Richardson DE 1994 Contextual Transformations and Generalizations of Remotely Sensed Imagery for Map Generation. In: Molenaar M, De Hoop S (Eds) Advanced geographic data modelling, Netherlands Geodetic Commission, New Series, Nr. 40, Delft, pp170-178

U.S.BUREAU OF THE CENSUS 1990 Technical description of the DIME System. In: Peuquet DJ, Marble DF (Eds) Introductory readings in GIS, London: Taylor and Francis, pp100-111

3.2 Modern GIS and model linking

Michael F. Goodchild

Department of Geography, and National Center for Geographic Information and Analysis
University of California Santa Barbara, CA 93106-4060, USA
Email: good@ncgia.ucsb.edu

The traditional approach to analysis sees it as part of a larger scheme that begins with problem formulation and ends with interpretation of results. Techniques of spatial analysis form a well-defined subset of the larger set of analytic methods, defined by an invariance property. For many reasons this view of spatial analysis, and the larger field of analysis in general, is undergoing profound change, brought on in part by the advent of integrated computing environments such as geographic information systems (GIS). The paper reviews the trends contributing to this change, and its possible effects on the role of spatial analysis, and the broader context of GIS, in the future.

1. Introduction

GIS and spatial analysis have enjoyed a long and productive relationship over the past decades (for reviews see Fotheringham and Rogerson 1994; Goodchild 1988; Goodchild et al. 1992). GIS has been seen as the key to implementing methods of spatial analysis, making them more accessible to a broader range of users, and hopefully more widely used in making effective decisions and in supporting scientific research. It has been argued (e.g. Goodchild 1988) that in this sense the relationship between spatial analysis and GIS is analogous to that between statistics and the statistical packages. Much has been written about the need to extend the range of spatial analytic functions available in GIS, and about the competition for the attention of GIS developers between spatial analysis and other GIS uses, many of which are more powerful and better able to command funding. Specialized GIS packages directed specifically at spatial analysis have emerged (e.g. IDRISI, and see Bailey and Gatrell 1995). Finally, implementation of spatial analysis methods in GIS is leading to a new, exploratory emphasis.

The purpose of this paper is to explore new directions that have emerged recently, or are currently emerging, in the general area of GIS and spatial analysis, with particular emphasis on the practical issues that arise in making use of today's capabilities for spatial analysis in GIS. In the next section, it is argued that in the past GIS and spatial analysis have followed a very clearly and narrowly defined path, one that has more to do with the world of spatial analysis prior to the advent of GIS than with making the most of both fields—the path is, in other words, a legacy of prior conditions and an earlier era. The following section identifies a number of trends, some related to GIS but some much more broadly-based, that have changed the context of GIS and spatial analysis over the past few years, and continue to do so at an increasing rate. The third section identifies some of the consequences of these trends, and the problems that are arising in the development of a new approach to spatial analysis. The paper concludes with some comments about the complexity of the interactions between analysis, data and tools, and speculation on what the future may hold, and what forms of spatial analysis it is likely to favor. A further elaboration of these ideas will appear in a forthcoming chapter written jointly with Paul Longley (Goodchild and Longley 1997).

2. Traditions in Spatial Analysis

2.1 The Linear Project Design

In the best of all possible worlds, a research project (the term 'research' will be interpreted very broadly to include both scientific and decision-making activities) begins with a clearly stated problem. Some decision must be made, some question of scientific theory resolved by resorting to experiment or real-world evidence. An experimental design is developed to resolve the problem, data are collected, analyses are performed, and the results are interpreted and reported. This simple structure has underlain generations of student dissertations, government reports, and research papers. The sequence is strictly linear, implying that the availability of data has no influence on problem definition; availability or awareness of methods of analysis no influence on collection of data; etc. Indeed, the terms 'data-driven' and 'technique-driven' are highly perjorative in research generally, as are such phrases as 'a technique in search of a problem'—in this ideal world, the statement of the problem strictly precedes the collection of data and the performance of analysis.

In this simple, sequential world the selection of methods of analysis can be reduced to a few simple rules (in the context of statistical analysis, see for example Levine 1981: Ch 17; Marascuilo and Levin 1983: inside cover; Siegel 1956: inside cover). Choice of analytic method depends on the type of decision to be made (e.g. whether two samples are drawn from the same, unknown population, or whether two variables are correlated), and on the characteristics of the available data (e.g. scale of measurement—nominal, ordinal, interval or ratio; but see Chrisman 1997 for a discussion of this simple four-way classification in the context of GIS).

