EDGE) 42–3
- Enhanced Forward Link Trilateration (E-FLT) 197
- Enhanced-Observed Time Difference (E-OTD) 196–7

## Index

- Environment as context 216, 217–21, 310–11
- E-OTD (Enhanced-Observed Time Difference) 196–7
- Ergonomic factors, *see* Communication modes; Human factors
- Errors, *see* Accuracy
- e-shops 315
- ESRI 97
- European Geostationary Navigation Overlay Service (EGNOS) 185
- Events-of-interest data 143
- Exhibition guiding 213
- Extended SQL 266–7, 268
- Face-to-face communication 12–13
- Field surveys 148–9
- ‘Fingerprinting’ 199
- Fixed relay networks (FRNs) 46
- Full-motion video 43
- Gaming 317
- Gateway Mobile Location Centre (GMLC) 328
- Gateway Mobile Switching Centre (GMSC) 34–5, 130
- Gaussian Minimum Shift Keying (GMSK) 42
- GDOP (Geometric Dilution of Precision) 178
- General Packet Radio Service (GPRS) 41, 42, 43
- Geocomputation 101
- Geodemographics 325
- Geofencing 132
- Geographical agglomeration 13
- Geographical differences 8, 11
- Geographical dispersal 13
- Geographical information  
  cognition 101  
  computational geometry 101  
  copyright 332–4  
  data representation 100, 235–6  
  liability 334–7  
  ontology 100  
  in relation to institutions and society 102  
  spatial analysis 102
- standards 327
- uncertainty 101–2, 157–9
- visualization 102, 290–3, 310–11
- Geographical information systems (GIS)  
  Anti-Spoofing (AS) 180  
  assisted 205–7  
  attribute data 91–4  
  background 56–7  
  carrier phase ranging 181  
  cognition 106  
  coordinate systems 82–3  
  corporate/enterprise systems 97  
  databases 89, 97  
  data layers 91  
  data modelling 88  
  data primitives 89–91  
  data resolution 84–5  
  data structures 105  
  data types 82–3  
    interval data 83  
    nominal data 82  
    ordinal data 83  
    ratio data 83  
  definition 73–4  
  design principles 173–4  
  Differential GPS (DGPS) 182–5  
  geographical spaces 10  
  Geography Markup Language (GML) 78  
  geometric data 237–43  
  Geometric Dilution of Precision (GDOP) 178  
  GeoMobility Server (GMS) 328–9  
  geoportals 167  
  gesture mode of input 283  
  GI-Engineering 103–7  
  GI-Science 98–102  
  GIS, *see* Geographical information systems (GIS)  
  global cities 12  
  Global Navigation Satellite Systems (GNSS) 172, 186  
  Global Positioning System (GPS) 145–6, 173–9  
    accuracy 145, 174, 181–2  
    sources of error 176–8  
    ‘wander’ 187–9  
  graphic design 87–9

