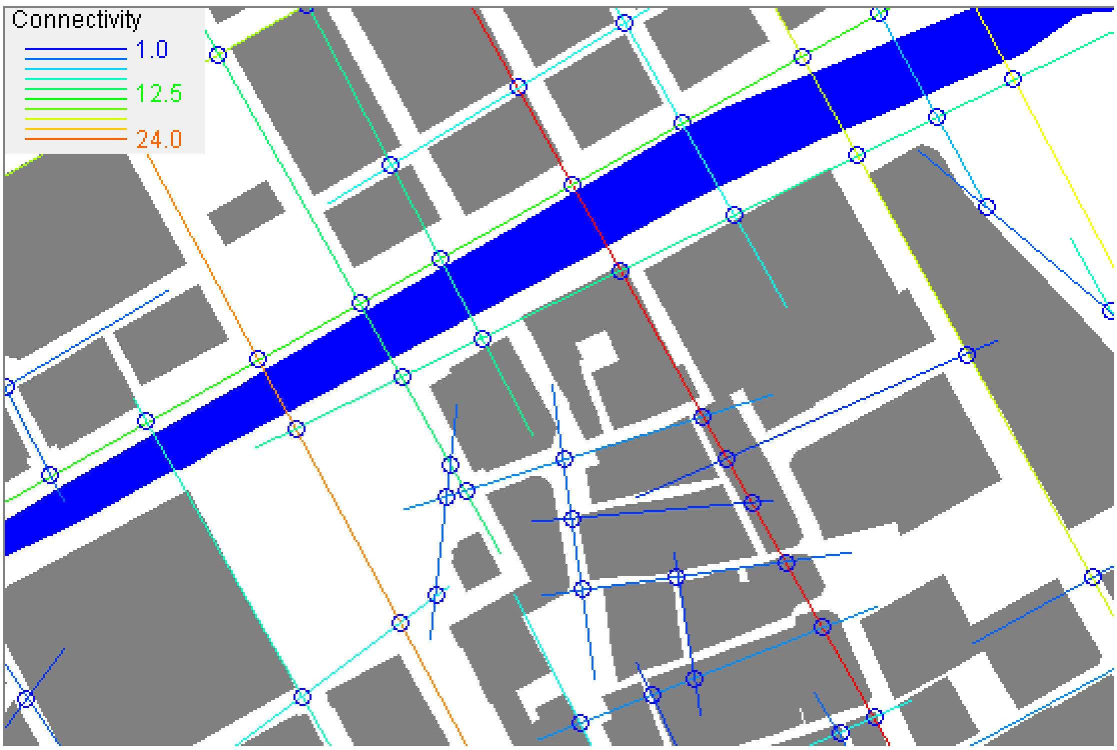
Geospatial Analysis

A Comprehensive Guide to Principles
Techniques and Software Tools

6th edition, 2018

Dr Michael J de Smith, Prof Michael F Goodchild
Prof Paul A Longley & Associates

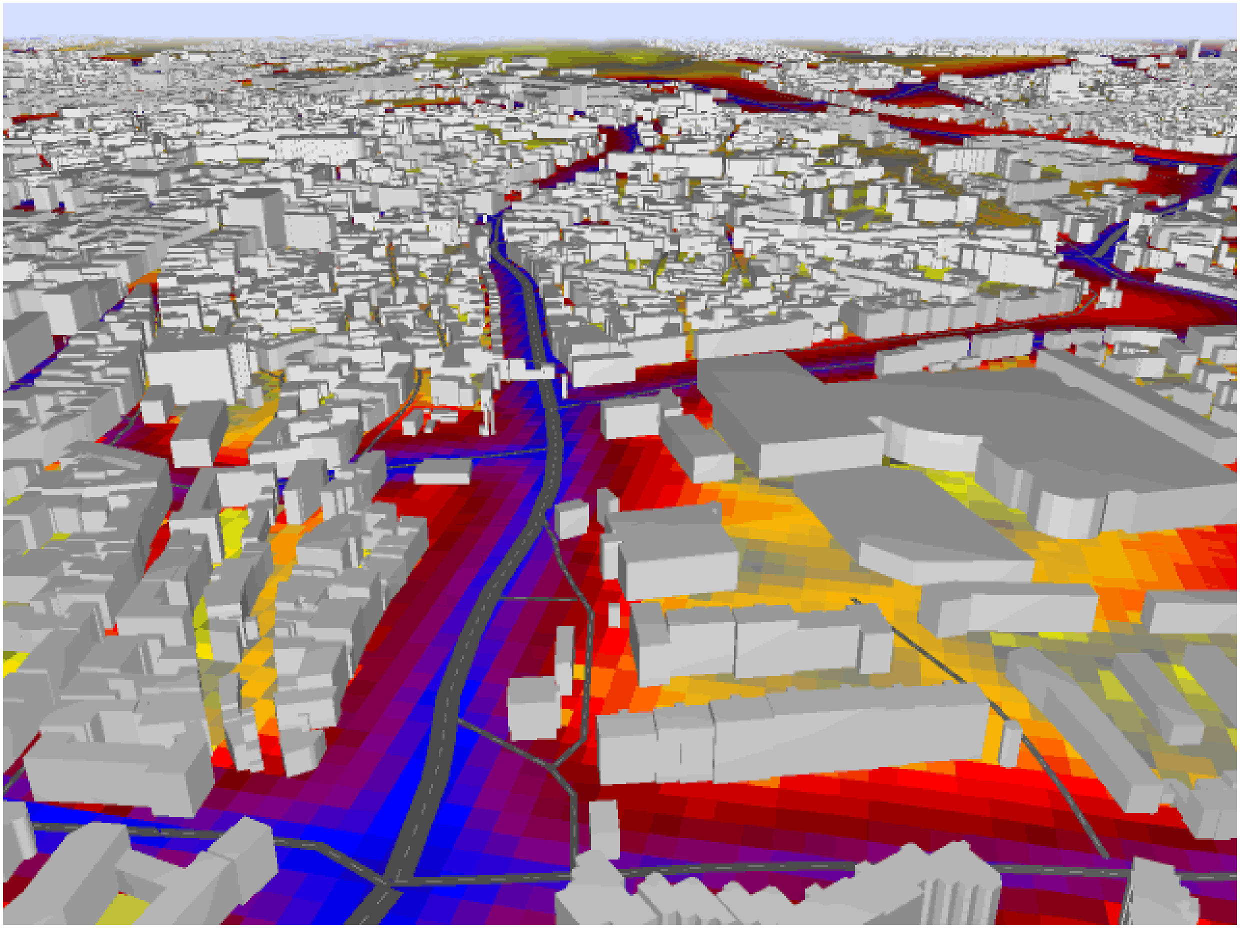


For more details please refer to the Guide’s website: www.spatialanalysisonline.com

ISBN-13

978-1-912556-03-8 Hardback
978-1-912556-04-5 Paperback
978-1-912556-05-2 ebook

3D visualization of modeled road-related noise levels in an urbanized area



Visualization using CadnaA software, courtesy of Accon GmbH & DataKustik GmbH

Optimized service center location and allocated demand
Tripolis, in Arcadia, Greece

Coverage or p center location optimization problem. See Section 7.4.2 for more details. Map produced using S-Distance software (2006), courtesy of S A Sirigos

Acknowledgements

The authors would like to express their particular thanks to the following individuals and organizations: Accon GmbH, Greifenberg, Germany for permission to use the noise mapping images on the inside cover of this Guide and in Figure 3-4; Prof D Martin for permission to use Figure 4-19 and Figure 4-20; Prof D Dorling and colleagues for permission to use Figure 4-50 and Figure 4-52; Dr K McGarigal for permission to use the Fragstats summary in Section 5.3.4; Prof Wei Li and colleagues, China National Laboratory for High Speed Railway Construction for permission to use figures 4-69A and 4-69B; Dr H Kristinsson, Faculty of Engineering, University of Iceland for permission to use Figure 4-70; Dr S Rana, formerly of the Center for Transport Studies, University College London for permission to use Figure 6-24; Prof B Jiang, Department of Technology and Built Environment of University of Gävle, Sweden for permission to use the Axwoman software and sample data in Section 6.3.3.2 and the inside cover page; Dr G Dubois, European Commission (EC), Joint Research Center Directorate (DG JRC) for comments on parts of Chapter 6 and permission to use material from the original AI-Geostats website; Geovariances (France) for provision of an evaluation copy of their Isatis geostatistical software; F O'Sullivan for use of Figure 6-41; Profs A Okabe, K Okunuki and S Shiode (Center for Spatial Information Science, Tokyo University, Japan) for use of their SANET software and sample data; and S A Sirigos, University of Thesally, Greece for permission to use his Tripolis dataset, the provision of his S-Distance software, and comments on part of Chapter 7.

Sections 8.1 and 8.2 of Chapter 8 are substantially derived from material researched and written by Christian Castle and Andrew Crooks (updated for the latest editions by Andrew) with the financial support of the Economic and Social Research Council (ESRC), Camden Primary Care Trust (PCT), and the Greater London Authority (GLA) Economics Unit. Chapter 9 has been substantially written by Guy Lansley of the UCL Consumer Data Research Centre. We would also like to express our thanks to the many users of the book and website for their comments, suggestions and occasionally, corrections. Particular thanks for corrections go to Bryan Thrall, Juanita Francis-Begay, Paul Johnson and Prof Liem Tran.

Table of Contents

Foreword

This 6th edition includes the following principal changes from the 2015 edition: weblinks and associated information have been updated and/or added or removed as appropriate; errata identified in the 5th edition have been corrected; new material in Chapter 4.4 has been added with thanks to China National Railways; and the final Chapter has been substantially expanded, with a focus on Big Data and the issues such data raise for geospatial research and analysis. Note that new versions of software tools referenced in the text have not been re-run and re-tested so readers should refer to the latest versions of these software tools and their documentation where appropriate.

Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools originated as material to accompany the spatial analysis module of MSc programmes at University College London delivered by the principal author, Dr Mike de Smith. The project was discussed with Professors Longley and Goodchild. They kindly agreed to contribute to the contents of the Guide itself. As such, this Guide may be seen as a companion to the pioneering book on Geographic Information Systems and Science (now changed to Science and Systems) by Longley, Goodchild, Maguire and Rhind, particularly the chapters that deal with spatial analysis and modeling. Their participation has also facilitated links with broader "spatial literacy" and spatial analysis programmes. Notable amongst these are the GIS&T Body of Knowledge materials provided by the Association of American Geographers together with the spatial educational programmes provided through UCL and UCSB. The formats in which this Guide has been published have proved to be extremely popular, encouraging us to seek to improve and extend the material and associated resources further. Many academics and industry professionals have provided helpful comments on previous editions, and universities in several parts of the world have now developed courses which make use of the Guide and the accompanying resources. Workshops based on these materials have been run in Ireland, the USA, East Africa, Italy and Japan, and a Chinese version of the Guide (2nd ed.) has been published by the Publishing House of Electronics Industry, Beijing, PRC, www.phei.com.cn in 2009.

A unique, ongoing, feature of this Guide is its independent evaluation of software, in particular the set of readily available tools and packages for conducting various forms of geospatial analysis. To our knowledge, there is no similarly extensive resource that is available in printed or electronic form. We remain convinced that there is a need for guidance on where to find and how to apply selected tools. Inevitably, some topics have been omitted, primarily where there is little or no readily available commercial or open source software to support particular analytical operations. Other topics, whilst included, have been covered relatively briefly and/or with limited examples, reflecting the inevitable constraints of time and the authors' limited access to some of the available software resources. Every effort has been made to ensure the information provided is up-to-date, accurate, compact, comprehensive and representative - we do not claim it to be exhaustive. However, with fast-moving changes in the software industry and in the development of new techniques and data sources it would be impractical and uneconomic to publish the material in a conventional manner. Accordingly the Guide has been prepared without intermediary typesetting. We would like to thank all those users of the book, for their comments and suggestions which have assisted us in producing this latest edition.

Mike de Smith, UK, Mike Goodchild, USA, Paul Longley, UK, 2018 (6th edition)

Chapter 1

1 Introduction and terminology

In this Guide we address the full spectrum of spatial analysis and associated modeling techniques that are provided within currently available and widely used geographic information systems (GIS) and associated software. Collectively such techniques and tools are often now described as *geospatial analysis*, although we use the more common form, *spatial analysis*, in most of our discussions.

The term 'GIS' is widely attributed to Roger Tomlinson and colleagues, who used it in 1963 to describe their activities in building a digital natural resource inventory system for Canada (Tomlinson 1967, 1970). The history of the field has been charted in an edited volume by Foresman (1998) containing contributions by many of its early protagonists. A timeline of many of the formative influences upon the field is provided in Longley *et al.* (2015, p20). The research makes the unassailable point that the success of GIS as an area of activity has been driven by the success of its applications in solving real world problems.

In order to cover such a wide range of topics, this Guide has been divided into a number of main sections or chapters. These are then further subdivided, in part to identify distinct topics as closely as possible, facilitating the creation of a web site from the text of the Guide. Hyperlinks embedded within the document enable users of the web and PDF versions of this document to navigate around the Guide and to external sources of information, data, software, maps, and reading materials.

Chapter 2 provides an introduction to spatial thinking, described by some as "spatial literacy", and addresses the central issues and problems associated with spatial data that need to be considered in any analytical exercise. In practice, real-world applications are likely to be governed by the organizational *practices* and *procedures* that prevail with respect to particular *places*. Not only are there wide differences in the volume and remit of data that the public sector collects about population characteristics in different parts of the world, but there are differences in the ways in which data are collected, assembled and disseminated (e.g. general purpose censuses versus statistical modeling of social surveys, property registers and tax payments). Data collected by the private sector, often as a result of the use of services that automatically gather information on events, locations and individuals, present a further challenge (see Chapter 9 for an extended discussion of this issue).