2.2 Spatial analysis

Spatial analysis, or spatial data analysis, is a well-defined subset of the methods of analysis available to a project. One might define spatial analysis as a set of methods useful when the data are spatial, in other words when the data are referenced to a two-dimensional frame. More narrowly, the Earth's surface provides a particular instance of such a frame, the geographic frame, with its peculiar properties of curvature. This definition of spatial analysis is arguably too broad, because in basing the definition on the properties of data it does not address the question of whether the two-dimensional frame actually matters—could the same results have been obtained if the frame were distorted in some way, or if objects were repositioned in the frame? More precisely, then, spatial analysis can be defined as that subset of analytic techniques whose results depend on the frame, or will change if the frame changes, or if objects are repositioned within it. To distinguish analytic methods from more mundane operations they might be defined as methods for processing data with the objective of solving some scientific or decision-making problem.

Methods of spatial analysis have accumulated in a literature that spans many decades, indeed centuries. They have been invented in many disciplines, including mathematics, and particularly geometry; statistics, and particularly spatial statistics and statistical geometry; and in geography and other Earth sciences. Compendia have been published (among others, see Bailey and Gatrell 1995; Berry and Marble 1968; Haining 1990; Taylor 1977; Unwin 1981), and various approaches proposed for structuring this body of technique. Many of the earlier methods could be described as confirmatory, mirroring the hypothesis-testing tradition of statistics in seeking to confirm or deny some formally stated hypothesis through the analysis of empirical data. Others are better described as exploratory, subjecting data to manipulations selected for their ability to expose patterns and anomalies that might not otherwise be evident to the analyst, or manipulating the data in ways designed to enhance the investigator's intuition.

2.3 The well-informed analyst

Traditionally, the responsibilities of the inventor of a technique ended when the technique had been tested and described. Even the testing of a technique can be suspect in an academic world that often values theory over empiricism, and is suspicious of empirical results that cannot be demonstrated to be generally true. The advent of the digital computer changed this world fundamentally, because it became possible for a scientist to perform a method of analysis automatically, without taking personal responsibility for every aspect of the performance. It was now possible using the 'black box' of the computer to perform an analysis that one did not know everything about—that one could not perform by hand. Methods emerged, beginning in the 1970s and particularly in the area of multivariate statistics, that would be impossibly impractical to perform by hand. Pedagogically, a fundamental shift became possible in how analysis was taught—that one might learn about a technique by studying the nature of its response to particular inputs, rather than by studying how the response was generated. But there is a fundamental difference between these two positions: between whether one understands the results of a principal components analysis, for example, as the extraction of eigenvalues from a specific matrix, or the generation of statistics that broadly indicate some concept of 'relative importance'.

Exactly where this change occurred is open to debate, of course. It may have occurred when students were no longer required to perform statistical analyses by hand before being let loose on computer packages; or when Fortran appeared, making it necessary to understand less about how instructions were actually carried out; or when the growth of the scientific enterprise had reached such a level that potential replication of every result was a practical impossibility.

In the early days of statistical analysis all calculations had to be carried out by hand. Although the intensity of the necessary calculations must clearly have had some influence on the choice of method, in principle the paradigm had no way of including this factor as a criterion that could affect the choice of method. In this somewhat monastic world the cost of the scientist's labor was simply not a factor in his or her science. Highly routine tasks could be assigned to an apparently inexhaustible supply of unpaid or very cheap student labour. Thus the intense numerical calculations needed in the early days of factor analysis seem to have had surprisingly little negative impact on the development or adoption of the technique (Harman 1976).

Of course the digital computers that were introduced to the scientific community beginning in the late 1950s produced rapid change in the labour demands of many statistical methods. The intricate calculations of factor analysis could be performed by a fully automatic machine, provided the researcher could command sufficient computer time, and provided labour was available to punch the necessary cards. Computers and the brains of young children are similar in many ways; both begin essentially empty; both must acquire the primitive elements of reasoning; but having done so, both can build enormously complex structures out of simpler ones, apparently ad infinitum. What began in the 1960s as a set of uncoordinated efforts by individual scientists writing their own programs had developed by the 1990s into a complex of enormously sophisticated tools, each integrating a large number of methods into an easy-to-use whole.