- historical perspective 74–6
- indoor 185–6
- internet access 98
- latency 179, 187
- map projection 85–6
- measurement and scale 82–5
- monitoring stations 176, 179
- multipath interference 177–8
- open systems 77–8
- operating principles 174–8
- Precise Positioning Service (PPS) 180, 182
- pseudorange 180–1
- receiver clock offset 176
- receivers 145
- reliability 103
- requirements 104–6
- response times 105–6, 135
- route planning 95–6
- scalability 96–8, 104–5
- scientific principles 98–102
- Selective Availability (SA) 178
- signal measurement 180–1
- signal structure 179–80
- software 75
  - agent-based technologies 78–9, 80
  - distributed components 98
  - interoperability 76–7
  - middleware 78, 97
  - plug-ins and applets 98
  - products 81, 97
  - scripting languages 96
- Standard Positioning Service (SPS) 180
- symbology 87–9
- syntax 105
- system components 178–9
- thematic mapping 94–6
- user requirements 179
- web-based maps 68–73
- wireless 64
- Geographic tracking data 34
- Global System for Mobile communications (GSM) 40, 195, 196
- GMLC (Gateway Mobile Location Centre) 328
- GML (Geography Markup Language) 78
- GMSC (Gateway Mobile Switching Centre) 34–5, 130
- GMS (GeoMobility Server) 328–9
- GMSK (Gaussian Minimum Shift Keying) 42
- GNSS (Global Navigation Satellite Systems) 172, 186
- GNU General Public License 334
- GPRS (General Packet Radio Service) 41, 42, 43
- A-GPS (Assisted-GPS) 205
- GPS, *see* Global Positioning System (GPS)
- Granularity, data 139–40
  - and geometric accuracy 159–60
  - and position finding 170–1
  - and update frequency 153
- Graph data structures 267–74
- Graphic presentation 87–9, 283–8
- Ground survey 148–50
- GSM Association 327
- Guidance systems, *see* Route planning; Tourist guiding; Wayfinding
- GUIDE 213
- Gulliver's Genie 214
- Haptic interface 283
- High-end service functions 13
- High level operating systems (OS) 55
- High Performance Radio Local Area Network (HIPERLAN) 48
- HIPERLAN (High Performance Radio Local Area Network) 48
- Home Location Register (HLR) 33–4, 40, 41
- Hosting services 315
- HTTP (Hypertext Transfer Protocol) 25
- Human factors 299
  - see also* Communication modes; Social issues
- Hypertext Transfer Protocol (HTTP) 25
- IEEE standards 46–7, 48, 52
- Imagery 143, 283–8

## Index

- IMEI (International Mobile Equipment Identity) 342
- i-mode 61–2, 127
- In-car navigation systems, *see* In-vehicle navigation systems
- Indexing 250–3
- Individual privacy, *see* Privacy concerns
- Indoor GPS 185–6
- Industrial Scientific and Medical (ISM) bands 47
- Inequalities 7–8, 11
- Inertial navigation systems (INS) 145–6
- Informational economy 5–6
- Informational mobility 17
- Information brokerage 316
- Information concept 137
- Information society 2–9
- ‘Information superhighway’ 10
- Infrared Data Association (IrDA) 203
- INS (inertial navigation systems) 145–6
- Integrated Transport Network (ITN) 141–2, 153–5
- Intellectual mobility 17
- Interactivity 299–300
  - context awareness 215–17, 228–34
  - maps 291–2, 293
  - see also* Communication modes
- Interim Standard 136 (IS-136) 40, 41
- International Mobile Equipment Identity (IMEI) 342
- International standards 327
- Internet
  - development 22–4
  - e-commerce 314–18
  - GIS access 62–3, 98
  - mapping 331–2
  - service providers 23, 24
  - wireless access 58–62
- Internet Service Providers (ISPs) 23, 24
  - wireless 51–2
- Internet Society 23
- Interoperability 136
  - geographical information systems (GIS) 76–7
  - location-based services (LBS) 136
  - standards 326–9
- In-vehicle navigation systems 106, 118–19
  - data collection for 149–50, 151, 153
  - data presentation 287
  - as passive data collectors 150
  - security 338–43
- iPhone 55
- IS-95B 43
- ISO standards 327
- ISPs, *see* Internet Service Providers (ISPs)
- ITN (Integrated Transport Network) 141–2, 153–5
- Keitai 126–8
- LAAS (Local Area Augmentation System) 185
- Landline 142
- Landmark knowledge 301
- Land surveying 148–50
- Laser-based remote sensing 147–8
- Latency 179, 187
- Legal issues 136
  - copyright 332–4
  - liability 334–7
  - patents 330–2
- Liability 334–7
- LiDAR (light detection and ranging) 147–8
- LMUs (Location Measurement Units) 195, 196
- Local Area Augmentation System (LAAS) 185
- Local Area DGPS 185
- Location awareness 216, 218–20
- Location-based services (LBS)
  - definition 2, 122
  - development 79–80
  - historical perspective 110–14
  - as opposed to location-specific services 117, 121–2
  - query processing times 135
  - technical and social context 2–9
- Location-based tariffs 132
- Location finding, *see* Position finding
- Location Measurement Units (LMUs) 195, 196