There are also differences in the ways in which different data holdings can legally be merged and the purposes for which data may be used — particularly with regard to health and law enforcement data. Finally, there are geographical differences in the cost of geographically referenced data. Some organizations, such as the US Geological Survey, are bound by statute to limit charges for data to sundry costs such as media used for delivering data while others, such as most national mapping organizations in Europe, are required to exact much heavier charges in order to recoup much or all of the cost of data creation. Analysts may already be aware of these contextual considerations through local knowledge, and other considerations may become apparent through browsing metadata catalogs. GIS applications must by definition be sensitive to context, since they represent unique locations on the Earth's surface.

This initial discussion is followed in Chapter 3 by an examination of the methodological background to GIS analysis. Initially we examine a number of formal methodologies and then apply ideas drawn from these to the specific case of spatial analysis. A process known by its initials, PPDAC (Problem, Plan, Data, Analysis, Conclusions) is described as a methodological framework that may be applied to a very wide range of spatial analysis problems and projects. We conclude Chapter 3 with a brief discussion on model-building, with particular reference to the various types of model that can be constructed to address geospatial problems.

Subsequent Chapters present the various analytical methods supported within widely available software tools. The majority of the methods described in Chapter 4 (Building blocks of spatial analysis) and many of those in Chapter 6 (Surface and field analysis) are implemented as standard facilities in modern commercial GIS packages such as ArcGIS, MapInfo, Manifold, TNTMips and Intergraph. Many are also provided in more specialized GIS products such as Idrisi, GRASS, QGIS and ENVI. Note that GRASS and QGIS (which includes GRASS in its download kit) are OpenSource.

In addition we discuss a number of more specialized tools, designed to address the needs of specific sectors or technical problems that are otherwise not well-supported within the core GIS packages at present. Chapter 5, which focuses on statistical methods, and Chapter 7 and Chapter 8 which address Network and Location Analysis, and Geocomputation, are much less commonly supported in GIS packages, but may provide loose- or close-coupling with such systems, depending upon the application area. In all instances we provide detailed examples and commentary on software tools that are readily available. The final Chapter (Chapter 9) addresses issues associated with so-called Big Data.

As noted above, throughout this Guide examples are drawn from and refer to specific products — these have been selected purely as examples and are not intended as recommendations. Extensive use has also been made of tabulated information, providing abbreviated summaries of techniques and formulas for reasons of both compactness and coverage. These tables are designed to provide a quick reference to the various topics covered and are, therefore, not intended as a substitute for fuller details on the various items covered. We provide limited discussion of novel 2D and 3D mapping facilities, and the support for digital globe formats (e.g. KML and KMZ), which is increasingly being embedded into general-purpose and specialized data analysis toolsets. These developments confirm the trend towards integration of geospatial data and presentation layers into mainstream software systems and services, both terrestrial and planetary (see, for example, the KML images of Mars DEMs developed by Google as part of the Google Earth project).

Just as all datasets and software packages contain errors, known and unknown, so too do all books and websites, and the authors of this Guide expect that there will be errors despite our best efforts to remove these! Some may be genuine errors or misprints, whilst others may reflect our use of specific versions of software packages and their documentation. Inevitably with respect to the latter, new versions of the packages that we have used to illustrate this Guide will have appeared even before publication, so specific examples, illustrations and comments on scope or restrictions may have been superseded. In all cases the user should review the documentation provided with the software version they plan to use, check release notes for changes and known bugs, and look at any relevant online services (e.g. user/developer forums and blogs on the web) for additional materials and insights.

The web version of this Guide may be accessed via the associated Internet site: www.spatialanalysisonline.com. The contents and sample sections of the PDF version may also be accessed from this site. In both cases the information is regularly updated. The Internet is now well established as society's principal mode of information exchange and most GIS users are accustomed to searching for material that can easily be customized to specific needs. Our objective for such users is to provide an independent, reliable and authoritative first port of call for conceptual, technical, software and applications material that addresses the panoply of new user requirements.

1.1 Spatial analysis, GIS and software tools

Our objective in producing this Guide is to be comprehensive in terms of concepts and techniques (but not necessarily exhaustive), representative and independent in terms of software tools, and above all practical in terms of application and implementation. However, we believe that it is no longer appropriate to think of a standard, discipline-specific textbook as capable of satisfying every kind of new user need. Accordingly, an innovative feature of our approach here is the range of formats and channels through which we disseminate the material.

Given the vast range of spatial analysis techniques that have been developed over the past half century many topics can only be covered to a limited depth, whilst others have been omitted because they are not implemented in current mainstream GIS products. This is a rapidly changing field and increasingly GIS packages are including analytical tools as standard built-in facilities or as optional *toolsets*, *add-ins* or *analysts*. In many instances such facilities are provided by the original software suppliers (commercial vendors or collaborative non-commercial development teams) whilst in other cases facilities have been developed and are provided by third parties. Many products offer software development kits (SDKs), programming languages and language support, scripting facilities and/or special interfaces for developing one's own analytical tools or variants.

In addition, a wide variety of web-based or web-deployed tools have become available, enabling datasets to be analyzed and mapped, including dynamic interaction and drill-down capabilities, without the need for local GIS software installation. These tools include the widespread use of web-based Java, Javascript, AJAX and HTML5 applications, and interactive Virtual Globe explorers, some of which are described in this Guide. They provide an illustration of the direction that many toolset and service providers are taking.

Throughout this Guide there are numerous examples of the use of software tools that facilitate geospatial analysis. In addition, some subsections of the Guide and the software section of the accompanying website, provide summary information about such tools and links to their suppliers. Commercial software products rarely provide access to source code or full details of the algorithms employed. Typically they provide references to books and articles on which procedures are based, coupled with online help and "white papers" describing their parameters and applications. This means that results produced using one package on a given dataset can rarely be exactly matched to those produced using any other package or through hand-crafted coding. There are many reasons for these inconsistencies including: differences in the software architectures of the various packages and the algorithms used to implement individual methods; errors in the source materials or their interpretation; coding errors; inconsistencies arising out of the ways in which different GIS packages model, store and manipulate information; and differing treatments of special cases (e.g. missing values, boundaries, adjacency, obstacles, distance computations etc.).

Non-commercial packages sometimes provide source code and test data for some or all of the analytical functions provided, although it is important to understand that "non-commercial" often does not mean that users can download the full source code. Source code greatly aids understanding, reproducibility and further development. Such software will often also provide details of known bugs and restrictions associated with functions — although this information may also be provided with commercial products it is generally less transparent. In this respect non-commercial software may meet the requirements of scientific rigor more fully than many commercial offerings, but is often provided with limited documentation, training tools, cross-platform testing and/or technical support, and thus is generally more demanding on the users and system administrators. In many instances open source and similar not-for-profit GIS software may also be less generic, focusing on a particular form of spatial representation (e.g. a grid or raster spatial model). Like some commercial software, it may also be designed with particular application areas in mind, such as addressing problems in hydrology or epidemiology.

The process of selecting software tools encourages us to ask: (i) "what is meant by geospatial analysis techniques?" and (ii) "what should we consider to be GIS software?" To some extent the answer to the second question is the simpler, if we are prepared to be guided by self-selection. For our purposes we focus principally on products that claim to provide geographic information systems capabilities, supporting at least 2D mapping (display and output) of raster (grid based) and/or vector (point/line/polygon based) data, with a minimum of basic map manipulation facilities. We concentrate our review on a number of the products most widely used or with the most readily accessible analytical facilities. This leads us beyond the realm of pure GIS. For example: we use examples drawn from packages that do not directly provide mapping facilities (e.g. the now rather outdated software called Crimestat) but which provide input and/or output in widely used GIS map-able formats; products that include some mapping facilities but whose primary purpose is spatial or spatio-temporal data exploration and analysis (e.g. GS+, GeoDa, PySal); and products that are general- or special-purpose analytical engines incorporating mapping capabilities (e.g. MATLab with the Mapping Toolbox, WinBUGS with GeoBUGS) — for more details on these and other example software tools, please see the website page:

www.spatialanalysisonline.com/software.html

The more difficult of the two questions above is the first — what should be considered as "geospatial analysis"? In conceptual terms, the phrase identifies the subset of techniques that are applicable when, as a minimum, data can be referenced on a two-dimensional frame and relate to terrestrial activities. The results of geospatial analysis will change if the location or extent of the frame changes, or if objects are repositioned within it: if they do not, then "everywhere is nowhere", location is unimportant, and it is simpler and more appropriate to use conventional, *aspatial*, techniques.

Many GIS products apply the term (geo)spatial analysis in a very narrow context. In the case of vector-based GIS this typically means operations such as: map overlay (combining two or more maps or map layers according to predefined rules); simple buffering (identifying regions of a map within a specified distance of one or more features, such as towns, roads or rivers); and similar basic operations. This reflects (and is reflected in) the use of the term *spatial analysis* within the Open Geospatial Consortium (OGC) "simple feature specifications" (see further Table 4-2). For raster-based GIS, widely used in the environmental sciences and remote sensing, this typically means a range of actions applied to the grid cells of one or more maps (or images) often involving filtering and/or algebraic operations (*map algebra*). These techniques involve processing one or more raster layers according to simple rules resulting in a new map layer, for example replacing each cell value with some combination of its neighbors' values, or computing the sum or difference of specific attribute values for each grid cell in two matching raster datasets. Descriptive statistics, such as cell counts, means, variances, maxima, minima, cumulative values, frequencies and a number of other measures and distance computations are also often included in this generic term "spatial analysis".

However, at this point only the most basic of facilities have been included, albeit those that may be the most frequently used by the greatest number of GIS professionals. To this initial set must be added a large variety of statistical techniques (descriptive, exploratory, explanatory and predictive) that have been designed specifically for spatial and spatio-temporal data. Today such techniques are of great importance in social and political sciences, despite the fact that their origins may often be traced back to problems in the environmental and life sciences, in particular ecology, geology and epidemiology. It is also to be noted that spatial statistics is largely an observational science (like astronomy) rather than an experimental science (like agronomy or pharmaceutical research). This aspect of geospatial science has important implications for analysis, particularly the application of a range of statistical methods to spatial problems.

Limiting the definition of geospatial analysis to 2D mapping operations and spatial statistics remains too restrictive for our purposes. There are other very important areas to be considered. These include: surface analysis —in particular analyzing the properties of physical surfaces, such as gradient, aspect and visibility, and analyzing surface-like data "fields"; network analysis — examining the properties of natural and man-made networks in order to understand the behavior of flows within and around such networks; and locational analysis. GIS-based network analysis may be used to address a wide range of practical problems such as route selection and facility location, and problems involving flows such as those found in hydrology. In many instances location problems relate to networks and as such are often best addressed with tools designed for this purpose, but in others existing networks may have little or no relevance or may be impractical to incorporate within the modeling process. Problems that are not specifically network constrained, such as new road or pipeline routing, regional warehouse location, mobile phone mast positioning, pedestrian movement or the selection of rural community health care sites, may be effectively analyzed (at least initially) without reference to existing physical networks. Locational analysis "in the plane" is also applicable where suitable network datasets are not available, or are too large or expensive to be utilized, or where the location algorithm is very complex or involves the examination or simulation of a very large number of alternative configurations.

A further important aspect of geospatial analysis is visualization (or *geovisualization*) — the use, creation and manipulation of images, maps, diagrams, charts, 3D static and dynamic views, high resolution satellite imagery and digital globes, and their associated tabular datasets (see further Slocum *et al.*, 2008, Dodge *et al.*, 2008, Longley *et al.*, 2015, Ch12 and the work of the GeoVista project team). For early insights into how some of these developments may be applied, see Andrew Hudson-Smith (2008) "Digital Geography: Geographic visualization for urban environments" and Martin Dodge and Rob Kitchin's earlier "Atlas of Cyberspace" which is available as a free downloadable document. A more recent Working Paper on this topic from the Centre for Advanced Spatial Analysis (CASA) can be found here:

discovery.ucl.ac.uk/1436767/2/CASAWorkingPaper_paper190.pdf

GIS packages and web-based services increasingly incorporate a range of such tools, providing static or rotating views, draping images over 2.5D surface representations, providing animations and fly-throughs, dynamic linking and brushing and spatio-temporal visualizations. This latter class of tools has been, until recently, the least developed, reflecting in part the limited range of suitable compatible datasets and the limited set of analytical methods available, although this picture is changing rapidly. One recent example is the availability of image time series from NASA's Earth Observation Satellites, yielding vast quantities of data on a daily basis (e.g. Aqua mission, commenced 2002; Terra mission, commenced 1999).

Geovisualization is the subject of ongoing research by the International Cartographic Association (ICA), Commission on Visual Analytics, who have organized a series of workshops and publications addressing developments in geovisualization, notably with a cartographic focus.

As datasets, software tools and processing capabilities develop, 3D geometric and photo-realistic visualization are becoming a *sine qua non* of modern geospatial systems and services — see Andy Hudson-Smith's "Digital Urban" blog for a regularly updated commentary on this field. We expect to see an explosion of tools and services and datasets in this area over the coming years — many examples are included as illustrations in this Guide. Other examples readers may wish to explore include: the static and dynamic visualizations at 3DNature and similar sites; the 2D and 3D Atlas of Switzerland; Urban 3D modeling programmes such as CityGML; and the integration of GIS technologies and data with digital globe software, e.g. data from Digital Globe and GeoEye/Satellite Imaging, and Earth-based frameworks such as Google Earth, Microsoft Virtual Earth, NASA Worldwind and Edushi (Chinese). There are also automated translators between GIS packages such as ArcGIS and digital Earth models (see for example Arc2Earth).