2.4 Extending the functions of analytic software

Although they show clear evidence of their roots, the packages used by the scientists of the 1990s are different in fundamental respects from the programs of the 1960s. Besides implementing large numbers of statistical methods, today's packages also provide support for the creation and maintenance of data. There will be tools for documenting data sets, and describing their properties, such as accuracy and history. Other tools will support the sharing of data, in the form of format converters or interfaces to the Internet. In short, the functions of today's digital computers in

supporting research go far beyond those of a simple calculating machine, carrying out well-defined methods of analysis. The same digital computer may now be involved in the selection and formulation of a problem, by providing access to automated library catalogs and on-line literature; in the collection of data through support for real-time data acquisition; in management of data, performance of analysis, visualization of results, writing of conclusions; and even in publication through access to the Internet and the World Wide Web. The computer is no longer part of the research environment—we are rapidly approaching a world in which the computer *is* the research environment.

These trends are echoed strongly in geographical information systems. Although a particular scientist might use a GIS in ways that are more analogous to the early days of statistical computing, by performing a single buffering operation, for example, scientific applications are much more likely to include integration of many GIS functions. Today's scientist or decision-maker is likely to see a GIS as an environment for research, rather than as a means of automating analysis. The GIS is likely to be involved in the project from beginning to end, and to be integrated with other tools and environments when these are needed. GIS will be used for collecting, assembling, verifying and editing the data; performing some of the analyses required by the project; and presenting and interpreting the results. Moreover, much GIS use may not be tied to a specific project—GIS finds extensive use in the collection of data for purposes that may be generic, or not well-defined, or may be justified in anticipation of future demand. Even though these may not be projects in the sense of the earlier discussion, analysis may still be necessary as part of the data production process—for example, when a soil scientist must analyze data to produce a soil map.

2.5 When to choose GIS

If GIS has multiple roles in support of science and problem-solving, then one might not be surprised to find that the choice between GIS alternatives is complex and often daunting. The many GIS packages offer a wide range of combinations of analysis functions, housekeeping support, alternative ways of representing the same phenomena, different levels of sophistication in visual display, and performance. In addition, choice is often driven by the available hardware, since not all GIS run on all platforms; on the format in which the necessary data has been supplied, the personal preferences and background of the user, and so forth. Even the extensive and frequently updated comparative surveys published by groups such as GIS World Inc can be of little help to the uninitiated user.

The existence of other classes of analytic software complicates the scene even more. Under what circumstances is a problem better solved using a package that identifies itself as a GIS, or using a statistical package, or a mathematical package, or a scientific visualization package? Under what circumstances is it better to fit the square peg of a real problem into the round GIS hole? GIS are distinguished by their ability to handle data referenced to a two-dimensional frame, but such capabilities also exist to a more limited extent in many other types of software environment. For example, it is possible to store a map in a spreadsheet array, and with a little ingenuity to produce a passable 'map' output; and many statistical packages support data in the form of images.

Under what circumstances, then, is an analyst likely to choose a GIS? The following conditions are suggested, although the list is certainly not complete, and the items are not intended to be mutually exclusive:

- when the data are geographically referenced, and when geographical referencing is essential to the analysis (see earlier discussion of the definition of spatial analysis);
- when the data include a range of vector data types (support for vector analysis among non-GIS packages appears to be much less common than support for raster analysis);
- when topology—representation of the connections between objects—is important to the analysis;

- when the curvature of the Earth's surface is important to the analysis, requiring support for projections and for methods of spatial analysis on curved surfaces;
- when the volume of data is large, since alternatives like spreadsheets tend to work only for small data sets;
- when data must be integrated from a variety of sources, requiring extensive support for reformatting, resampling, and other forms of format change;
- when geographical objects under analysis have large numbers of attributes, requiring support from integrated database management systems, since many alternatives lack such integration;
- when the background of the investigator is in geography, or a discipline with strong interest in geographical data;
- when the project involves several disciplines, and must therefore transcend the software traditions and preferences of each;
- when visual display is important, and when the results must be presented to varied audiences;
- when the results of the analysis are likely to be used as input by other projects, or when the data are being extensively shared.

3. Elements of a New Perspective

This section reviews some of the changes that are altering the context and face of spatial analysis using GIS. Some are driven by technological change, and others by larger trends affecting society as we approach the millennium.

3.1 The costs of data creation

The collection of geographical data can be extremely labor-intensive. Early topographic mapping required the map-maker to walk large parts of the ground being mapped; soil mapping requires the exhausting work of digging soil pits, followed often by laborious chemical analysis; census data collection requires repeated visits to a substantial proportion of all households; and forest mapping requires 'operational cruise', the intensive observation of conditions along transects. Although many new methods of geographical data creation have replaced the human observer on the ground with various forms of automated sensing, there is no alternative in those areas that require the presence of expert interpreters in the field.