- Logic layer 28–9, 30
- LTK information 139–40
- Maintenance tracking 132
- MapInfo 97, 265, 266–7
- MapPoint 81–2
- MapQuest 68–73
- Maps 283–4, 288–97
  - descriptive information 290
  - generalization 293–7
  - interactivity 291–2, 293
  - patents 331–2
  - scalability 82–5, 292–7
  - symbolology 289–90, 292
  - visualization 290–3, 310–11
- Market growth 321–3
- Market segmentation 325
- MAUP (modifiable areal unit problem) 102
- MBR (minimum bounding rectangle) 252–3
- MBWA (Mobile Broadband Wireless Access) 46–7
- Metadata 27, 144, 157, 161–6
  - standards 161, 327
  - statements 162–5
- Metric knowledge 301
- Microsoft Access 265
- Microsoft MapPoint 81–2
- Microsoft SQL-Server 265
- Middleware 78, 97, 328
- ‘Milieu of innovation’ 12–13
- MIMO (Multiple-Input-Multiple-Output) 46
- Minimum bounding rectangle (MBR) 252–3
- Mobile Broadband Wireless Access (MBWA) 46–7
- Mobile communication 15–17
- Mobile e-Commerce (MeC) 314
  - business models 319–21
  - emerging products 323–5
  - implications of using mobile devices 318–19
  - market growth 321–3
- Mobile GIS 63–4
- Mobile Guide 213
- ‘Mobile network society’ 17
- Mobile networks, *see* Mobile wireless networks
- Mobile phones 15–16
  - data collection from 152
  - functionality 53
  - internet access 42, 43, 53, 58–62, 127
  - location-based services 119–21, 127
  - position finding 124–6
  - security 342
  - standards 53
  - subscriber statistics 52–3
- Mobile Service Switching Centre (MSSC) 33–4, 35, 40, 130
- Mobile Station (MS) 40
- Mobile wireless networks 32–47
  - 1G 37–8
  - 2.5G networks 41–3
  - 2G 38–41
  - 3G 43–5, 57–8
  - 4G 45–7
  - authentication database 34
  - basic concepts 33–6
  - cellular concept 35–6
  - connectivity 223–4
  - convergence with Bluetooth 58
  - convergence with WiFi 57–8
  - early development 32–3
  - frequency re-use 35–6
  - internet access 58–62
  - subscriber database 34, 41
  - technical standards 40–52
  - wireless GIS 64
- Modes of communication, *see* Communication modes
- Modifiable areal unit problem (MAUP) 102
- Moving image clips 143
- MSSC (Mobile Service Switching Centre) 33–4, 35, 40, 130
- Multi-agent systems 78–9
- Multidimensional, *see* 3-d
- Multimap 331–2
- Multimedia applications 43
- Multimodal communication 287–8, 311
- Multipath interference 177–8
- Multiple-Input-Multiple-Output (MIMO) 46

## Index

- National GPS Network 185
- Navigation, *see* In-vehicle navigation systems; Wayfinding
- Navteq Standard Streets 141–2, 150
- Network analysis 270–4
- Network Cell Identification (Cell-ID) 189–92
- Network data structures 240, 267–74
- New mobility 15–20
  
- Object-oriented modelling 78, 242–4, 296
- OFDM (Orthogonal Frequency Division Multiplexing) 46, 48
- OmniSTAR 185
- Ontology 28, 30
- Open Geospatial Consortium (OGC) 77–8, 327
- Open GIS 77–8
- OpenLS specification 78, 328
- Open systems 77–8, 327–8
- Operating systems (OS) 55
- Operations and Support System (OSS) 40
- Oracle Spatial ORDBMS 267
- Ordnance Survey
  - copyright 333–4
  - data sets 141–2, 153–5
- Orthogonal Frequency Division Multiplexing (OFDM) 46, 48
- Outsourcing 317
- OWL (Web Ontology Language) 28
  