These novel visualization tools and facilities augment the core tools utilized in spatial analysis throughout many parts of the analytical process: exploration of data; identification of patterns and relationships; construction of models; dynamic interaction with models; and communication of results — see, for example, the work of the city of Portland, Oregon, who have used 3D visualization to communicate the results of zoning, crime analysis and other key local variables to the public. Another example is the 3D visualizations provided as part of the web-accessible London Air Quality network. These are designed to enable:

- users to visualize air pollution in the areas that they work, live or walk
- transport planners to identify the most polluted parts of London
- urban planners to see how building density affects pollution concentrations in the City and other high density areas, and
- students to understand pollution sources and dispersion characteristics

Physical 3D models and hybrid physical-digital models have also been developed and applied to practical analysis problems. For example: 3D physical models constructed from plaster, wood, paper and plastics have been used for many years in architectural and engineering planning projects; hybrid sand tables are used to help firefighters in California visualize the progress of wildfires (see Figure 1-1A, below); very large sculptured solid terrain models (e.g. see STM) are being used for educational purposes, to assist land use modeling programmes, and to facilitate participatory 3D modeling in less-developed communities (P3DM); and 3D digital printing technology is being used to generate 3D landscapes and cityscapes from GIS, CAD and/or VRML files with planning, security, architectural, archaeological and geological applications (see Figure 1-1B, below and the websites of Z corporation and Stratasys for more details). To create large landscape models multiple individual prints, which are typically only around 20cm x 20cm x 5cm, are made, in much the same manner as raster file mosaics.

Figure 1-1A: 3D Physical GIS models: Sand-in-a-box model, Albuquerque, USA

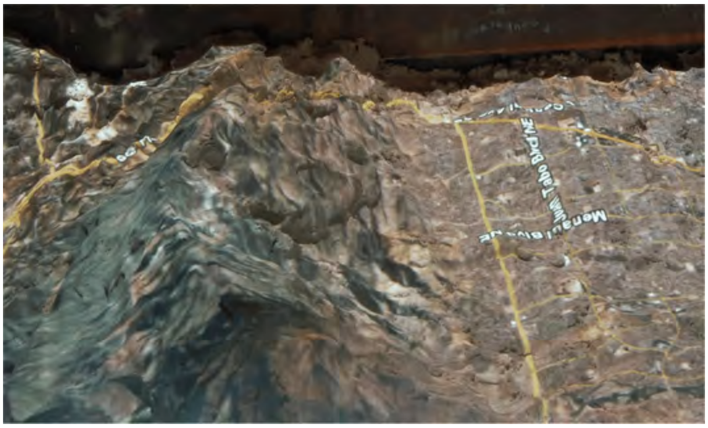


Figure 1-1B: 3D Physical GIS models: 3D GIS printing

GIS software, notably in the commercial sphere, is driven primarily by demand and applicability, as manifest in willingness to pay. Hence, to an extent, the facilities available often reflect commercial and resourcing realities (including the development of improvements in processing and display hardware, and the ready availability of high quality datasets) rather than the status of development in geospatial science. Indeed, there may be many capabilities available in software packages that are provided simply because it is extremely easy for the designers and programmers to implement them, especially those employing object-oriented programming and data models. For example, a given operation may be provided for polygonal features in response to a well-understood application requirement, which is then easily enabled for other features (e.g. point sets, polylines) despite the fact that there may be no known or likely requirement for the facility.

Despite this cautionary note, for specific well-defined or *core* problems, software developers will frequently utilize the most up-to-date research on algorithms in order to improve the quality (accuracy, optimality) and efficiency (speed, memory usage) of their products. For further information on algorithms and data structures, see the online NIST Dictionary of algorithms and data structures.

Furthermore, the quality, variety and efficiency of spatial analysis facilities provide an important discriminator between commercial offerings in an increasingly competitive and open market for software. However, the ready availability of analysis tools does not imply that one product is necessarily better or more complete than another — it is the selection and application of *appropriate* tools in a manner that is *fit for purpose* that is important. Guidance documents exist in some disciplines that assist users in this process, e.g. Perry *et al*. (2002) dealing with ecological data analysis, and to a significant degree we hope that this Guide will assist users from many disciplines in the selection process.

1.2 Intended audience and scope

This Guide has been designed to be accessible to a wide range of readers — from undergraduates and postgraduates studying GIS and spatial analysis, to GIS practitioners and professional analysts. It is intended to be much more than a cookbook of formulas, algorithms and techniques — its aim is to provide an explanation of the key techniques of spatial analysis using examples from widely available software packages. It stops short, however, of attempting a systematic evaluation of competing software products. A substantial range of application examples are provided, but any specific selection inevitably illustrates only a small subset of the huge range of facilities available. Wherever possible, examples have been drawn from non-academic sources, highlighting the growing understanding and acceptance of GIS technology in the commercial and government sectors.

The scope of this Guide incorporates the various spatial analysis topics included within the seminal NCGIA Core Curriculum (Goodchild and Kemp, 1990) and as such may provide a useful accompaniment to GIS Analysis courses based closely or loosely on this programme. More recently the Education Committee of the University Consortium for Geographic Information Science (UCGIS) in conjunction with the Association of American Geographers (AAG) has produced a comprehensive "Body of Knowledge" (BoK) document, which is available from the AAG bookstore (www.aag.org/bok) and is in the process of being updated to a second edition. This Guide covers materials that primarily relate to the BoK sections **CF**: Conceptual Foundations; **AM**: Analytical Methods and **GC**: Geocomputation. In the general introduction to the **AM** knowledge area the authors of the BoK summarize this component as follows:

> *"This knowledge area encompasses a wide variety of operations whose objective is to derive analytical results from geospatial data. Data analysis seeks to understand both first-order (environmental) effects and second-order (interaction) effects. Approaches that are both data-driven (exploration of geospatial data) and model-driven (testing hypotheses and creating models) are included. Data-driven techniques derive summary descriptions of data, evoke insights about characteristics of data, contribute to the development of research hypotheses, and lead to the derivation of analytical results. The goal of model-driven analysis is to create and test geospatial process models. In general, model-driven analysis is an advanced knowledge area where previous experience with exploratory spatial data analysis would constitute a desired prerequisite." (BoK, p83 of the e-book version, first edition).*

1.3 Software tools and Companion Materials

In this section you will find the following topics:

GIS and related software tools

Suggested reading

1.3.1 GIS and related software tools

The GIS software and analysis tools that an individual, group or corporate body chooses to use will depend very much on the purposes to which they will be put. There is an enormous difference between the requirements of academic researchers and educators, and those with responsibility for planning and delivery of emergency control systems or large scale physical infrastructure projects. The spectrum of products that may be described as a GIS includes (amongst others):

- highly specialized, sector specific packages: for example civil engineering design and costing systems; satellite image processing systems; and utility infrastructure management systems
- transportation and logistics management systems
- civil and military control room systems
- systems for visualizing the built environment for architectural purposes, for public consultation or as part of simulated environments for interactive gaming
- land registration systems
- census data management systems
- commercial location services and Digital Earth models
- Geospatial data visualization tools

The list of software functions and applications is long and in some instances suppliers would not describe their offerings as a GIS. In many cases such systems fulfill specific operational needs, solving a well-defined subset of spatial problems and providing mapped output as an incidental but essential part of their operation. Many of the capabilities may be found in generic GIS products. In other instances a specialized package may utilize a GIS engine for the display and in some cases processing of spatial data (directly, or indirectly through interfacing or file input/output mechanisms). For this reason, and in order to draw a boundary around the present work, reference to application-specific GIS will be limited.

A number of GIS packages and related toolsets have particularly strong facilities for processing and analyzing binary, grayscale and color images. They may have been designed originally for the processing of remote sensed data from satellite and aerial surveys, but many have developed into much more sophisticated and complete GIS tools, e.g. Clark Lab's Idrisi software; MicroImage's TNTMips product set; the ERDAS suite of products; and ENVI with associated packages such as RiverTools. Alternatively, image handling may have been deliberately included within the original design parameters for a generic GIS package (e.g. Manifold), or simply be toolsets for image processing that may be combined with mapping tools (e.g. the MATLab Image Processing Toolbox). Whatever their origins, a central purpose of such tools has been the capture, manipulation and interpretation of image data, rather than spatial analysis *per se*, although the latter inevitably follows from the former.

In this Guide we do not provide a separate chapter on image processing, despite its considerable importance in GIS, focusing instead on those areas where image processing tools and concepts are applied for spatial analysis (e.g. surface analysis). We have adopted a similar position with respect to other forms of data capture, such as field and geodetic survey systems and data cleansing software — although these incorporate analytical tools, their primary function remains the recording and georeferencing of datasets, rather than the analysis of such datasets once stored.

For most GIS professionals, spatial analysis and associated modeling is an infrequent activity. Even for those whose job focuses on analysis the range of techniques employed tends to be quite narrow and application specific. GIS consultants, researchers and academics on the other hand are continually exploring and developing analytical techniques. For the first group and for consultants, especially in commercial environments, the imperatives of financial considerations, timeliness and corporate policy loom large, directing attention to: delivery of solutions within well-defined time and cost parameters; working within commercial constraints on the cost and availability of software, datasets and staffing; ensuring that solutions are fit for purpose/meet client and end-user expectations and agreed standards; and in some cases, meeting "political" expectations.

For the second group of users it is common to make use of a variety of tools, data and programming facilities developed in the academic sphere. Increasingly these make use of non-commercial wide-ranging spatial analysis software libraries, such as the R-Spatial project (in "R"); PySal (in "Python"); and Splancs (in "S").

Sample software products

The principal products we have included in this latest edition of the Guide are included on the accompanying website's software page. Many of these products are free whilst others are available (at least in some form) for a small fee for all or selected groups of users. Others are licensed at varying per user prices, from a few hundred to over a thousand US dollars per user. Our tests and examples have largely been carried out using desktop/Windows versions of these software products. Different versions that support Unix-based operating systems and more sophisticated back-end database engines have not been utilized. In the context of this Guide we do not believe these selections affect our discussions in any substantial manner, although such issues may have performance and systems architecture implications that are extremely important for many users. OGC compliant software products are listed on the OGC resources web page:

www.opengeospatial.org/resource/products/compliant

To quote from the OGC: "*The OGC Compliance Testing Program provides a formal process for testing compliance of products that implement OpenGIS® Standards. Compliance Testing determines that a specific product implementation of a particular OpenGIS® Standard complies with all mandatory elements as specified in the standard and that these elements operate as described in the standard.*"

Software performance

Suppliers should be able to provide advice on performance issues (e.g. see the ESRI web site, "Services" area for relevant documents relating to their products) and in some cases such information is provided within product Help files (e.g. see the Performance Tips section within the Manifold GIS help file). Some analytical tasks are very processor- and memory-hungry, particularly as the number of elements involved increases. For example, vector overlay and buffering is relatively fast with a few objects and layers, but slows appreciably as the number of elements involved increases. This increase is generally at least linear with the number of layers and features, but for some problems grows in a highly non-linear (i.e. geometric) manner. Many optimization tasks, such as optimal routing through networks or trip distribution modeling, are known to be extremely hard or impossible to solve optimally and methods to achieve a best solution with a large dataset can take a considerable time to run (see our discussion of Algorithms and computational complexity theory later in this Guide for a fuller discussion of this topic). Similar problems exist with the processing and display of raster files, especially large images or sets of images. Geocomputational methods, some of which are beginning to appear within GIS packages and related toolsets, are almost by definition computationally intensive. This certainly applies to large-scale (Monte Carlo) simulation models, cellular automata and agent-based models and some raster-based optimization techniques, especially where modeling extends into the time domain.

A frequent criticism of GIS software is that it is over-complicated, resource-hungry and requires specialist expertise to understand and use. Such criticisms are often valid and for many problems it may prove simpler, faster and more transparent to utilize specialized tools for the analytical work and draw on the strengths of GIS in data management and mapping to provide input/output and visualization functionality. Example approaches include: (i) using high-level programming facilities within a GIS (e.g. macros, scripts, VBA, Python) – many add-ins are developed in this way; (ii) using wide-ranging programmable spatial analysis software libraries and toolsets that incorporate GIS file reading, writing and display, such as the R-Spatial and PySal projects noted earlier; (iii) using general purpose data processing toolsets, e.g. MATLab, Excel, Python's Matplotlib, Numeric Python (Numpy) and other libraries from Enthought; or (iv) directly utilizing mainstream programming languages (e.g. Java, C++). The advantage of these approaches is control and transparency, the disadvantages are that software development is never trivial, is often subject to frustrating and unforeseen delays and errors, and generally requires ongoing maintenance. In some instances analytical applications may be well-suited to parallel or grid-enabled processing – as for example is the case with GWR (see Harris *et al.*, 2006).

At present there are no standardized tests for the quality, speed and accuracy of GIS procedures. It remains the buyer's and user's responsibility and duty to evaluate the software they wish to use for the specific task at hand, and by systematic controlled tests or by other means establish that the product and facility within that product they choose to use is truly fit for purpose — *caveat emptor*! Details of how to obtain these products are provided on the software page of the website that accompanies this book. The list maintained on Wikipedia is also a useful source of information and links, although is far from being complete or independent. A number of trade magazines and websites (such as Geoplace and Geocommunity) provide ad hoc reviews of GIS software offerings, especially new releases, although coverage of analytical functionality may be limited.

1.3.2 Suggested reading

There are numerous excellent modern books on GIS and spatial analysis, although few address software facilities and developments. Hypertext links are provided here, and throughout the text where they are cited, to the more recent publications and web resources listed.

As a background to this Guide any readers unfamiliar with GIS are encouraged to first tackle "Geographic Information Science and Systems" (GISSc) by Longley *et al.* (2015). GISSc seeks to provide a comprehensive and highly accessible introduction to the subject as a whole. The GB Ordnance Survey's "Understanding GIS" also provides an excellent brief introduction to GIS and its applications.

Some of the basic mathematics and statistics of relevance to GIS analysis is covered in Dale (2005) and Allan (2004). For detailed information on datums and map projections, see Iliffe and Lott (2008). Useful online resources for those involved in data analysis, particularly with a statistical content, include the StatsRef website and the e-Handbook of Statistical Methods produced by the US National Institute on Standards and Technology, NIST). The more informally produced set of articles on statistical topics provided under the Wikipedia umbrella are also an extremely useful resource. These sites, and the mathematics reference site, Mathworld, are referred to (with hypertext links) at various points throughout this document. For more specific sources on geostatistics and associated software packages, the European Commission's AI-GEOSTATS website (now a separate WIKI website) is highly recommended, as is the web site of the Center for Computational Geostatistics (CCG) at the University of Alberta. For those who find mathematics and statistics something of a mystery, de Smith (2018) and Bluman (2003) provide useful starting points. For guidance on how to avoid the many pitfalls of statistical data analysis readers are recommended the material in the classic work by Huff (1993) "How to lie with statistics", and the 2008 book by Blastland and Dilnot "The tiger that isn't".