Many of the remaining stages of geographical data creation are also highly labor-intensive. There is still no alternative to manual digitizing in cases where the source document is complex, compromised, or difficult to interpret. The processes of error detection and correction are difficult if not impossible to automate, and the methods of cartographic generalization used by expert cartographers have proven very difficult to formalize and replace. In short, despite much technical progress over the past few decades, geographical data creation remains an expensive process that is far from fully automated.

Labor costs continue to rise at a time when the resources available to government, the traditional source of geographical data, continue to shrink. Many geographical data sets are collected for purposes which may be far from immediate, and it is difficult therefore to convince taxpayers that they represent an essential investment of public funds, especially in peacetime. Governments in financial straits call for evidence of need, and many have moved their mapping operations onto a semi-commercial basis in order to allow demand to be expressed through willingness to pay. To date, the U.S. Federal mapping agencies have resisted the trend, but internationally there is more and more evidence of the emergence of a market in geographical information.

Within the domain of geographical data the pressures of increased labor costs favor data that can be collected and processed automatically. Given a choice between the labor-intensive production of vector topographic data, and the semi-automated generation of such raster products as digital elevation models and digital orthophotos, economic pressures can lead only in one direction. It is easy to imagine a user trading off the ability to identify features by name against the order of magnitude lower cost, and thus greater potential update frequency, of raster data.

Of course, the principle of information commerce is alien to the scientific community, which is likely to resist strongly any attempt to charge for data that is of interest to science, even peripherally. But here too there are pressures to make better use of the resources invested in scientific data collection. Research funding agencies now increasingly require evidence that data collected for a project have been disseminated, or made accessible to others, while recognizing the need to protect the interests of the collector.

But trends such as these, while they may be eminently rational to dispensers of public funds, nevertheless fly directly in the face of the traditional model of science presented earlier. How can projects fail to be driven by data, if data are forced to obey the economic laws of supply and demand? Where in traditional science are the rules and standards that allow scientists to trade off economic cost against scientific truth? It seems that economic necessity has forced the practice of science to move well beyond the traditions that are reflected in accepted scientific methodologies and philosophies of science.

3.2 The life of a data set

In the traditional model presented earlier data were collected or created to solve a particular problem, and had no use afterwards except perhaps to historians of science. But many types of geographical data are collected and maintained for generic purposes, and may be used many times by completely unrelated projects. For other types, the creation of data is itself a form of science, involving the field skills of a soil scientist, for example, or a biologist. Thus a data set can be simultaneously the output of one person's science, and the input to another's. These relationships have become further complicated by the rise of multidisciplinary science, which combines the strengths and expertise of many different sciences, and partitions the work among them. Once again, the linear model of science is in trouble, unable to reflect the complex relationships between projects, data sets, and analytic techniques that exist in modern science. The notion that data are somehow subsidiary to problems, methods and results is challenged, and traditional dicta about not including technical detail in scientific reports may be counterproductive.

In this new world a given set of data is likely to fall into many different hands during its life. It may be assembled from a mixture of field and remote sensing sources, interpreted by a specialist, cataloged by an archivist or librarian, used by scientists and problem-solvers, and passed between its custodians using a range of technologies. It is quite possible in today's world that the various creators and users share little in the way of common disciplinary background, leaving the data set open to misunderstanding and misinterpretation. Recent interest in metadata, or ways of describing the contents of data sets, is directed at reducing some of these problems, but the easy access to data provided by the Internet and various geographical data archives has tended to make the problem worse.

These issues are particularly prominent in the case of data quality, and the ability of the user of a data set to understand its limitations, and the uncertainty that exists about the real phenomena the data are intended to represent. To take a simple example, suppose information on the geodetic datum underlying a particular data set—potentially a very significant component of its metadata—was lost in transmission between source and user; or alternatively suppose that the user simply assumed the wrong datum, or was unaware of its significance. This loss of metadata, or specification

of the data content, is equivalent in every respect to an actual loss of accuracy equal to the difference between the true datum and the datum assumed by the user, which can be several hundreds of meters. In short, the quality of a data set to a user is a function of the difference between its contents and the *user's* understanding of its meaning, not the creator's.

