SPATIAL DATA HANDLING AND GIS

Atkinson, P.M.
School of Geography, University of Southampton, UK

Keywords: data models, data transformation, GIS cycle, sampling, GIS functionality

Contents

1. Background
2. Geographical Data
3. Data Models
4. Measurement and Sampling
4.1 The Support
4.2 Measurement Error and Accuracy
4.3 Sampling
5. Data Entry, Archiving and Retrieval
6. Data Organization
7. Analysis
8. Accuracy assessment
9. Conclusion
Acknowledgements
Glossary
Bibliography
Biographical Sketch

Summary

This chapter is a pre-cursor to subsequent articles in under the subject topic: *Statistical Analysis in the Geosciences*. It introduces the general concepts underpinning spatial data handling and geographical information systems (GIS) at a non-technical level. This article starts by distinguishing spatial data from aspatial data. The raster and vector data models are discussed with particular emphasis on the relevance of the raster data model to spatial continua. Sampling processes are considered. In particular, the support is shown to be a key concept for spatial data handling and errors of measurement are introduced. The rôle of GIS as a tool for organising geospatial data is emphasized. An example of GIS analysis is given involving overlay. The need to extend overlay to more statistically rigorous methods is emphasized. Finally, the fundamentals of accuracy assessment are given.

1. Background

In the simplest terms, a geographical information system (GIS) is a computer software package for the handling and analysis of geographical data. Geographical data are distinguished from aspatial data in that every value is associated with a location. When used properly, a geographical information system becomes much more than the computer package: it involves the geographical data, the computer hardware, the computer peripherals (such as scanning and plotting devices), the computer operator(s)

and even the decision makers who will use the information output by the GIS. Most often, a GIS is used as a decision-making tool.

A GIS is often used in a sequence of processes or operations that is referred to as the GIS cycle (Figure 1). This cycle starts with the real world and then moves through measurement, data entry, retrieval, analysis, uncertainty analysis, output, decision-making and action. At the end of this sequence, when action has been taken with regard to managing the environment, the cycle begins again. Thus, GIS has great utility as a decision-making and, more generally, management tool. Many organisations now run GIS on an operational basis as part of their management infrastructure.