A relatively new development has been the increasing availability of out-of-print published books, articles and guides as free downloads in PDF format. These include: the series of 59 short guides published under the CATMOG umbrella (Concepts and Methods in Modern Geography), published between 1975 and 1995, most of which are now available at the QMRG website (a full list of all the guides is provided at the end of this book); the Atlas of Cyberspace by Dodge and Kitchin; and Fractal Cities, by Batty and Longley.

Undergraduates and MSc programme students will find Burrough and McDonnell (1998, 2015) provides excellent coverage of many aspects of geospatial analysis, especially from an environmental sciences perspective. Valuable guidance on the relationship between spatial process and spatial modeling may be found in Cliff and Ord (1981) and Bailey and Gatrell (1995). The latter provides an excellent introduction to the application of statistical methods to spatial data analysis. O'Sullivan and Unwin (2010, 2nd ed.) is a more broad-ranging book covering the topic the authors describe as "Geographic Information Analysis". This work is best suited to advanced undergraduates and first year postgraduate students. In many respects a deeper and more challenging work is Haining's (2003) "Spatial Data Analysis — Theory and Practice". This book is strongly recommended as a companion to the present Guide for postgraduate researchers and professional analysts involved in using GIS in conjunction with statistical analysis.

However, these authors do not address the broader spectrum of geospatial analysis and associated modeling as we have defined it. For example, problems relating to networks and location are often not covered and the literature relating to this area is scattered across many disciplines, being founded upon the mathematics of graph theory, with applications ranging from electronic circuit design to computer networking and from transport planning to the design of complex molecular structures. Useful books addressing this field include Miller and Shaw (2001) "Geographic Information Systems for Transportation" (especially Chapters 3, 5 and 6), and Rodrigue *et al.* (2006) "The geography of transport systems" (see further: people.hofstra.edu/geotrans/).

As companion reading on these topics for the present Guide we suggest the two volumes from the Handbooks in Operations Research and Management Science series by Ball *et al.* (1995): "Network Models", and "Network Routing". These rather expensive volumes provide collections of reviews covering many classes of network problems, from the core optimization problems of shortest paths and arc routing (e.g. street cleaning), to the complex problems of dynamic routing in variable networks, and a great deal more besides. This is challenging material and many readers may prefer to seek out more approachable material, available in a number of other books and articles, e.g. Ahuja *et al.* (1993), Mark Daskin's excellent book "Network and Discrete Location" (1995) and the earlier seminal works by Haggett and Chorley (1969), and Scott (1971), together with the widely available online materials accessible via the Internet. Final recommendations here are Stephen Wise's excellent GIS Basics (2002) and Worboys and Duckham (2004) which address GIS from a computing perspective. Both these volumes covers many topics, including the central issues of data modeling and data structures, key algorithms, system architectures and interfaces.

Many recent books described as covering (geo)spatial analysis are essentially edited collections of papers or brief articles. As such most do not seek to provide comprehensive coverage of the field, but tend to cover information on recent developments, often with a specific application focus (e.g. health, transport, archaeology). The latter is particularly common where these works are selections from sector- or discipline-specific conference proceedings, whilst in other cases they are carefully chosen or specially written papers. Classic amongst these is Berry and Marble (1968) "Spatial Analysis: A reader in statistical geography". More recent examples include "GIS, Spatial Analysis and Modeling" edited by Maguire, Batty and Goodchild (2005), and the excellent (but costly) compendium work "The SAGE handbook of Spatial Analysis" edited by Fotheringham and Rogerson (2008).

A second category of companion materials to the present work is the extensive product-specific documentation available from software suppliers. Some of the online help files and product manuals are excellent, as are associated example data files, tutorials, worked examples and white papers (see for example, ESRI's What is GIS?), which provides a wide-ranging guide to GIS. In many instances we utilize these to illustrate the capabilities of specific pieces of software and to enable readers to replicate our results using readily available materials. In addition some suppliers, notably ESRI, have a substantial publishing operation, including more general (i.e. not product specific) books of relevance to the present work. Amongst their publications we strongly recommend the "ESRI Guide to GIS Analysis Volume 1: Geographic patterns and relationships" (1999) by Andy Mitchell, which is full of valuable tips and examples. This is a basic introduction to GIS Analysis, which he defines in this context as "a process for looking at geographic patterns and relationships between features". Mitchell's Volume 2 (July 2005) covers more advanced techniques of data analysis, notably some of the more accessible and widely supported methods of spatial statistics, and is equally highly recommended. A number of the topics covered in his Volume 2 also appear in this Guide. David Allen has produced a tutorial book and DVD (GIS Tutorial II: Spatial Analysis Workbook) to go alongside Mitchell's volumes, and these are obtainable from ESRI Press. Those considering using Open Source software should investigate the books by Neteler and Mitasova (2008), Tyler Mitchell (2005) and Sherman (2008).

In parallel with the increasing range and sophistication of spatial analysis facilities to be found within GIS packages, there has been a major change in spatial analytical techniques. In large measure this has come about as a result of technological developments and the related availability of software tools and detailed publicly available datasets. One aspect of this has been noted already — the move towards network-based location modeling where in the past this would have been unfeasible. More general shifts can be seen in the move towards local rather than simply global analysis, for example in the field of exploratory data analysis; in the increasing use of advanced forms of visualization as an aid to analysis and communication; and in the development of a wide range of computationally intensive and simulation methods that address problems through micro-scale processes (geocomputational methods). These trends are addressed at many points throughout this Guide.

1.4 Terminology and Abbreviations

GIS, like all disciplines, utilizes a wide range of terms and abbreviations, many of which have well-understood and recognized meanings. For a large number of commonly used terms online dictionaries have been developed, for example: those created by the Association for Geographic Information (AGI); the Open Geospatial Consortium (OGC); and by various software suppliers. The latter includes many terms and definitions that are particular to specific products, but remain a valuable resource. Web site details for each of these are provided at the end of this Guide.

1.4.1 Definitions

Geospatial analysis utilizes many of these terms, but many others are drawn from disciplines such as mathematics and statistics. The result that the same terms may mean entirely different things depending on their context and in many cases, on the software provider utilizing them. In most instances terms used in this Guide are defined on the first occasion they are used, but a number warrant defining at this stage. Table 1-1, below, provides a selection of such terms, utilizing definitions from widely recognized sources where available and appropriate.

Table 1-1 Selected terminology

Term	Definition
Adjacency	The sharing of a common side or boundary by two or more polygons (AGI). Note that adjacency may also apply to features that lie either side of a common boundary where these features are not necessarily polygons
Arc	Commonly used to refer to a straight line segment connecting two nodes or vertices of a polyline or polygon. Arcs may include segments or circles, spline functions or other forms of smooth curve. In connection with graphs and networks, arcs may be directed or undirected, and may have other attributes (e.g. cost, capacity etc.)
Artifact	A result (observation or set of observations) that appears to show something unusual (e.g. a spike in the surface of a 3D plot) but which is of no significance. Artifacts may be generated by the way in which data have been collected, defined or re-computed (e.g. resolution changing), or as a result of a computational operation (e.g. rounding error or substantive software error). Linear artifacts are sometimes referred to as “ghost lines”
Aspect	The direction in which slope is maximized for a selected point on a surface (see also, Gradient and Slope)
Attribute	A data item associated with an individual object (record) in a spatial database. Attributes may be explicit, in which case they are typically stored as one or more fields in tables linked to a set of objects, or they may be implicit (sometimes referred to as *intrinsic*), being either stored but hidden or computed as and when required (e.g. polyline length, polygon centroid). Raster/grid datasets typically have a single explicit attribute (a value) associated with each cell, rather than an attribute table containing as many records as there are cells in the grid
Azimuth	The horizontal direction of a vector, measured clockwise in degrees of rotation from the positive Y-axis, for example, degrees on a compass (AGI)
Azimuthal Projection	A type of map projection constructed as if a plane were to be placed at a tangent to the Earth's surface and the area to be mapped were projected onto the plane. All points on this projection keep their true compass bearing (AGI)
(Spatial) Autocorrelation	The degree of relationship that exists between two or more (spatial) variables, such that when one changes, the other(s) also change. This change can either be in the same direction, which is a positive autocorrelation, or in the opposite direction, which is a negative autocorrelation (AGI). The term autocorrelation is usually applied to ordered datasets, such as those relating to time series or spatial data ordered by distance band. The existence of such a relationship suggests but does not definitely establish causality
Cartogram	A cartogram is a form of map in which some variable such as Population Size or Gross National Product typically is substituted for land area. The geometry or space of the map is distorted in order to convey the information of this alternate variable. Cartograms use a

Term	Definition
	variety of approaches to map distortion, including the use of continuous and discrete regions. The term cartogram (or *linear cartogram*) is also used on occasion to refer to maps that distort distance for particular display purposes, such as the London Underground map
Choropleth	A thematic map [i.e. a map showing a theme, such as soil types or rainfall levels] portraying properties of a surface using area symbols such as shading [or color]. Area symbols on a choropleth map usually represent categorized classes of the mapped phenomenon (AGI)
Conflation	A term used to describe the process of combining (merging) information from two data sources into a single source, reconciling disparities where possible (e.g. by rubber-sheeting — see below). The term is distinct from *concatenation* which refers to combinations of data sources (e.g. by overlaying one upon another) but retaining access to their distinct components
Contiguity	The topological identification of adjacent polygons by recording the left and right polygons of each arc. Contiguity is not concerned with the exact locations of polygons, only their relative positions. Contiguity data can be stored in a table, matrix or simply as [i.e. in] a list, that can be cross-referenced to the relevant co-ordinate data if required (AGI).
Curve	A one-dimensional geometric object stored as a sequence of points, with the subtype of curve specifying the form of interpolation between points. A curve is simple if it does not pass through the same point twice (OGC). A LineString (or polyline — see below) is a subtype of a curve
Datum	Strictly speaking, the singular of *data*. In GIS the word datum usually relates to a reference level (surface) applying on a nationally or internationally defined basis from which elevation is to be calculated. In the context of terrestrial geodesy datum is usually defined by a model of the Earth or section of the Earth, such as WGS84 (see below). The term is also used for horizontal referencing of measurements; see Iliffe and Lott (2008) for full details
DEM	Digital elevation model (a DEM is a particular kind of DTM, see below)
DTM	Digital terrain model
EDM	Electronic distance measurement
EDA, ESDA	Exploratory data analysis/Exploratory spatial data analysis
Ellipsoid/Spheroid	An ellipse rotated about its minor axis determines a spheroid (sphere-like object), also known as an ellipsoid of revolution (see also, WGS84)

Term	Definition
Feature	Frequently used within GIS referring to point, line (including polyline and mathematical functions defining arcs), polygon and sometimes text (annotation) objects (see also, vector)
Geoid	An imaginary shape for the Earth defined by mean sea level and its imagined continuation under the continents at the same level of gravitational potential (AGI)
Geodemographics	The analysis of people by where they live, in particular by type of neighborhood. Such localized classifications have been shown to be powerful discriminators of consumer behavior and related social and behavioral patterns
Geospatial	Referring to location relative to the Earth's surface. "Geospatial" is more precise in many GI contexts than "geographic," because geospatial information is often used in ways that do not involve a graphic representation, or map, of the information. OGC
Geostatistics	Statistical methods developed for and applied to geographic data. These statistical methods are required because geographic data do not usually conform to the requirements of standard statistical procedures, due to spatial autocorrelation and other problems associated with spatial data (AGI). The term is widely used to refer to a family of tools used in connection with spatial interpolation (prediction) of (piecewise) continuous datasets and is widely applied in the environmental sciences. Spatial statistics is a term more commonly applied to the analysis of discrete objects (e.g. points, areas) and is particularly associated with the social and health sciences
Geovisualization	A family of techniques that provide visualizations of spatial and spatio-temporal datasets, extending from static, 2D maps and cartograms, to representations of 3D using perspective and shading, solid terrain modeling and increasingly extending into dynamic visualization interfaces such as linked windows, digital globes, fly-throughs, animations, virtual reality and immersive systems. Geovisualization is the subject of ongoing research by the International Cartographic Association (ICA), Commission on Geovisualization
GIS-T	GIS applied to transportation problems
GPS/ DGPS	Global positioning system; Differential global positioning system — DGPS provides improved accuracy over standard GPS by the use of one or more fixed reference stations that provide corrections to GPS data
Gradient	Used in spatial analysis with reference to surfaces (scalar fields). Gradient is a vector field comprised of the *aspect* (direction of maximum slope) and *slope* computed in this direction (magnitude of rise over run) at each point of the surface. The magnitude of the gradient (the slope or inclination) is sometimes itself referred to as the gradient (see also, Slope and Aspect)