3.3 Data sharing

In this new world of shared data the term *metadata* has come to function as the equivalent of documentation, catalog, handling instructions, and production control. The U.S. Federal Geographic Data Committee's Content Standards for Digital Geospatial Metadata (FGDC 1994) have been very influential in providing a standard, and have been emulated frequently. If the custodian of a large collection of geographical data sets provides metadata in this form, it is possible for others to search its records for those that match their needs. The FGDC's National Geospatial Data Clearinghouse (http://www.fgdc.gov) is one such directory (and see also the Alexandria Digital Library project to provide distributed library services for geographically referenced data sets, Smith et al. 1996; and see http://alexandria.ucsb.edu).

The user of a traditional library will rarely know the exact subject of a search—instead, library search has an essential fuzziness, which is supported by the traditional library in several essential ways. By assigning similar call numbers to books on similar subjects, and shelving by call number, the traditional library is able to provide an environment that allows the user to browse the collection in a chosen area. But this support is missing when the records of a metadata file are searched using simple Boolean methods. It would make better sense to model the search process as one of finding the best fit between a metadata record representing the user's ideal, and metadata records representing the data sets available. It is very unlikely, after all, that data exist that perfectly match the needs of a given problem, especially in the ideal world of problem-solving represented earlier.

3.4 New techniques for analysis

Many new methods of spatial analysis have emerged in the rich computational environment now available to scientists. These include neural nets, new methods of optimization such as simulated annealing and genetic techniques, and computationally intensive simulation. The term *geocomputation* has been suggested. Methods of exploratory spatial data analysis have extended the principles of exploratory data analysis (Tukey 1977) to spatial data.

In science generally, the combination of vast new sources of data and high-speed computation have led to an interest in methods of *data mining*, which implies the ability to *dredge* data at very high speed in a search for patterns of scientific interest. In a geographical context, the very vague notion of 'scientific interest' might suggest the need for methods to detect features or measurements that are inconsistent with their surroundings, in apparent violation of Tobler's 'first law of geography' (Tobler 1970). Linearities in images are of potential interest in geological prospecting; and one can imagine circumstances in which atmospheric scientists might want to search large numbers of images for patterns consistent with weather events. Such techniques of pattern recognition were pioneered many years ago in particle physics, to search vast numbers of bubble chamber photographs for the tracks characteristic of rare new particles.

One might argue that such techniques represent a renewal of interest in inductive science—the search for regularities or patterns in the world that would then stimulate new explanatory theories. Inductivism has fallen out of fashion in recent decades, at least in disciplines that focus on geographical data, leading one to ask whether a renewal of interest represents a fundamental shift in science, or merely a response to the opportunities offered by more powerful technology. On thi

issue the jury is clearly still 'out'—geocomputation has not yet provided the kinds of new insights that might support a broad shift to inductivism.

3.5 New computer architectures

The communication technologies that have emerged in the past decade have allowed a fundamental change in the architecture of computing systems. Instead of the early mainframes and later stand-alone desktop systems, today's computers are linked with high-speed networks that allow data, software, and storage capacity located in widely scattered systems to be integrated into functioning wholes. Data can now be 'served' from central sites on demand, avoiding the need to disseminate many copies, with subsequent confusion when updates are needed.

The new approaches to computing that are possible in this interconnected environment are having a profound effect on spatial analysis. Because it is no longer possible to assume a lifetime association between a user and a particular system design, there are mounting pressures for standards and interoperability between systems to counter the high costs of retraining of staff and reformatting of data.

The proprietary GIS that once dominated the industry attempted to provide a full range of GIS services in one homogeneous environment. Data were stored in proprietary formats, often kept secret by vendors to maintain market position, but making it difficult for others to expand the capabilities of the system by programming extra modules. The 'open GIS' movement (Buehler and McKee 1996; and see http://www.ogis.org) mirrors efforts in other areas of the electronic data processing world to promote interoperability, open standards and formats, and easy exchange from one system to another. While such ideas were often regarded as counter to the commercial interests of vendors, there is now widespread acceptance in the industry that they represent the way of the future.

The implications of open systems for spatial analysis are likely to be profound. First, they offer the potential of a uniform working environment, in which knowledge of one system is readily transferrable to another. To make this work, however, it will be necessary to achieve a uniform view, and its acceptance across a very heterogeneous user community. There is no prospect of interoperability and open systems without agreement on the fundamental data models, terminology and objectives of GIS-based analysis. Thus much effort will be needed on the part of the inventors and implementors of spatial analysis to develop this uniform view.

Second, the possibility of easy sharing of data across systems gives even greater momentum to efforts to make geographical information more shareable, and even greater demands on the existence and effectiveness of metadata.