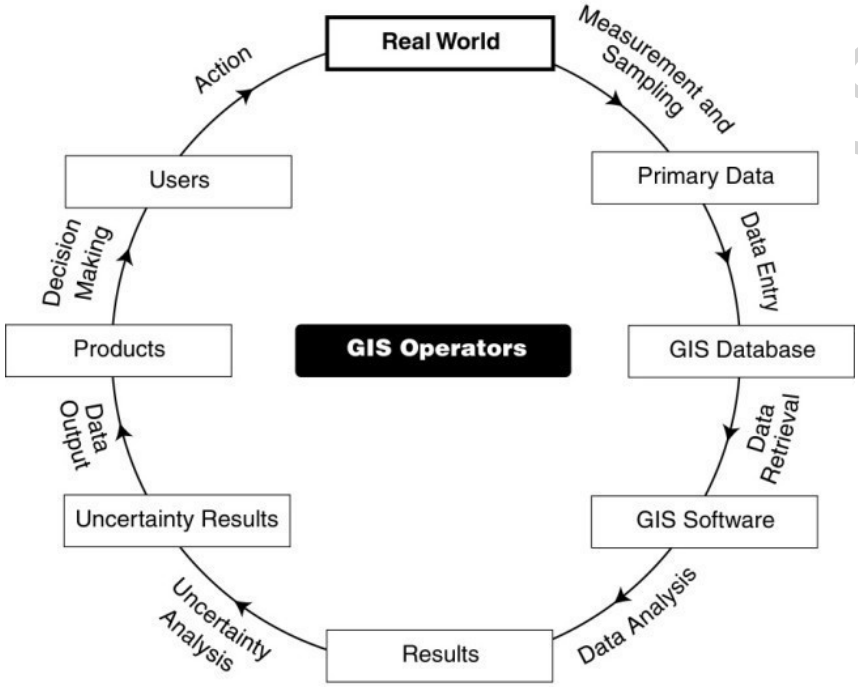

Figure 1. The GIS cycle

In the early days of GIS, in the 1960s and 1970s, a GIS was a single software package run on a powerful hardware platform such as a mainframe computer. The high costs associated with both the software and the hardware, not to mention the high levels of expertise required to set up, run and maintain such systems, meant that GIS were implemented in very few organisations. Further, academics became increasingly frustrated with GIS as the limitations of single software packages were realized. If the software package did not allow a particular kind of analysis then there was no option, but to export the geographical data and perform the analysis outside the GIS. With rapid increases in the performance of office PCs (and the associated decreases in cost) and increases in the availability of diverse software packages, many of which are free and accessible readily via the internet, GIS have changed over the years from being stand-alone monolithic expensive systems to being flexible, inexpensive collections of different software components. Thus, GIS users have been able to build their own

systems to meet their needs at relatively little cost. Facilitated by these changes, GIS are now commonly placed at the heart of the environmental management function of many organisations.

From an academic point-of-view, standard off-the-shelf GIS are now seen as a useful tool for organising data into suitable formats for further analysis, some of which may be conducted within the GIS but much of which may be conducted outside it. This analysis may be achieved using other software, some of which may be written (i.e., in a programming language) by the user. After all, academic research must advance knowledge, not simply re-apply it.

The above view of GIS as somewhat limited and inflexible has spurred the field of GIScience, which is concerned primarily with the advancement of the tools of GIS and their application to particular scientific problems. The fact that this endeavour actually involves a technological research methodology rather than a deductive scientific research methodology (i.e., repeatable controlled experiment on the real world employing deductive logic to test hypotheses) need not detain us here. The important point is that GI research is not constrained to GIS packages and neither should it be. However, GIS provide a useful framework for handling geographical data and for performing basic operations. Such data handling and analysis can be seen as a pre-cursor, necessary to subsequent statistical and analytical analysis of the data, perhaps outside the GIS. It is in this light that GIS is presented in this chapter.

2. Geographical Data

As suggested above, geographical data are distinctive because each value of the property of interest (referred to as an attribute) has associated with it a location defined in a given space. This space is usually two-dimensional (2-D) for geographical properties, and a location within it is defined by an (x, y) pair of values, where x refers to the x-coordinate and y to the y-coordinate in a (x, y) coordinate system. Such coordinate systems include the Universal Tranverse Mercator (UTM) projection, which defines a position for every point on the surface of the Earth, and the British National Grid (BNG) which defines position in terms of "Eastings" (x) and "Northings" (y) relative to a fixed position (0, 0) off the South West coast of Great Britain.

The important point from the above is that geographical data are often presented as an (x, y, z) triplet, where z is the actual attribute value (e.g., elevation, forest biomass, soil moisture, amount of rainfall, type of geology) and x and y represent location in a 2-D coordinate system. In fact, since there often exist many properties that are of interest at a single location (e.g., snow depth may be related to elevation and we are interested in their association), we can generalize to (x, y, z_1, z_2,…z_n) where n is the number of attributes of interest.

The association of attribute and locational data in a GIS is what differentiates geographical analysis from aspatial analysis. Knowledge of the locations of attribute values and, in particular, their *relative* positions adds both opportunities for and complex problems to geographical analysis. For example, techniques that utilize the available spatial information can be used to increase the precision of prediction above

that attainable with equivalent aspatial techniques. A variety of such techniques have been developed to exploit the available spatial information. At the same time, many of the available aspatial techniques (e.g., simple linear regression) depend on statistical independence between data, an assumption that is clearly not met for spatial data where proximate observations are most often likely to be similar, a phenomenon known as spatial dependence (see *Geostatistical Analysis of Spatial Data*). Thus, standard models such as linear regression are not applicable to geographical data without modification. Such opportunities and problems set geographical data analysis apart from traditional statistical analysis.