Term	Definition
Graph	A collection of vertices and edges (links between vertices) constitutes a graph. The mathematical study of the properties of graphs and paths through graphs is known as graph theory
Heuristic	A term derived from the same Greek root as Eureka, heuristic refers to procedures for finding solutions to problems that may be difficult or impossible to solve by direct means. In the context of optimization heuristic algorithms are systematic procedures that seek a good or near optimal solution to a well-defined problem, but not one that is necessarily optimal. They are often based on some form of intelligent trial and error or search procedure
iid	An abbreviation for "independently and identically distributed". Used in statistical analysis in connection with the distribution of errors or residuals
Invariance	In the context of GIS invariance refers to properties of features that remain unchanged under one or more (spatial) transformations
Kernel	Literally, the core or central part of an item. Often used in computer science to refer to the central part of an operating system, the term kernel in geospatial analysis refers to methods (e.g. density modeling, local grid analysis) that involve calculations using a well-defined local neighborhood (block of cells, radially symmetric function)
Layer	A collection of geographic entities of the same type (e.g. points, lines or polygons). Grouped layers may combine layers of different geometric types
Map algebra	A range of actions applied to the grid cells of one or more maps (or images) often involving filtering and/or algebraic operations. These techniques involve processing one or more raster layers according to simple rules resulting in a new map layer, for example replacing each cell value with some combination of its neighbors' values, or computing the sum or difference of specific attribute values for each grid cell in two matching raster datasets
Mashup	A recently coined term used to describe websites whose content is composed from multiple (often distinct) data sources, such as a mapping service and property price information, constructed using programmable interfaces to these sources (as opposed to simple compositing or embedding)
MBR/ MER	Minimum bounding rectangle/Minimum enclosing (or envelope) rectangle (of a feature set)
Planar/non-planar/planar enforced	Literally, lying entirely within a plane surface. A polygon set is said to be planar enforced if every point in the set lies in exactly one polygon, or on the boundary between two or more polygons. See also, planar graph. A graph or network with edges crossing (e.g. bridges/underpasses) is non-planar

Term	Definition
Planar graph	If a graph can be drawn in the plane (embedded) in such a way as to ensure edges only intersect at points that are vertices then the graph is described as planar
Pixel/image	Picture element — a single defined point of an image. Pixels have a “color” attribute whose value will depend on the encoding method used. They are typically either binary (0/1 values), grayscale (effectively a color mapping with values, typically in the integer range [0,255]), or color with values from 0 upwards depending on the number of colors supported. Image files can be regarded as a particular form of raster or grid file
Polygon	A closed figure in the plane, typically comprised of an ordered set of connected vertices, $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots\boldsymbol{v}_{n-1},\boldsymbol{v}_n=\boldsymbol{v}_1$ where the connections (edges) are provided by straight line segments. If the sequence of edges is not self-crossing it is called a simple polygon. A point is inside a simple polygon if traversing the boundary in a clockwise direction the point is always on the right of the observer. If every pair of points inside a polygon can be joined by a straight line that also lies inside the polygon then the polygon is described as being convex (i.e. the interior is a connected point set). The OGC definition of a polygon is “a planar surface defined by 1 exterior boundary and 0 or more interior boundaries. Each interior boundary defines a hole in the polygon”
Polyhedral surface	A Polyhedral surface is a contiguous collection of polygons, which share common boundary segments (OGC). See also, Tesseral/Tessellation
Polyline	An ordered set of connected vertices, $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots\boldsymbol{v}_{n-1},\boldsymbol{v}_n\neq\boldsymbol{v}_1$ where the connections (edges) are provided by straight line segments. The vertex $\boldsymbol{v}_1$ is referred to as the start of the polyline and $\boldsymbol{v}_n$ as the end of the polyline. The OGC specification uses the term LineString which it defines as: "a curve with linear interpolation between points. Each consecutive pair of points defines a line segment"
Raster/grid	A data model in which geographic features are represented using discrete cells, generally squares, arranged as a (contiguous) rectangular grid. A single grid is essentially the same as a two-dimensional matrix, but is typically referenced from the lower left corner rather than the norm for matrices, which are referenced from the upper left. Raster files may have one or more values (attributes or bands) associated with each cell position or pixel
Resampling	1. Procedures for (automatically) adjusting one or more raster datasets to ensure that the grid resolutions of all sets match when carrying out combination operations. Resampling is often performed to match the coarsest resolution of a set of input rasters. Increasing resolution rather than decreasing requires an interpolation procedure such as bicubic spline. 2. The process of reducing image dataset size by representing a group of pixels with a single pixel. Thus, pixel count is lowered, individual pixel size is increased, and overall image geographic extent is retained. Resampled images are “coarse” and have less information than the images from which they are taken. Conversely, this process can also be executed in the reverse (AGI)

Term	Definition
	3. In a statistical context the term resampling (or re-sampling) is sometimes used to describe the process of selecting a subset of the original data, such that the samples can reasonably be expected to be independent
Rubber sheeting	A procedure to adjust the co-ordinates all of the data points in a dataset to allow a more accurate match between known locations and a few data points within the dataset. Rubber sheeting … preserves the interconnectivity or topology, between points and objects through stretching, shrinking or re-orienting their interconnecting lines (AGI). Rubber-sheeting techniques are widely used in the production of Cartograms (*op. cit.*)
Slope	The amount of *rise* of a surface (change in elevation) divided by the distance over which this rise is computed (the *run*), along a straight line transect in a specified direction. The run is usually defined as the *planar* distance, in which case the slope is the tan() function. Unless the surface is flat the slope at a given point on a surface will (typically) have a maximum value in a particular direction (depending on the surface and the way in which the calculations are carried out). This direction is known as the *aspect*. The *vector* consisting of the slope and aspect is the *gradient* of the surface at that point (see also, Gradient and Aspect)
Spatial econometrics	A subset of econometric methods that is concerned with spatial aspects present in cross-sectional and space-time observations. These methods focus in particular on two forms of so-called spatial effects in econometric models, referred to as spatial dependence and spatial heterogeneity (Anselin, 1988, 2006)
Spheroid	A flattened (oblate) form of a sphere, or ellipse of revolution. The most widely used model of the Earth is that of a spheroid, although the detailed form is slightly different from a true spheroid
SQL/Structured Query Language	Within GIS software SQL extensions known as spatial queries are frequently implemented. These support queries that are based on spatial relationships rather than simply attribute values
Surface	A 2D geometric object. A simple surface consists of a single 'patch' that is associated with one exterior boundary and 0 or more interior boundaries. Simple surfaces in 3D are isomorphic to planar surfaces. Polyhedral surfaces are formed by 'stitching' together simple surfaces along their boundaries (OGC). Surfaces may be regarded as scalar fields, i.e. fields with a single value, e.g. elevation or temperature, at every point
Tesseral/Tessellation	A gridded representation of a plane surface into disjoint polygons. These polygons are normally either square (raster), triangular (TIN — see below), or hexagonal. These models can be built into hierarchical structures, and have a range of algorithms available to navigate through them. A (regular or irregular) 2D tessellation involves the subdivision of a 2-dimensional plane into polygonal tiles (polyhedral blocks) that completely cover a plane (AGI). The term lattice is sometimes used to describe the complete division of the plane

Term	Definition
	into regular or irregular disjoint polygons. More generally the subdivision of the plane may be achieved using arcs that are not necessarily straight lines
TIN	Triangulated irregular network. A form of the tesseral model based on triangles. The vertices of the triangles form irregularly spaced nodes. Unlike the grid, the TIN allows dense information in complex areas, and sparse information in simpler or more homogeneous areas. The TIN dataset includes topological relationships between points and their neighboring triangles. Each sample point has an *X,Y* co-ordinate and a surface, or *Z*-Value. These points are connected by edges to form a set of non-overlapping triangles used to represent the surface. TINs are also called irregular triangular mesh or irregular triangular surface models (AGI)
Topology	The relative location of geographic phenomena independent of their exact position. In digital data, topological relationships such as connectivity, adjacency and relative position are usually expressed as relationships between nodes, links and polygons. For example, the topology of a line includes its from- and to-nodes, and its left and right polygons (AGI). In mathematics, a property is said to be *topological* if it survives stretching and distorting of space
Transformation 1. Map	Map transformation: A computational process of converting an image or map from one coordinate system to another. Transformation … typically involves rotation and scaling of grid cells, and thus requires resampling of values (AGI)
Transformation 2. Affine	Affine transformation: When a map is digitized, the *X* and *Y* coordinates are initially held in digitizer measurements. To make these *X,Y* pairs useful they must be converted to a real world coordinate system. The affine transformation is a combination of linear transformations that converts digitizer coordinates into Cartesian coordinates. The basic property of an affine transformation is that parallel lines remain parallel (AGI, with modifications). The principal affine transformations are contraction, expansion, dilation, reflection, rotation, shear and translation
Transformation 3. Data	Data transformation (see also, subsection 6.7.1): A mathematical procedure (usually a one-to-one mapping or function) applied to an initial dataset to produce a result dataset. An example might be the transformation of a set of sampled values $\{x_i\}$ using the log() function, to create the set $\{\log(x_i)\}$. Affine and map transformations are examples of mathematical transformations applied to coordinate datasets. Note that operations on transformed data, e.g. checking whether a value is within 10% of a target value, is not equivalent to the same operation on untransformed data, even after back transformation
Transformation 4. Back	Back transformation: If a set of sampled values $\{x_i\}$ has been transformed by a one-to-one mapping function $f()$ into the set $\{f(x_i)\}$, and $f()$ has a one-to-one inverse mapping function $f^{-1}()$, then the process of computing $f^{-1}\{f(x_i)\}=\{x_i\}$ is known as back transformation. Example $f()=\ln()$ and $f^{-1}=\exp()$

Term	Definition
Vector	1. Within GIS the term vector refers to data that are comprised of lines or arcs, defined by beginning and end points, which meet at nodes. The locations of these nodes and the topological structure are usually stored explicitly. Features are defined by their boundaries only and curved lines are represented as a series of connecting arcs. Vector storage involves the storage of explicit topology, which raises overheads, however it only stores those points which define a feature and all space outside these features is "non-existent" (AGI) 2. In mathematics the term refers to a directed line, i.e. a line with a defined origin, direction and orientation. The same term is used to refer to a single column or row of a matrix, in which case it is denoted by a bold letter, usually in lower case
Viewshed	Regions of visibility observable from one or more observation points. Typically a viewshed will be defined by the numerical or color coding of a raster image, indicating whether the (target) cell can be seen from (or probably seen from) the (source) observation points. By definition a cell that can be viewed from a specific observation point is inter-visible with that point (each location can see the other). Viewsheds are usually determined for optically defined visibility within a maximum range
WGS84	World Geodetic System, 1984 version. This models the Earth as a spheroid with major axis 6378.137 kms and flattening factor of 1:298.257, i.e. roughly 0.3% flatter at the poles than a perfect sphere. One of a number of such global models

Note: Where cited, references are drawn from the Association for Geographic Information (AGI), and the Open Geospatial Consortium (OGC). Square bracketed text denotes insertion by the present authors into these definitions. For OGC definitions see: Open Geospatial Consortium Inc (2006) in References section

1.5 Common Measures and Notation

Throughout this Guide a number of terms and associated formulas are used that are common to many analytical procedures. In this section we provide a brief summary of those that fall into this category. Others, that are more specific to a particular field of analysis, are treated within the section to which they primarily apply. Many of the measures we list will be familiar to readers, since they originate from standard single variable (univariate) statistics. For brevity we provide details of these in tabular form. In order to clarify the expressions used here and elsewhere in the text, we use the notation shown in Table 1-2. *Italics* are used within the text and formulas to denote variables and parameters, as well as selected terms.

1.5.1 Notation

Table 1-2 Notation and symbology

$[a,b]$	A closed interval of the Real line, for example [0,1] means the set of all values between 0 and 1, including 0 and 1
(a,b)	An open interval of the Real line, for example (0,1) means the set of all values between 0 and 1, NOT including 0 and 1. This should not be confused with the notation for coordinate pairs, (x,y), or its use

	within bivariate functions such as f(x,y), or in connection with graph edges (see below) — the meaning should be clear from the context
(i,j)	In the context of graph theory, which forms the basis for network analysis, this pairwise notation is often used to define an edge connecting the two vertices i and j
(x,y)	A (spatial) data pair, usually representing a pair of coordinates in two dimensions. Terrestrial coordinates are typically Cartesian (i.e. in the plane, or *planar*) based on a pre-specified projection of the sphere, or Spherical (latitude, longitude). Spherical coordinates are often quoted in positive or negative degrees from the Equator and the Greenwich meridian, so may have the ranges [-90,+90] for latitude (north-south measurement) and [-180,180] for longitude (east-west measurement)
(x,y,z)	A (spatial) data triple, usually representing a pair of coordinates in two dimensions, plus a third coordinate (usually height or depth) or an attribute value, such as soil type or household income
$\{x_i\}$	A *set* of n values x_1, x_2, x_3, … x_n, typically continuous ratio-scaled variables in the range (-∞,∞) or [0,∞). The values may represent measurements or attributes of distinct objects, or values that represent a collection of objects (for example the population of a census tract)
$\{X_i\}$	An ordered set of n values X_1, X_2, X_3, … X_n, such that $X_i \leq X_{i+1}$ for all i
X,x	The use of bold symbols in expressions indicates matrices (upper case) and vectors (lower case)
$\{f_i\}$	A set of k frequencies (k<=n), derived from a dataset $\{x_i\}$. If $\{x_i\}$ contains discrete values, some of which occur multiple times, then $\{f_i\}$ represents the number of occurrences or the *count* of each distinct value. $\{f_i\}$ may also represent the number of occurrences of values that lie in a range or set of ranges, $\{r_i\}$. If a dataset contains n values, then the sum $\Sigma f_i=n$. The set $\{f_i\}$ can also be written $f(x_i)$. If $\{f_i\}$ is regarded as a set of weights (for example attribute values) associated with the $\{x_i\}$, it may be written as the set $\{w_i\}$ or $w(x_i)$
$\{p_i\}$	A set of k probabilities (k<=n), estimated from a dataset or theoretically derived. With a finite set of values $\{x_i\}$, $p_i=f_i/n$. If $\{x_i\}$ represents a set of k classes or ranges then p_i is the probability of finding an occurrence in the i^{th} class or range, i.e. the proportion of events or values occurring in that class or range. The sum $\Sigma p_i=1$. If a set of frequencies, $\{f_i\}$, have been standardized by dividing each value f_i by their sum, Σf_i, then $\{p_i\}$ is equivalent to $\{f_i\}$
Σ	Summation symbol, e.g. $x_1+x_2+x_3+ \ldots +x_n$. If no limits are shown the sum is assumed to apply to all subsequent elements, otherwise upper and/or lower limits for summation are provided
Π	Product symbol, e.g. $x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n$. If no limits are shown the product is assumed to apply to all subsequent elements, otherwise upper and/or lower limits for multiplication are provided
^	Used here in conjunction with Greek symbols (directly above) to indicate a value is an estimate of the true population value. Sometimes referred to as “hat”

~	Is distributed as, for example *y*~N(0,1) means the variable *y* has a distribution that is Normal with a mean of 0 and standard deviation of 1
!	Factorial symbol. *z*=*x*! means *z*=*x*(*x*-1)(*x*-2)…1. *x*>=0. Usually applied to integer values of *x*. May be defined for fractional values of *x* using the Gamma function (Table 1-3)
≡	'Equivalent to' symbol
≈	'Approximately equal to' symbol
∈	'Belongs to' symbol, e.g. *x* ∈[0,2] means that *x* belongs to/is drawn from the set of all values in the closed interval [0,2]; *x*∈{0,1} means that *x* can take the values 0 and 1
≤	Less than or equal to, represented in the text where necessary by <= (provided in this form to support display by some web browsers)
≥	Greater than or equal to, represented in the text where necessary by >= (provided in this form to support display by some web browsers)

1.5.2 Statistical measures and related formulas

Table 1-3, below, provides a list of common measures (univariate statistics) applied to datasets, and associated formulas for calculating the measure from a sample dataset in summation form (rather than integral form) where necessary. In some instances these formulas are adjusted to provide estimates of the population values rather than those obtained from the sample of data one is working on.