Third, interoperability is likely to create an environment in which it is much easier to implement methods of spatial analysis in GIS. Traditionally, vendors of monolithic systems have added functions when market demand appears to justify the development costs. It has been impossible, in a world of proprietary systems, for third parties to add significant functionality. Thus expansion of spatial analytic capabilities has been slow, and has tended to reflect the needs of the commercial market, rather than those of science and problem-solving, when these diverge. In a world of open systems it will be much easier to add functions, and the new environment will encourage the emergence of small companies offering specialized functionality in niche markets.

Finally, new interoperable approaches to software will encourage the modularization of code. It is already possible in some mainstream software environments to launch one specialized application within another—for example, to apply spreadsheet functions to information in a word processing package. This 'plug and play' environment offers enormous scope to GIS, since it will lead ultimately to a greater integration of GIS functions, and map and imagery data in general, into mainstream electronic data processing applications.

The scientific world has grown used to a more or less complete separation between data, and the functions that operate on and manipulate data. Functions are part of 'analysis', which plays a role in the traditional approach to problem-solving outlined earlier that is clearly distinct from that of data. But it has already been argued that in a world of extensive data sharing and interaction between disciplines it is impossible to think of data in isolation from its description, or metadata, which allows the meaning of information to be shared.

In the abstract world of object-oriented methods it is argued that the meaning of data lies ultimately in the operations that can be performed. If data sets exist in two systems, and pairs of functions exist in both systems that produce the same answers, then the two data sets are the same in information content, irrespective of their specific formats and arrangements of bits. It makes sense, then, to *encapsulate* methods with data. When more than one method is available to perform a given function, it makes sense for the choice to be made by the person best able to do so, and for the method thereafter to travel with the data. For example, a climatologist might encapsulate an appropriate method for spatial interpolation with a set of point weather records, because the climatologist is arguably better able to select the best method of spatial interpolation, given his or her knowledge of atmospheric processes.

In future, and especially given the current trend in computing to object-oriented methods, it is likely that the distinction between data and methods will become increasingly blurred. Commonly used techniques of spatial analysis, such as spatial interpolation, may become encapsulated with data in an extension of the concept of metadata to include methods. Of course this assumes that methods are capable of running in a wide variety of host systems, which takes the discussion back to the issue of interoperability introduced earlier.

4. Spatial Analysis in Practice

At this stage, it seems useful to introduce a discussion of the practical problems which face the users of today's GIS. While it is now possible to undertake a wide range of forms of spatial analysis, and to integrate data from a range of sources that would have seemed inconceivable as little as five years ago, there continue to be abundant limitations that impede the complete fulfilment of the technology's promise. The following subsections discuss several of these current impediments.

4.1 Absolute and relative position

First, and perhaps foremost, are problems of varying data quality. In science generally it is common to express quality in terms such as 'accurate to plus or minus one degree'. But while such methods are useful for many types of data, they are much less so when the data are geographical. The individual items of information in a geographical data set are typically the result of a long and complex series of processing and interpretation steps, and bear little relationship to the independent measurements of traditional error analysis. The following discussion is limited to the particular problems encountered when merging data sets.

While projections and geodetic datums are commonly well-documented for the data sets produced by government agencies, the individual scientist digitizing a map may well not be in a position identify either. The idea that lack of specification could contribute to uncertainty was discussed earlier, and its effects will be immediately apparent if a data set is merged with one based on another projection or datum. In practice, therefore, users of GIS frequently encounter the need for methods of *conflation*, a topic discussed in detail below.

The individual items of information in a geographical data set often share lineage, in the sense that more than one item is affected by the same error. This happens, for example, when a map or photograph is registered poorly—all of the data derived from it will have the same error. One indicator of shared lineage, then, is the persistence of error—all points derived from or dependent on the same misregistration will be displaced by the same or a similar amount. Because neighboring points are more likely to share lineage than distant points, errors tend to show strong positive spatial autocorrelation (Goodchild and Gopal 1989).

Rubber-sheeting is the term used to describe methods for removing such errors on the assumption that strong spatial autocorrelations exist. If errors tend to be spatially autocorrelated up to a distance of x, say, then rubber-sheeting will be successful at removing them, at least partially, provided control points can be found that are spaced less than x apart. For the same reason, the shapes of features that are less than x across will tend to have little distortion, while very large shapes may be badly distorted. The results of calculating areas, or other geometric operations that rely only on relative position, will be accurate as long as the areas are small, but will grow rapidly with feature size. Thus it is important for the user of a GIS to know which operations depend on *relative* position, and over what distance; and where *absolute* position is important (of course the term *absolute* simply means relative to the Earth frame, defined by the Equator and the Greenwich meridian, or relative over a very long distance).