- Pacific Digital Cellular (PDC) 40, 41
- PANs (Personal Area Networks) 51, 58
- Participative content 317
- Passive data collectors 150
- Patents 330–2
- ‘Pay for placement’ 18
- Personal Area Networks (PANs) 51, 58
- Personal Digital Assistants (PDAs)
  - 53–4
  - data input 283
  - in ground survey 148–9
  - in wayfinding application 304–9
- Personal privacy, *see* Privacy concerns
- Personal profiles 227, 228
- Photogrammetry 146
  
- Points of interest (POI) 91–2, 143
- PointX 142
- Polygons 240, 246–7, 252–3, 266–7
- Portals 316–17
- Position finding 130–1, 133–4
  - accuracy 119–20, 134, 169–70
  - needs 170–1
  - active or passive 128
  - Bluetooth 200–1
  - device based 171–2
  - GPS, *see* Global Positioning System (GPS)
  - hybrid approaches 205–7
  - latency 170
  - network-based 172, 189–97
  - power consumption 170, 172
  - privacy concerns 127, 136
  - short range 197–204
  - technologies 171–3
- Positioning services provider (PSP) 131
- Postcodes 106–7
- Power consumption 170, 172, 298, 317–18
- Precise Positioning Service (PPS) 180, 182
- Presentation 135–6, 280–8
  - constraints 298
- Privacy concerns 19, 127, 136, 282, 337, 338
- PRN (Pseudo-Random Noise)
  - code 175, 181
- Procedural knowledge 301
- Product development 13
- Proof layer 30
- Pseudolites 186
- Pseudo-Random Noise (PRN)
  - code 175, 181
- Public safety 19
- Public spaces 13–14
- Publisher/subscribe model 320–1
  
- Quality assurance 159–61
- Query language 135, 262–3, 265–7, 268
- Query processing 224–5, 262–77
  - latency 179, 187
  - optimization 274–7
  - processing time 235–6
  - response times 135, 236–7

- Radar imaging 147
- Radio Frequency Identification (RFID) 201–2
- Raster data 240–2
- RDBMS, *see* Relational database management systems (RDBMS)
- RDF Schema 28, 29–30
- Receiver clock offset 176
- Recursive tessellation 270
- Relational algebra (RA) 263
- Relational database management systems (RDBMS) 97, 250
  - geometric data 237, 238
  - query optimization 274–7
  - query processing 262–77
  - time representation 261–2
- Relational knowledge 301
- Remote sensing 146–8
  - active 147–8
  - passive 147
- Resource Description Framework (RDF) 28, 29–30
- Response times 135
- Response types, *see* Modes of communication; Presentation
- RFID (Radio Frequency Identification) 201–2
- Route knowledge 301
- Route planning 95–6
  - SatNavs, *see* In-vehicle navigation systems
- Routers 49
- R-tree index 252
- Satellite imaging 146–7
- SatNavs, *see* In-vehicle navigation systems
- Scalability
  - applications 96–8
  - maps 82–5, 292–7
  - technology 104–5
- Scripting languages 96
- SDI (Spatial Data Infrastructure) 166–7
- SDSS (spatial decision support systems) 101
- Search engines 18, 26
  - companies 316
  - for data collection 151–2
  - gazetteer-driven 115–16
- Secondary data 158
- Secure Digital (SD) card keys 342
- Security issues 19, 338–42
  - see also* Privacy concerns
- Selective Availability (SA) 178
- Semantics 299
- Semantic Web 26–31
- Service brokers 320
- Services, business, *see* Mobile e-Commerce (MeC)
- Services-of-interest data 143
- Short Message Service (SMS) 34, 53, 282
- Signal measurement 180–1
- Signal reception 185–6, 223–4
- Signal structure 179–80
- SIM (Subscriber Identity Module) 40
- Situation context 220–1
- Smartphones 54–5
- SMS (Short Message Service) 34, 53, 282
- Social context 2–9
- Social exclusion 7–8
- Social issues 102, 136, 337–42
- Social mobility 15
- Social networking 17, 25–6, 317
- ‘soft clipping’ 232, 275, 276, 320
- Software 75, 97
  - agent-based technologies 78–9, 80
  - distributed components 98
  - interoperability 76–7
  - middleware 78, 97
  - plug-ins and applets 98
  - products 81, 97
  - scripting languages 96
- Sound clips 143
- Space–time envelopes 232–4, 320
- Space–time paths 257–60
- Space versus time views 254–6
- Spacial databases, *see* Databases
- Spacial queries 262–77
  - optimization 274–7
  - processing time 235–6
  - response times 236–7
- Spanning trees 270–4
- Spatial analysis 102