Locational and attribute data are often separated in a GIS such that the (x, y) and z data are held in different tables. To link them (i.e., to associate the locational and attribute values point by point) each table contains a unique identifier (ID) for each point (e.g., (ID, x, y), $(\mathrm{ID}, z_1, z_2, z_3, ...z_n)$ or (ID, x, y), (ID, z_1) (ID, z_2) (ID, z_3)). This method of data encoding is efficient because location is recorded only once (not once per attribute). Through the unique identifier, it is possible to find the location of a given value and the value at a given location.

-
-
-

Bibliography

Atkinson, P.M. (1996). Optimal sampling strategies for raster-based geographical information systems. *Global Ecology and Biogeography Letters* **5**, 271-280. [An article on sampling design for data entry into GIS].

Atkinson, P.M. and N.J. Tate. (2000). 'Spatial scale problems and geostatistical solutions: a review' Professional Geographer 52, 607-623. [Introduces basic sampling issues in a geographical context].

Burrough, P.A. and McDonnell, R.A., (1998). Principles of Geographical Information Systems. Spatial Information Systems and Geostatistics. New York: Oxford University Press. [The standard reference on GIS from a physical geography viewpoint].

Csillag, F., Kertész, M. and Kummert, Á. (1996). Sampling and mapping of heterogeneous surfaces: multi-resolution tiling adjusted to spatial variability. *International Journal of Geographical Information Systems* **10**, 851-875. [A landmark paper on multi-resolution tiling and quadtrees. That is, matching sampling resolution to frequency of spatial variation].

Fotheringham, S., Brunsdon, C. and Charlton, M. (2000). Quantitative Geography. Perspectives on Spatial Analysis. London: Sage. [A wonderfully easy-to-read book on quantitative geography in general].

Goodchild, M.F. and Gopal, S. (1989). Accuracy of Spatial Databases. London: Taylor and Francis. [One of the first books to consider errors in GIS in an explicit way].

Heuvelink, G.B.M. (1998). *Error Propagation in Environmental Modelling with GIS*. London: Taylor and Francis. [An excellent monograph on models for how errors propagate through the GIS cycle].

Heywood, I., Cornelius, S. and Carver, S. (1998). An Introduction to Geographical Information Systems. Longman: Harlow. [An excellent introduction to the basic principles of GIS].

Longley, P.A., Brooks, S.M., McDonnell, R. and MacMillan, W. (editors) (1998). *Geocomputation: a Primer*. Chichester: Wiley. [One of the first volumes on GeoComputation, which emphasises computational analysis].

Masser, I., Campell, H. and Craglia, M. (1996). *GISDATA3. GIS Diffusion. The Adoption and Use of Geographical Information Systems in Local Government in Europe*. London: Taylor and Francis. [This, and several other similar books, detail the extent to which GIS have become commonplace in everyday life].

Openshaw, S. and Abrahart, R. (eds.) (2000). *Geocomputation*. London: Taylor and Francis.

Biographical Sketch

Peter Atkinson is full Professor of Geography at the University of Southampton. Prior to joining the University of Southampton, he spent three years as a post-doctoral researcher at the Department of Geography, University of Bristol. He obtained his B.Sc. degree from the Department of Geography, University of Nottingham in 1986, and his Ph.D. from the Department of Geography, University of Sheffield (NERC CASE award with Rothamsted Experimental Station) in 1990. His main research interests are in spatial statistics and spatial modelling, with particular emphasis on geostatistics, GIS, remote sensing and spatially distributed dynamic modelling. His substantive interests are varied and include biogeography, ecology, epidemiology, soil survey, geomorphology, land surface hydrology, and a range of natural hazards within these (e.g., disease, landsliding, flooding). Peter Atkinson has published numerous refereed journal articles. In addition, he has edited five books on remote sensing or GIS and seven journal special issues. He is joint Editor of *International Journal of Remote Sensing Letters*, and is an Editorial Board member for several journals. He has been a PI on several grants and contracts, and currently sits on numerous international scientific committees.