Many of the measures can be extended to two-dimensional forms in a very straightforward manner, and thus they provide the basis for numerous standard formulas in spatial statistics. For a number of univariate statistics (variance, skewness, kurtosis) we refer to the notion of (estimated) *moments* about the mean. These are computations of the form

$$\sum \left(x_i - \bar{x}\right)^r, r = 1,2,3\ldots$$

When r=1 this summation will be 0, since this is just the difference of all values from the mean. For values of r>1 the expression provides measures that are useful for describing the shape (spread, skewness, peakedness) of a distribution, and simple variations on the formula are used to define the correlation between two or more datasets (the *product moment* correlation). The term *moment* in this context comes from physics, i.e. like 'momentum' and 'moment of inertia', and in a spatial (2D) context provides the basis for the definition of a centroid — the center of mass or center of gravity of an object, such as a polygon (see further, Section 4.2.5, Centroids and centers).

Table 1-3 Common formulas and statistical measures

This table of measures has been divided into 9 subsections for ease of use. Each is provided with its own subheading. For more details on these topics, see the relevant topic within the StatsRef website.

Counts and specific values

Measure	Definition	Expression(s)
Count	The number of data values in a set	$Count(\{x_i\})=n$
Top *m*, Bottom *m*	The set of the largest (smallest) *m* values from a set. May be generated via an SQL command	$Top_m\{x_i\}=\{X_{n-m+1},\dots X_{n-1},X_n\}$; $Bot_m\{x_i\}=\{X_1,X_2,\dots X_m\}$;
Variety	The number of distinct i.e. different data values in a set. Some packages refer to the variety as diversity, which should not be confused with information theoretic and other diversity measures	
Majority	The most common i.e. most frequent data values in a set. Similar to mode (see below), but often applied to raster datasets at the neighborhood or zonal level. For general datasets the term should only be applied to cases where a given class is 50%+ of the total	
Minority	The least common i.e. least frequently occurring data values in a set. Often applied to raster datasets at the neighborhood or zonal level	
Maximum, *Max*	The maximum value of a set of values. May not be unique	$Max\{x_i\}=X_n$
Minimum, *Min*	The minimum value of a set of values. May not be unique	$Min\{x_i\}=X_1$
Sum	The sum of a set of data values	$\sum_{i=1}^{n} x_i$

Measures of centrality

Measure	Definition	Expression(s)
Mean (arithmetic)	The arithmetic average of a set of data values (also known as the *sample mean* where the data are a sample from a larger population). Note that if the set $\{f_i\}$ are regarded as weights rather than frequencies the result is known as the *weighted mean*. Other mean values include the geometric and harmonic mean. The population mean is often denoted by the symbol μ. In many instances the sample mean is the best (unbiased) estimate of the population mean and is sometimes denoted by μ with a ^ symbol above it) or as a variable such as *x* with a bar above it.	$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ $\bar{x}=\sum_{i=1}^{n}f_ix_i\bigg/\sum_{i=1}^{n}f_i$ $\bar{x}=\sum_{i=1}^{n}p_ix_i$
Mean (harmonic)	The harmonic mean, *H*, is the mean of the reciprocals of the data values, which is then adjusted by taking the reciprocal of the result. The harmonic mean is less than or equal to the geometric mean, which is less than or equal to the arithmetic mean	$H=\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$
Mean (geometric)	The geometric mean, *G*, is the mean defined by taking the products of the data values and then adjusting the value by taking the n^{th} root of the result. The geometric mean is greater than or equal to the harmonic mean and is less than or equal to the arithmetic mean	$G=\left(\prod_{i=1}^{n}x_i\right)^{1/n}$ hence $\log(G)=\frac{1}{n}\sum_{i=1}^{n}\log(x_i)$
Mean (power)	The general (limit) expression for mean values. Values for *p* give the following means: *p*=1 arithmetic; *p*=2 root mean square; *p*=-1 harmonic. Limit values for *p* (i.e. as *p* tends to these values) give the following means: *p*=0 geometric; *p*=-∞ minimum; *p*=∞ maximum	$M=\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{p}\right)^{1/p}$
Trim-mean, *TM*, *t*, Olympic mean	The mean value computed with a specified percentage (proportion), *t*/2, of values removed from each tail to eliminate the highest and lowest outliers and extreme values. For small samples a specific number of observations (e.g. 1) rather than a percentage, may be ignored. In general an equal number, *k*, of high and low values should be removed and the number of observations summed should equal	$TM=\frac{1}{n(1-t)}\sum_{i=nt/2}^{n(1-t/2)}X_i$ $t\in[0,1]$

Measure	Definition	Expression(s)
	$n(1-t)$ expressed as an integer. This variant is sometimes described as the Olympic mean, as has been used in scoring Olympic gymnastics for example	
Mode	The most common or frequently occurring value in a set. Where a set has one dominant value or range of values it is said to be unimodal; if there are several commonly occurring values or ranges it is described as multi-modal. Note that arithmetic mean-mode≈3 (arithmetic mean-median) for many unimodal distributions	
Median, *Med*	The middle value in an ordered set of data if the set contains an odd number of values, or the average of the two middle values if the set contains an even number of values. For a continuous distribution the median is the 50% point (0.5) obtained from the cumulative distribution of the values or function	$Med\{x_i\}=X_{(n+1)/2}$; n odd $Med\{x_i\}=(X_{n/2}+X_{n/2+1})/2$; n even
Mid-range, *MR*	The middle value of the Range	$MR\{x_i\}=\text{Range}/2$
Root mean square (*RMS*)	The root of the mean of squared data values. Squaring removes negative values	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$

Measures of spread

Measure	Definition	Expression(s)
Range	The difference between the maximum and minimum values of a set	$Range\{x_i\}=X_n-X_1$
Lower quartile (25%), *LQ*	In an ordered set, 25% of data items are less than or equal to the upper bound of this range. For a continuous distribution the LQ is the set of values from 0% to 25% (0.25) obtained from the cumulative distribution of the values or function. Treatment of cases where *n* is even and *n* is odd, and when *i* runs from 1 to n or 0 to *n* vary	$LQ=\{X_1, \ldots X_{(n+1)/4}\}$
Upper quartile (75%), *UQ*	In an ordered set 75% of data items are less than or equal to the upper bound of this range.	$UQ=\{X_{3(n+1)/4}, \ldots X_n\}$

Measure	Definition	Expression(s)
	For a continuous distribution the UQ is the set of values from 75% (0.75) to 100% obtained from the cumulative distribution of the values or function. Treatment of cases where *n* is even and *n* is odd, and when *i* runs from 1 to *n* or 0 to *n* vary	
Inter-quartile range, *IQR*	The difference between the lower and upper quartile values, hence covering the middle 50% of the distribution. The inter-quartile range can be obtained by taking the median of the dataset, then finding the median of the upper and lower halves of the set. The IQR is then the difference between these two secondary medians	$IQR=UQ-LQ$
Trim-range, *TR*, *t*	The range computed with a specified percentage (proportion), $t/2$, of the highest and lowest values removed to eliminate outliers and extreme values. For small samples a specific number of observations (e.g. 1) rather than a percentage, may be ignored. In general an equal number, *k*, of high and low values are removed (if possible)	$TR_t=X_{n(1-t/2)}-X_{nt/2},\ t\in[0,1]$ $TR_{50\%}=IQR$
Variance, *Var*, σ^2, s^2, μ_2	The average squared difference of values in a dataset from their population mean, μ, or from the sample mean (also known as the sample variance where the data are a sample from a larger population). Differences are squared to remove the effect of negative values (the summation would otherwise be 0). The third formula is the frequency form, where frequencies have been standardized, i.e. $\Sigma f_i=1$. *Var* is a function of the 2nd moment about the mean. The population variance is often denoted by the symbol μ_2 or σ^2. The estimated population variance is often denoted by s^2 or by σ^2 with a ^ symbol above it	$Var=\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ $Var=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ $Var=\sum_{i=1}^{n}f_i(x_i-\bar{x})^2$ $Var=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})$ $s^2=\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Measure	Definition	Expression(s)
Standard deviation, *SD*, *s* or *RMSD*	The square root of the variance, hence it is the Root Mean Squared Deviation (RMSD). The population standard deviation is often denoted by the symbol σ. SD* shows the estimated population standard deviation (sometimes denoted by σ with a ^ symbol above it or by *s*)	$SD=\sqrt{Var}=\sigma$ $SD=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ $SD^*=\hat{\sigma}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
Standard error of the mean, *SE*	The estimated standard deviation of the mean values of *n* samples from the same population. It is simply the sample standard deviation reduced by a factor equal to the square root of the number of samples, *n*>=1	$SE=\frac{SD}{\sqrt{n}}$
Root mean squared error, *RMSE*	The standard deviation of samples from a known set of true values, x_i*. If x_i* are estimated by the mean of sampled values RMSE is equivalent to RMSD	$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-x_i^*\right)^2}$
Mean deviation/error, *MD* or *ME*	The mean deviation of samples from the known set of true values, x_i*	$MD=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-x_i^*\right)$
Mean absolute deviation/error, *MAD* or *MAE*	The mean absolute deviation of samples from the known set of true values, x_i*	$MAE=\frac{1}{n}\sum_{i=1}^{n}\left\|x_i-x_i^*\right\|$
Covariance, *Cov*	Literally the pattern of common (or co-) variation observed in a collection of two (or more) datasets, or partitions of a single dataset. Note that if the two sets are the same the covariance is the same as the variance	$Cov(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ $Cov(x,x)=Var(x)$
Correlation/ product moment or Pearson's correlation coefficient, *r*	A measure of the similarity between two (or more) paired datasets. The correlation coefficient is the ratio of the covariance to the product of the standard deviations. If the two datasets are the same or perfectly matched this will give a result=1	$r=Cov(x,y)/SD_xSD_y$ $r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

Measure	Definition	Expression(s)
Coefficient of variation, *CV*	The ratio of the standard deviation to the mean, sometime computed as a percentage. If this ratio is close to 1, and the distribution is strongly left skewed, it may suggest the underlying distribution is Exponential. Note, mean values close to 0 may produce unstable results	$SD / \bar{x}$
Variance mean ratio, *VMR*	The ratio of the variance to the mean, sometime computed as a percentage. If this ratio is close to 1, and the distribution is unimodal and relates to count data, it may suggest the underlying distribution is Poisson. Note, mean values close to 0 may produce unstable results	$Var / \bar{x}$

Measures of distribution shape

Measure	Definition	Expression(s)
Skewness, α_3	If a frequency distribution is unimodal and symmetric about the mean it has a skewness of 0. Values greater than 0 suggest skewness of a unimodal distribution to the right, whilst values less than 0 indicate skewness to the left. A function of the 3rd moment about the mean (denoted by α_3 with a ^ symbol above it for the sample skewness)	$\alpha_3 = \frac{1}{n\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^3$ $\alpha_3 = \frac{1}{n\hat{\sigma}^3}\sum_{i=1}^{n}(x_i-\bar{x})^3$ $\hat{\alpha}_3 = \frac{n}{(n-1)(n-2)\hat{\sigma}^3}\sum_{i=1}^{n}(x_i-\bar{x})^3$
Kurtosis, α_4	A measure of the peakedness of a frequency distribution. More pointy distributions tend to have high kurtosis values. A function of the 4th moment about the mean. It is customary to subtract 3 from the raw kurtosis value (which is the kurtosis of the Normal distribution) to give a figure relative to the Normal (denoted by α_4 with a ^ symbol above it for the sample kurtosis)	$\alpha_4 = \frac{1}{n\hat{\sigma}^4}\sum_{i=1}^{n}(x_i-\bar{x})^4$ $\alpha_4 = \frac{1}{n\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^4$ $\hat{\alpha}_4 = \frac{a}{\hat{\sigma}^4}\sum_{i=1}^{n}(x_i-\bar{x})^4 - b$ where

Measure	Definition	Expression(s)
		$a=\dfrac{n(n+1)}{(n-1)(n-2)(n-3)}$, $b=\dfrac{3(n-1)^2}{(n-2)(n-3)}$