## Index

- Spatial Data Infrastructure (SDI) 166–7
- Spatial decision support systems (SDSS) 101
- Spatial dependency 102
- Spatial knowledge 301–4
- Spatial literacy skills 214
- Speech recognition 281
- Speech synthesis 281
- SQL (structured query language) 263, 265–7, 268
- Standard Positioning Service (SPS) 180, 182
- Standards 326–9
  - metadata 161
  - mobile wireless networks 40–52
- Storage capacity 55–6
- Structured query language (SQL) 263, 265–7, 268
- Stylus mode of input 283
- Subscriber Identity Module (SIM) 40
- Subscription-based services 320
- Surveillance 18
- Survey knowledge 301
- Symbology 87–9, 289–90, 292
- System context 223
- Systems architecture 97
  - interoperable, open, distributed systems 143–4
  - main components 128–31
  - Wireless Application Protocol (WAP) 60–1
- Taxonomies 27
- TCP/IP, *see* Transmission Control Protocol/Internet Protocol (TCP/IP)
- TDMA (Time Division Multiple Access) 39–40
- TDOA (Time Difference of Arrival) 195–6
- Technical constraints 297–8
- Technological determinism 10
- Technology as context 104–5, 216, 221–4
- Telecommunications
  - infrastructure 11–12
- TeleGeoInformation 151
- Terrorism 19
- Texting 16–17
- Text mode of communication 282
- Thematic mapping 94–6
- 3-d
  - maps 293
  - position fixing 145
  - representation 284–8
  - visualization 146, 147
- Time of Arrival (TOA) 194–5
- Time as context 216, 220, 253–62
- Time Difference of Arrival (TDOA) 195–6
- Time Division Multiple Access (TDMA) 39–40
- Time Division-Synchronous Code Division Multiple Access (TD-SCDMA) 44, 45
- Time representation 261–2
- Time-series data 256
- TOA (Time of Arrival) 194–5
- Topological data 245–50, 259–60
- Topological location context 219
- Tourist guiding 213, 214
- Tracking systems 34, 132
- Transmission Control Protocol/Internet Protocol (TCP/IP) 22, 24
  - over wireless mobile networks 59
- Travel MATE 213
- Trust layer 30
- Trust services 317
- Ultrasound-based positioning 204
- Ultra Wide Band (UWB) 204
- UMTS Terrestrial Radio Access Network (UTRAN) 44
- Unicode 29
- Uniform Resource Identifier (URI) 29
- Universal Mobile Telecommunication System (UMTS) 43–4
- Updating data 150, 152–5, 262
- Urbanization 9–10
- URL (Uniform Resource Locator) 25
- Usability 157
  - see also* Human factors
- User-adaptive content 225–8
- User characteristics 216, 225–6
- User context 123, 134–5, 216, 220–1, 225–8