When two data sets are merged that share no common lineage (for example, they have not been subject to the same misregistration), then the relative positions of objects inherit the absolute positional errors of both, even over the shortest distances. While the shapes of objects in each data set may be accurate, the relative locations of pairs of neighboring objects may be wildly inaccurate when drawn from different data sets. The anecdotal history of GIS is full of such examples—data sets which were perfectly adequate for one application, but failed completely when an application required that they be merged with some new data set that had no common lineage. For example, merging GPS measurements of point positions with streets derived from the U.S. Bureau of the Census TIGER files may lead to surprises where points appear on the wrong sides of streets. If the absolute positional accuracy of a data set is 50m, as it is with parts of TIGER, then such surprises will be common for points located less than 50m from the nearest street.

4.2 Semantic integration

Some of the most challenging problems in GIS practice occur in the area of semantic integration, where integration relies on an understanding of meaning. Such problems can occur between geographic jurisdictions, if definitions of feature types, or classifications, or methods of measurement vary between them. It is common, for example, for schemes of vegetation classification to vary from one country to another, making it difficult to produce horizontally merged data (Mounsey 1991). 'Vertical' integration can also be problematic, as in the problems of merging maps produced of the same area by different agencies.

While some of these problems may disappear with more enlightened standards, others are eminently reasonable. The problems of management of ecosystems in Florida are clearly different from those of Montana, and it is reasonable that standards adopted by the two states should be different. Even if it were possible to standardize for the entire U.S., one would be no further ahead in standardizing between the U.S. and other countries. Instead, it seems a more reasonable approach is to achieve interoperability without standardization, by more intelligent approaches to system design.

4.3 Conflation

Conflation appears to be the term of choice in the GIS community for functions that attempt to overcome differences between data sets, or to merge their contents. Conflation attempts to replace two or more versions of the same information with a single version that reflects the pooling of the sources; it may help to think of it as a process of weighted averaging. The complementary term *concatenation* refers to the integration of the sources, so that the contents of both are accessible in the product. The polygon overlay operation familiar to many GIS users is thus a form of concatenation.

Two distinct forms of conflation can be identified, depending on the context: (1) conflation of feature geometry and topology, and concatenation of feature attributes; and (2) conflation of geometry, topology and attributes. As an example of the first case, suppose information is available on the railroad network at two scales, 1:100,000 and 1:2 million. The set of attributes available is richer at the 1:2 million scale, but the geometry and topology are more accurate at 1:100,000. Thus it would be desirable to combine the two, discarding the coarser geometry and topology. As an example of the second case, consider a situation in which soils have been mapped for two adjacent counties, by two different teams of scientists. At the common border there is an obvious problem, because although the county boundary was defined by a process that was in no way dependent on soils, the border nevertheless appears in the combined map. Thus it would be desirable to 'average' the data at and near the boundary by combining the information from both maps in compatible fashion. As these two examples illustrate, the need for conflation occurs both horizontally, in the form of edgematching, and 'vertically'.

4.4 Perfect positioning

It is easy to imagine that the need for conflation and for discussions of relative and absolute positional accuracy will eventually go away, as positioning becomes more and more accurate, leading eventually to 'perfect' positioning. Unfortunately there are good reasons why that happy state will never be reached. Although the positions of the Greenwich meridian and various geodetic control points have been established by fixing monuments, seismic motions, continental drift, and the wobbling of the Earth's axis all lead to fundamental uncertainty in position. Any mathematical representation of the Earth's shape must be an approximation, and different approximations have been adopted for different purposes. Moreover, there will always be a legacy of earlier, less accurate measurements to deal with. Thus it seems GIS will always have to deal with uncertainty of position, and with the distinctions between relative and absolute accuracy, and their complex implications for analysis.

Instead, strategies must be found for overcoming the inevitable differences between databases, either prior to analysis or in some cases 'on the fly'. Consider, for example, the problems caused by use of different map databases for vehicle routing. Systems are already available on an experimental basis that broadcast information on street congestion and road maintenance to vehicles equipped with map databases and systems to display such information for the driver. In a world of many competing vendors such systems will have to overcome problems of mismatch between different databases, in terms both of position and of attributes. For example, two databases may disagree over the exact location of 100 Main St, or whether there *is* a 100 Main St, with potentially disastrous consequences for emergency vehicles, and expensive consequences for deliveries. Recent trends suggest that the prospects for central standardization of street naming by a single authority are diminishing, rather than growing.