Measures of complexity and dimensionality

Measure	Definition	Expression(s)
Information statistic (Entropy), ***I*** (Shannon's)	A measure of the amount of pattern, disorder or *information*, in a set $\{x_i\}$ where p_i is the proportion of events or values occurring in the i^{th} class or range. Note that if $p_i=0$ then $p_i\log_2(p_i)$ is 0. ***I*** takes values in the range $[0,\log_2(k)]$. The lower value means all data falls into 1 category, whilst the upper means all data are evenly spread	$I=-\sum_{i=1}^{k}p_i\log_2(p_i)$
Information statistic (Diversity), *Div*	Shannon's entropy statistic (see above) standardized by the number of classes, *k*, to give a range of values from 0 to 1	$Div=\dfrac{-\sum_{i=1}^{k}p_i\log_2(p_i)}{\log_2(k)}$
Dimension (topological), D_T	Broadly, the number of (intrinsic) coordinates needed to refer to a single point anywhere on the object. The dimension of a point=0, a rectifiable line=1, a surface=2 and a solid=3. See text for fuller explanation. The value 2.5 (often denoted 2.5D) is used in GIS to denote a planar region over which a single-valued attribute has been defined at each point (e.g. height). In mathematics topological dimension is now equated to a definition similar to cover dimension (see below)	D_T=0,1,2,3,…
Dimension (capacity, cover or fractal), D_c	Let $N(h)$ represent the number of small elements of edge length *h* required to cover an object. For a line, length 1, each element has length $1/h$. For a plane surface each element (small square of side length $1/h$) has area $1/h^2$, and for a volume, each element is a cube with volume $1/h^3$. More generally $N(h)=1/h^D$, where D is the topological dimension, so $N(h)=h^{-D}$ and thus	$D_c=-\lim\dfrac{\ln N(h)}{\ln(h)},\ h\to 0^+$ $D_c\geq 0$

Measure	Definition	Expression(s)
	log(*N*(*h*))=-*D*log(*h*) and so D_c=-log(*N*(*h*))/log(*h*). D_c may be fractional, in which case the term *fractal* is used	

Common distributions

Measure	Definition	Expression(s)
Uniform (continuous)	All values in the range are equally likely. Mean=*a*/2, variance=a^2/12. Here we use *f*(*x*) to denote the probability distribution associated with continuous valued variables *x*, also described as a *probability density function*	$f(x)=\frac{1}{a};x\in[0,a]$
Binomial (discrete)	The terms of the Binomial give the probability of *x* successes out of *n* trials, for example 3 heads in 10 tosses of a coin, where *p*=probability of success and *q*=1-*p*=probability of failure. Mean, *m*=*np*, variance=*npq*. Here we use *p*(*x*) to denote the probability distribution associated with discrete valued variables *x*	$p(x)=\frac{n!}{(n-x)!x!}p^x q^{1-x}\ ;\ x=1,2,\ldots n$
Poisson (discrete)	An approximation to the Binomial when *p* is very small and *n* is large (>100), but the mean *m*=*np* is fixed and finite (usually not large). Mean=variance=*m*	$p(x)=\frac{m^x}{x!}e^{-m}\ ;\ x=1,2,\ldots n$
Normal (continuous)	The distribution of a measurement, *x*, that is subject to a large number of independent, random, additive errors. The Normal distribution may also be derived as an approximation to the Binomial when *p* is not small (e.g. *p*≈1/2) and *n* is large. If μ=mean and σ=standard deviation, we write $N(\mu,\sigma)$ as the Normal distribution with these parameters. The Normal- or *z*-transform $z=(x-\mu)/\sigma$ changes (normalizes) the distribution so that it has a zero mean and unit variance, N(0,1). The distribution of *n* mean values of independent random variables drawn from *any* underlying distribution is also Normal (*Central Limit Theorem*)	$f(z)=\frac{1}{\sqrt{2\pi}}e^{-z/2};\ z\in[-\infty,\infty]$

Data transforms and back transforms

Measure	Definition	Expression(s)
Log	If the frequency distribution for a dataset is broadly unimodal and left-skewed, the natural log transform (logarithms base e) will adjust the pattern to make it more symmetric/similar to a Normal distribution. For variates whose values may range from 0 upwards a value of 1 is often added to the transform. Back transform with the exp() function	$z=\ln(x)$ or $z=\ln(x+1)$ n.b. $\ln(x)=\log_e(x)=\log_{10}(x)*\log_{10}(e)$ $x=\exp(z)$ or $x=\exp(z)-1$
Square root (Freeman-Tukey)	A transform that may adjust the dataset to make it more similar to a Normal distribution. For variates whose values may range from 0 upwards a value of 1 is often added to the transform. For 0<=*x*<=1 (e.g. rate data) the combined form of the transform is often used, and is known as the Freeman-Tukey (FT) transform	$z=\sqrt{x}$, or $z=\sqrt{x+1}$, or $z=\sqrt{x}+\sqrt{x+1}$ (FT) $x=z^2$, or $x=z^2-1$
Logit	Often used to transform binary response data, such as survival/non-survival or present/absent, to provide a continuous value in the range (-∞,∞), where *p* is the proportion of the sample that is 1 (or 0). The inverse or back-transform is shown as *p* in terms of *z*. This transform avoids concentration of values at the ends of the range. For samples where proportions *p* may take the values 0 or 1 a modified form of the transform may be used. This is typically achieved by adding 1/2*n* to the numerator and denominator, where *n* is the sample size. Often used to correct S-shaped (*logistic*) relationships between response and explanatory variables	$z=\ln\left(\frac{p}{1-p}\right), p\in[0,1]$ $p=\frac{e^z}{1+e^z}$
Normal, *z*-transform	This transform normalizes or standardizes the distribution so that it has a zero mean and unit variance. If $\{x_i\}$ is a set of *n* sample *mean* values from *any* probability distribution with mean μ and variance σ^2 then the *z*-transform shown here as z_2 will be distributed N(0,1) for large *n* (Central	$z_1=\frac{(x-\mu)}{\sigma}$ $z_2=\frac{(x-\mu)}{\sigma/\sqrt{n}}$

Measure	Definition	Expression(s)
	Limit Theorem). The divisor in this instance is the standard error. In both instances the standard deviation must be non-zero	
Box-Cox, power transforms	A family of transforms defined for positive data values only, that often can make datasets more Normal; *k* is a parameter. The inverse or back-transform is also shown as *x* in terms of *z*	$z=\frac{(x^k-1)}{k},\ k>0, x>0$ $x=(kz+1)^{1/k},\ k>0$
Angular transforms (Freeman-Tukey)	A transform for proportions, *p*, designed to spread the set of values near the end of the range. *k* is typically 0.5. Often used to correct S-shaped relationships between response and explanatory variables. If *p*=*x*/*n* then the Freeman-Tukey (FT) version of this transform is the averaged version shown. This is a variance-stabilizing transform	$z=\sin^{-1}\left(p^k\right) p=\sin(z)^{1/k}$ $z=\sin^{-1}\left(\sqrt{\frac{x}{n+1}}\right)+\sin^{-1}\left(\sqrt{\frac{x+1}{n+1}}\right)$ (FT)

Selected functions

Measure	Definition	Expression(s)
Bessel functions of the first kind	Bessel functions occur as the solution to specific differential equations. They are described with reference to a parameter known as the order, shown as a subscript. For non-negative real orders Bessel functions can be represented as an infinite series. Order 0 expansions are shown here for standard (*J*) and modified (*I*) Bessel functions. Usage in spatial analysis arises in connection with directional statistics and spline curve fitting. See the Mathworld website entry for more details	$J_0(\kappa)=\sum_{i=0}^{\infty}\frac{(-1)^i(\kappa/2)^{2i}}{(i!)^2}$ and $I_0(\kappa)=\sum_{i=0}^{\infty}\frac{(\kappa/2)^{2i+1}}{i!(i+1)!}$
Exponential integral function, $E_1(x)$	A definite integral function. Used in association with spline curve fitting. See the Mathworld website entry for more details	$E_1(x)=\int_1^{\infty}\frac{e^{-tx}}{t}dt$
Gamma function, Γ	A widely used definite integral function. For integer values of *x*: $\Gamma(x)=(x-1)!$ and $\Gamma(x/2)=(x/2-1)!$ so $\Gamma(3/2)=(1/2)!/2=(\sqrt{\pi})/2$	$\Gamma(x)=\int_0^{\infty}x^{1/2}e^{-x}dx$ $\Gamma(1/2)=\sqrt{\pi}$

Measure	Definition	Expression(s)
	See the Mathworld website entry for more details	

Matrix expressions

Measure	Definition	Expression(s)
Identity	A matrix with diagonal elements 1 and off-diagonal elements 0	$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .. & .. & .. & .. \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Determinant	Determinants are only defined for square matrices. Let **A** be an *n* by *n* matrix with elements $\{a_{ij}\}$. The matrix $\mathbf{M}_{ij}$ here is a subset of **A** known as the *minor*, formed by eliminating row *i* and column *j* from **A**. An *n* by *n* matrix, **A**, with Det=0 is described as *singular*, and such a matrix has no inverse. If Det(**A**) is very close to 0 it is described as *ill-conditioned*	$\lvert\mathbf{A}\rvert$, Det($\mathbf{A}$)
Inverse	The matrix equivalent of division in conventional algebra. For a matrix, **A**, to be invertible its determinant must be non-zero, and ideally not very close to zero. A matrix that has an inverse is by definition non-singular. A symmetric real-valued matrix is *positive definite* if all its eigenvalues are positive, whereas a *positive semi-definite* matrix allows for some eigenvalues to be 0. A matrix, **A**, that is invertible satisfies the relation $\mathbf{AA}^{-1}=\mathbf{I}$	$\mathbf{A}^{-1}$
Transpose	A matrix operation in which the rows and columns are transposed, i.e. in which elements a_{ij} are swapped with a_{ji} for all *i,j*. The inverse of a transposed matrix is the same as the transpose of the matrix inverse	$\mathbf{A}^{\mathbf{T}}$ or $\mathbf{A}'$ $(\mathbf{A}^{\mathbf{T}})^{-1}=(\mathbf{A}^{-1})^{\mathbf{T}}$
Symmetric	A matrix in which element $a_{ij}=a_{ji}$ for all *i,j*	$\mathbf{A}=\mathbf{A}^{\mathbf{T}}$
Trace	The sum of the diagonal elements of a matrix, a_{ii} — the sum of the eigenvalues of a matrix equals its trace	Tr($\mathbf{A}$)

Measure	Definition	Expression(s)
Eigenvalue, Eigenvector	If **A** is a real-valued *k* by *k* square matrix and **x** is a non-zero real-valued vector, then a scalar λ that satisfies the equation shown in the adjacent column is known as an eigenvalue of **A** and **x** is an eigenvector of **A**. There are *k* eigenvalues of **A**, each with a corresponding eigenvector. The matrix **A** can be decomposed into three parts, as shown, where **E** is a matrix of its eigenvectors and **D** is a diagonal matrix of its eigenvalues	$(\mathbf{A}-\lambda\mathbf{I})\mathbf{x}=\mathbf{0}$ $\mathbf{A}=\mathbf{EDE}^{-1}$ (diagonalization)

Chapter 2

2 Conceptual Frameworks for Spatial Analysis

Geospatial analysis provides a distinct perspective on the world, a unique lens through which to examine events, patterns, and processes that operate on or near the surface of our planet. It makes sense, then, to introduce the main elements of this perspective, the conceptual framework that provides the background to spatial analysis, as a preliminary to the main body of this Guide's material. This chapter provides that introduction. It is divided into four main sections. The first, Basic Primitives, describes the basic components of this view of the world — the classes of things that a spatial analyst recognizes in the world, and the beginnings of a system of organization of geographic knowledge. The second section, Spatial Relationships, describes some of the structures that are built with these basic components and the relationships between them that interest geographers and others. The third section, Spatial Statistics, introduces the concepts of spatial statistics, including probability, that provide perhaps the most sophisticated elements of the conceptual framework. Finally, the fourth section, Spatial Data Infrastructure, discusses some of the basic components of the data infrastructure that increasingly provides the essential facilities for spatial analysis.

The domain of geospatial analysis is the surface of the Earth, extending upwards in the analysis of topography and the atmosphere, and downwards in the analysis of groundwater and geology. In scale it extends from the most local, when archaeologists record the locations of pieces of pottery to the nearest centimeter or property boundaries are surveyed to the nearest millimeter, to the global, in the analysis of sea surface temperatures or global warming. In time it extends backwards from the present into the analysis of historical population migrations, the discovery of patterns in archaeological sites, or the detailed mapping of the movement of continents, and into the future in attempts to predict the tracks of hurricanes, the melting of the Greenland ice-cap, or the likely growth of urban areas. Methods of spatial analysis are robust and capable of operating over a range of spatial and temporal scales.

Ultimately, geospatial analysis concerns *what* happens *where*, and makes use of geographic information that links features and phenomena on the Earth's surface to their locations. This sounds very simple and straightforward, and it is not so much the basic information as the structures and arguments that can be built on it that provide the richness of spatial analysis. In principle there is no limit to the complexity of spatial analytic techniques that might find some application in the world, and might be used to tease out interesting insights and support practical actions and decisions. In reality, some techniques are simpler, more useful, or more insightful than others, and the contents of this Guide reflect that reality. This chapter is about the underlying concepts that are employed, whether it be in simple, intuitive techniques or in advanced, complex mathematical or computational ones.

Spatial analysis exists at the interface between the human and the computer, and both play important roles. The concepts that humans use to understand, navigate, and exploit the world around them are mirrored in the concepts of spatial analysis. So the discussion that follows will often appear to be following parallel tracks — the track of human intuition on the one hand, with all its vagueness and informality, and the track of the formal, precise world of spatial analysis on the other. The relationship between these two tracks forms one of the recurring themes of this Guide.

2.1 Basic Primitives

The building blocks for any form of spatial analysis are a set of basic primitives that refer to the place or places of interest, their attributes and their arrangement. These basic primitives are discussed in the following subsections.

2.1.1 Place

At the center of all spatial analysis is the concept of *place*. The Earth's surface comprises some 500,000,000 sq km, so there would be room to pack half a billion industrial sites of 1 sq km each (assuming that nothing else required space, and that the two-thirds of the Earth's surface that is covered by water was as acceptable as the one-third that is land); and 500 trillion sites of 1 sq m each (roughly the space occupied by a sleeping human). People identify with places of various sizes and shapes, from the room to the parcel of land, to the neighborhood, the city, the county, the state or province, or the nation-state. Places may overlap, as when a watershed spans the boundary of two counties, and places may be nested hierarchically, as when counties combine to form a state or province.