- User preferences 226–8, 299–300, 304–9, 310–11
- UWB (Ultra Wide Band) 204
- Value-added selling 324
- Value-chain integration 317
- Variable-Spreading-Factor Spread OFDM (VSF-Spread OFDM) 46
- Vector data 237–40
- Version management 262  
*see also* Updating data
- ViaMichelin 68–73
- Video clips 143
- Video conferencing 43
- Video presentation 283
- Virtual reality (VR) 143, 284–5, 302, 304
- Visitor Location Registration (VLR) 40, 41
- Visualization
  - context-adaptive 310–11
  - geographical information 102
  - maps 85–6, 290–3, 310–11
  - 3-D 146, 147
  - user preferences 310–11
- VLR (Visitor Location Registration) 40, 41
- Voice communication 281–2, 305
- Voice over Internet Protocol (VoIP) 58
- Voice User Interface (VUI) 281
- VoiceXML 281
- VR (virtual reality) 143, 284–5, 302, 304
- VSF-Spread OFDM (Variable-Spreading-Factor Spread OFDM) 46
- VUI (Voice User Interface) 281
- WAAS (Wide Area Augmentation System) 185
- WADGPS (Wide Area DGPS) 184–5
- ‘Wander’ 187–9
- WAP Transaction Protocol (WTP) 60
- WAP (Wireless Application Protocol) 41, 53, 59–61
- Wayfinding 132, 230–1, 304–9
- W-CDMA (Wideband CDMA) 43–4
- Web Ontology Language (OWL) 28
- Web, *see* World Wide Web (WWW/ Web)
- Wide Area Augmentation System (WAAS) 185
- Wide Area DGPS (WADGPS) 184–5
- Wideband CDMA (W-CDMA) 43–4
- WiFi, *see* Wireless Local Area Networks (WLANs)
- WiMax (Worldwide Interoperability for Microwave Access) 51–2
- Wireless Application Protocol (WAP) 41, 53, 59–61
- Wireless Fidelity (WiFi), *see* Wireless Local Area Networks (WLANs)
- Wireless GIS 63–4
- Wireless Internet Service Providers (WISPs) 51–2
- Wireless Local Area Networks (WLANs) 47–9
  - Access Points (APs) 49
  - convergence with wireless mobile networks 57–8
  - ‘Hot Spots’ 49, 55
  - position finding 198–200
  - wireless GIS 64
  - in wireless mobile devices 56
- Wireless MAN (WMAN) 51–2
- Wireless Markup Language (WML) 60, 61
- Wireless mobile devices 52–6
  - battery technologies 55–6, 298, 317–18
  - PDAs, *see* Personal Digital Assistants (PDAs)
  - storage capacity 56
  - use of WiFi 56
  - see also* Mobile phones
- Wireless (mobile) GIS 63–4
- Wireless mobile internet 58–62
- Wireless mobile networks, *see* Mobile wireless networks
- Wireless Transport Layer Security (WTLS) 60
- Wireless USB sticks 55
- WISPs (Wireless Internet Service Providers) 51–2
- WLANs, *see* Wireless Local Area Networks (WLANs)

## Index

- WMAN (Wireless MAN) 51–2
- WMLScript 60, 61
- WML (Wireless Markup Language) 60, 61
- Worldwide Interoperability for Microwave Access (WiMax) 51–2
- World wide web (WWW/Web)
  - browsers 23, 60
  - crawlers 151–2
  - development 23
  - evolution 25–6
  - location-specific services 115–18
  - mobile phone display 127
  - search engines 18, 26
- Semantic Web 26–31
  - social networking 25–6
  - taxonomies 27
  - technology 24–5
- Web 2.0 26
  - web-based GIS 62–3, 98
  - web-based maps 68–73
- WTLS (Wireless Transport Layer Security) 60
- XML (eXtensible Markup Language) 27, 28, 29, 78
- Zip codes 106–7