Places often have names, and people use these to talk about and distinguish between places. Some names are official, having been recognized by national or state agencies charged with bringing order to geographic names. In the U.S., for example, the Board on Geographic Names exists to ensure that all agencies of the federal government use the same name in referring to a place, and to ensure as far as possible that duplicate names are removed from the landscape. A list of officially sanctioned names is termed a *gazetteer*, though that word has come to be used for any list of geographic names.

Places change continually, as people move, climate changes, cities expand, and a myriad of social and physical processes affect virtually every spot on the Earth's surface. For some purposes it is sufficient to treat places as if they were static, especially if the processes that affect them are comparatively slow to operate. It is difficult, for example, to come up with instances of the need to modify maps as continents move and mountains grow or shrink in response to earthquakes and erosion. On the other hand it would be foolish to ignore the rapid changes that occur in the social and economic makeup of cities, or the constant movement that characterizes modern life. Throughout this Guide, it will be important to distinguish between these two cases, and to judge whether time is or is not important.

People associate a vast amount of information with places. Three Mile Island, Sellafield, and Chernobyl are associated with nuclear reactors and accidents, while Tahiti and Waikiki conjure images of (perhaps somewhat faded) tropical paradise. One of the roles of places and their names is to link together what is known in useful ways. So for example the statements "I am going to London next week" and "There's always something going on in London" imply that I will be having an exciting time next week. But while "London" plays a useful role, it is nevertheless vague, since it might refer to the area administered by the Greater London Authority, the area inside the M25 motorway, or something even less precise and determined by the context in which the name is used. Science clearly needs something better if information is to be linked exactly to places, and if places are to be matched, measured, and subjected to the rigors of spatial analysis.

The basis of rigorous and precise definition of place is a *coordinate system*, a set of measurements that allows place to be specified unambiguously and in a way that is meaningful to everyone. The Meridian Convention of 1884 established the Greenwich Observatory in London as the basis of longitude, replacing a confusing multitude of earlier systems. Today, the World Geodetic System (WGS84) of 1984 and subsequent adjustments provide a highly accurate pair of coordinates for every location on the Earth's surface (and incidentally place the line of zero longitude about 100m east of the Greenwich Observatory). Elevation continues to be problematic, however, since countries and even agencies within countries insist on their own definitions of what marks zero elevation, or exactly how to define "sea level". Many other coordinate systems are in use, but most are easily converted to and from latitude/longitude. Today it is possible to measure location directly, using the Global Positioning System (GPS) or its Russian counterpart GLONASS (and in future its European counterpart Galileo). Spatial analysis is most often applied in a two-dimensional space. But applications that extend above or below the surface of the Earth must often be handled as three-dimensional. Time sometimes adds a fourth dimension, particularly in studies that examine the dynamic nature of phenomena.

2.1.2 Attributes

Attribute has become the preferred term for any recorded characteristic or property of a place (see Table 1-1 for a more formal definition). A place's name is an obvious example of an attribute, but a vast array of other options has proven useful for various purposes. Some are measured, including elevation, temperature, or rainfall. Others are the result of classification, including soil type, land-use or land cover type, or rock type. Government agencies provide a host of attributes in the form of statistics, for places ranging in size from countries all the way down to neighborhoods and streets. The characteristics that people assign rightly or mistakenly to places, such as "expensive", "exciting", "smelly", or "dangerous" are also examples of attributes. Attributes can be more than simple values or terms, and today it is possible to construct information systems that contain entire collections of images as attributes of hotels, or recordings of birdsong as attributes of natural areas. But while these are certainly feasible, they are beyond the bounds of most techniques of spatial analysis.

Within GIS the term *attribute* usually refers to records in a data table associated with individual features in a vector map or cells in a grid (*raster* or image file). Sample vector data attributes are illustrated in Figure 2-1A where details of major wildfires recorded in Alaska are listed. Each row relates to a single polygon feature that identifies the spatial extent of the fire recorded. Most GIS packages do not display a separate attribute table for raster data, since each grid cell contains a single data item, which is the value at that point and can be readily examined. ArcGIS is somewhat unusual in that it provides an attribute table for raster data (see Figure 2-1B).

Figure 2-1 Attribute tables – spatial datasets

A. Alaskan fire dataset - polygon attributes

	OBJECTID	HA_BURNED	ACRES_BURN	YEAR_	BOUNDARY	COUNTRY	SITENAME
1	1	2571	6352	2004	Y	AK	446
2	2	25518	63057	2004	Y	AK	342
3	3	281	695	2004	Y	AK	348
4	4	8789	21719	2004	Y	AK	578
5	5	1968	4864	2004	Y	AK	164
6	6	60	149	2004	Y	AK	533
7	7	75	186	2004	Y	AK	452
8	8	86953	214869	2004	Y	AK	158
9	9	135	333	2004	Y	AK	213

B. DEM dataset - raster file attribute table (ArcGIS)

ObjectID	Value	Count
275	449	167
276	450	212
277	451	166
278	452	159
279	453	144

Rows in this raster attribute table provide a count of the number of grid cells (pixels) in the raster that have a given value, e.g. 144 cells have a value of 453 meters. Furthermore, the linking between the attribute table visualization and mapped data enables all cells with elevation=453 to be selected and highlighted on the map.

Many terms have been adopted to describe attributes. From the perspective of spatial analysis the most useful divides attributes into scales or levels of measurement, as follows:

- **Nominal**. An attribute is nominal if it successfully distinguishes between locations, but without any implied ranking or potential for arithmetic. For example, a telephone number can be a useful attribute of a place, but the number itself generally has no numeric meaning. It would make no sense to add or divide telephone numbers, and there is no sense in which the number 9680244 is more or better than the number 8938049. Likewise, assigning arbitrary numerical values to classes of land type, e.g. 1=arable, 2=woodland, 3=marsh, 4=other is simply a convenient form of naming (the values are nominal). SITENAME in Figure 2-1A is an example of a nominal attribute, as is OBJECTID, even though both happen to be numeric

- **Ordinal**. An attribute is ordinal if it implies a ranking, in the sense that Class 1 may be better than Class 2, but as with nominal attributes no arithmetic operations make sense, and there is no implication that Class 3 is worse than Class 2 by the precise amount by which Class 2 is worse than Class 1. An example of an ordinal scale might be preferred locations for residences — an individual may prefer some areas of a city to others, but such differences between areas may be barely noticeable or quite profound. Note that although OBJECTID in Figure 2-1A appears to be an ordinal variable it is not, because the IDs are provided as unique names only, and could equally well be in any order and use any values that provided uniqueness (and typically, in this example, are required to be integers)
- **Interval**. The remaining three types of attributes are all quantitative, representing various types of measurements. Attributes are interval if differences make sense, as they do for example with measurements of temperature on the Celsius or Fahrenheit scales, or for measurements of elevation above sea level
- **Ratio**. Attributes are ratio if it makes sense to divide one measurement by another. For example, it makes sense to say that one person weighs twice as much as another person, but it makes no sense to say that a temperature of 20 Celsius is twice as warm as a temperature of 10 Celsius, because while weight has an absolute zero Celsius temperature does not (but on an absolute scale of temperature, such as the Kelvin scale, 200 degrees can indeed be said to be twice as warm as 100 degrees). It follows that negative values cannot exist on a ratio scale. HA_BURNED and ACRES_BURN in Figure 2-1A are examples of ratio attributes. Note that only one of these two attribute columns is required, since they are simple multiples of one another
- **Cyclic**. Finally, it is not uncommon to encounter measurements of attributes that represent directions or cyclic phenomena, and to encounter the awkward property that two distinct points on the scale can be equal — for example, 0 and 360 degrees are equal. Directional data are cyclic (Figure 2-2), as are calendar dates. Arithmetic operations are problematic with cyclic data, and special techniques are needed, such as the techniques used to overcome the Y2K problem, when the year after (19)99 was (20)00. For example, it makes no sense to average 1degree and 359degrees to get 180degrees, since the average of two directions close to north clearly is not south. Mardia and Jupp (1999) provide a comprehensive review of the analysis of directional or cyclic data (see further, Section 4.5.1, Directional analysis of linear datasets)

Figure 2-2 Cyclic attribute data — Wind direction, single location

	A	C	D	E
Type	Angles	Date	Month of year	Time
Name	Direction	Date	Month	Time
1	265.000	15/06/2002	June	12:00
2	248.000	15/06/2002	June	13:00
3	232.000	15/06/2002	June	14:00
4	186.000	15/06/2002	June	15:00
5	222.000	15/06/2002	June	16:00
6	252.000	15/06/2002	June	17:00
7	239.000	15/06/2002	June	18:00
8	250.000	15/06/2002	June	19:00
9	237.000	15/06/2002	June	20:00

While this terminology of measurement types is standard, spatial analysts find that another distinction is particularly important. This is the distinction between attributes that are termed *spatially intensive* and *spatially extensive*. Spatially extensive attributes include total population, measures of a place's area or perimeter length, and total income — they are true *only of the place as a whole*.

Spatially intensive attributes include population density, average income, and percent unemployed, and if the place is homogeneous they will be true of any part of the place as well as of the whole. For many purposes it is necessary to keep spatially intensive and spatially extensive attributes apart, because they respond very differently when places are merged or split, and when many types of spatial analysis are conducted.

Since attributes are essentially measured or computed data items associated with a given location or set of locations, they are subject to the same issues as any conventional dataset: sampling error; measurement errors and limitations; mistakes and miscalculations; missing values; temporal and thematic errors and similar issues. Metadata accompanying spatial datasets should assist in assessing the quality of such attribute data, but at least the same level of caution should be applied to spatial attribute data as with any other form of data that one might wish to use or analyze.

2.1.3 Objects

The places discussed in Section 2.1.1, Place, vary enormously in size and shape. Weather observations are obtained from stations that may occupy only a few square meters of the Earth's surface (from instruments that occupy only a small fraction of the station's area), whereas statistics published for Russia are based on a land area of more than 17 million sq km. In spatial analysis it is customary to refer to places as *objects*. In studies of roads or rivers the objects of interest are long and thin, and will often be represented as lines of zero width. In studies of climate the objects of interest may be weather stations of minimal extent, and will often be represented as points. On the other hand many studies of social or economic patterns may need to consider the two-dimensional extent of places, which will therefore be represented as areas, and in some studies where elevations or depths are important it may be appropriate to represent places as volumes. To a spatial statistician, these points, lines, areas, or volumes are known as the attributes' *spatial support*.

Each of these four classes of objects has its own techniques of representation in digital systems. The software for capturing and storing spatial data, analyzing and visualizing them, and reporting the results of analysis must recognize and handle each of these classes. But digital systems must ultimately represent everything in a language of just two characters, 0 and 1 or "off" and "on", and special techniques are required to represent complex objects in this way. In practice, points, lines, and areas are most often represented in the following standard forms:

- **Points** as pairs of coordinates, in latitude/longitude or some other standard system
- **Lines** as ordered sequences of points connected by straight lines
- **Areas** as ordered rings of points, also connected by straight lines to form polygons. In some cases areas may contain holes, and may include separate islands, such as in representing the State of Michigan with its separate Upper Peninsula, or the State of Georgia with its offshore islands. This use of polygons to represent areas is so pervasive that many spatial analysts refer to all areas as polygons, whether or not their edges are actually straight

Lines represented in this way are often termed *polylines*, by analogy to polygons (see Table 1-1 for a more formal definition). Three-dimensional volumes are represented in several different ways, and as yet no one method has become widely adopted as a standard. The related term *edge* is used in several ways within GIS. These include: to denote the border of polygonal regions; to identify the individual links connecting nodes or vertices in a network; and as a general term relating to the distinct or indistinct boundary of areas or zones. In many parts of spatial analysis the related term, *edge effect* is applied. This refers to possible bias in the analysis which arises specifically due to proximity of features to one or more edges. For example, in point pattern analysis computation of distances to the nearest neighboring point, or calculation of the density of points per unit area, may both be subject to edge effects.

Figure 2-3, below, shows a simple example of points, lines, and areas, as represented in a typical map display. The hospital, boat ramp, and swimming area will be stored in the database as points with associated attributes, and symbolized for display. The roads will be stored as polylines, and the road type symbols (U.S. Highway, Interstate Highway) generated from the attributes when each object is displayed. The lake will be stored as two polygons with appropriate attributes. Note how the lake consists of two geometrically disconnected pieces, linked in the database to a single set of attributes — objects in a GIS may consist of multiple parts, as long as each part is of the same type.

Figure 2-3 An example map showing points, lines, and areas appropriately symbolized

see text for explanation

It can be expensive and time-consuming to create the polygon representations of complex area objects, and so analysts often resort to simpler approaches, such as choosing a single *representative point*. But while this may be satisfactory for some purposes, there are obvious problems with representing the entirety of a large country such as Russia as a single point. For example, the distance from Canada to the U.S. computed between representative points in this way would be very misleading, given that they share a very long common boundary.

2.1.4 Maps

Historically, maps have been the primary means to store and communicate spatial data. Objects and their attributes can be readily depicted, and the human eye can quickly discern patterns and anomalies in a well-designed map. Points can be shown as symbols of various kinds, depicting anything from a windmill to a church; lines can be symbolized to distinguish between major roads, minor roads, and rivers; and areas can be symbolized with color, shading, or annotation.

Maps have traditionally existed on paper, as individual sheets or bound into atlases (a term that originated with Mercator, who produced one of the first atlases in the late 16th century). The advent of digital computers has broadened the concept of a map substantially, however. Maps can now take the form of images displayed on the screens of computers, mobile phones or in-vehicle navigation systems. They can be dynamic, showing the Earth spinning on its axis or tracking the movement of migrating birds. Their designs can now go far beyond what was traditionally possible when maps had to be drawn by hand, incorporating a far greater range of color and texture, and even integrating audio and video elements.