5. Conclusion

The prospects for spatial analysis have never been better. Data are available in unprecedented volume, and easily accessed over today's communication networks. More methods of spatial analysis are implemented in today's GIS than ever before, and GIS has made methods of analysis that were previously locked in obscure journals easy and straightforward to use. Nevertheless, today's environment for spatial analysis raises many issues, not the least of which is the ability of users to understand and to interpret correctly. Questions are being raised about the deeper implications of spatial analysis, and the development of databases that verge on invasion of individual. And our expectations may be unreasonable given the inevitable problems of spatial data quality.

Technological developments have further muddied the methodological waters, by confusing what was once a simple linear sequence of problem formulation, data collection, analysis, and conclusion. It seems clear that tomorrow's science will be increasingly driven by complex interactions, as data become increasingly commodified, technology increasingly indispensible to science, and conclusions increasingly consensual. New philosophies of science that reflect today's realities are already overdue.

If science and problem-solving are to be constrained by these new realities, then what kinds of spatial analysis are most likely to dominate in the coming years? The points raised in this chapter's discussion suggest that the future environment will favor the following:

- data whose meanings are widely understood, making it easier for multidisciplinary teams to collaborate;
- data with widespread use, generating demands that can justify the costs of creation;
- data with commercial as well as scientific and problem-solving value, allowing costs to be shared across many sectors;
- methods of analysis with commercial application, making it more likely that such methods will be implemented in widely available form;
- methods implemented using general standards, allowing them to be linked to other methods using common standards and protocols.

6. References

Bailey TC, Gatrell AC 1995 Interactive Spatial Data Analysis. New York: Wiley.

Berry BJL, Marble DF (eds) 1968 Spatial Analysis: A Reader in Statistical Geography. Englewood Cliffs, NJ: Prentice-Hall.

Buehler K, McKee L (eds) 1996 The OpenGIS Guide. Wayland, MA: The Open GIS Consortium Inc

Chrisman NR 1997 Exploring Geographic Information Systems. New York: Wiley.

Federal Geographic Data Committee 1994 Content Standards for Digital Geospatial Metadata. Washington, DC: Federal Geographic Data Committee, Department of the Interior. http://www.fgdc.gov.

Fotheringham AS, Rogerson PA (eds) 1994 Spatial Analysis and GIS. London: Taylor and Francis.

Goodchild MF 1988 A spatial analytic perspective on geographical information systems. International Journal of Geographical Information Systems 1: 327–334.

Goodchild MF, Gopal S 1989 Accuracy of Spatial Databases. London: Taylor and Francis.

Goodchild MF, Haining RP, Wise S and 12 others 1992 Integrating GIS and spatial analysis: problems and possibilities. International Journal of Geographical Information Systems 6(5): 407–423.

Goodchild MF, Longley P 1997 The future of GIS and spatial analysis. In Longley P, Goodchild MF, Maguire DJ, Rhind DW (eds) Geographical Information Systems: Principles, Techniques, Management, and Applications. Cambridge: Geoinformation International.

Haining RP 1990 Spatial Data Analysis in the Social and Environmental Sciences. New York: Cambridge University Press.

Harman HH 1976 Modern Factor Analysis. Chicago: University of Chicago Press.

Levine G 1981 Introductory Statistics for Psychology: The Logic and the Methods. New York: Academic Press.

Marascuilo LA, Levin JR 1983 Multivariate Statistics in the Social Sciences: A Researcher's Guide. Monterey, CA: Brooks/Cole.

Mounsey H 1991 Multisource, multinational environmental GIS: lessons learned from CORINE. In Maguire DJ, Goodchild MF, Rhind DW (eds) Geographical information systems: principles and applications. Harlow: Longman Scientific and Technical, 2: 185–200.

Siegel S 1956 Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill.

Smith TR, Andresen D, Carver L, Dolin R, and others 1996 A digital library for geographically referenced materials. Computer 29(5): 54, 29(7):14.

Taylor PJ 1977 Quantitative Methods in Geography: An Introduction to Spatial Analysis. Boston: Houghton Mifflin.

Tobler WR 1970 A computer movie simulating urban growth in the Detroit region. Economic Geography supplement 46: 234–240.

Tukey JW 1970 Exploratory Data Analysis. Reading, MA: Addison Wesley.

Unwin DJ 1981 Introductory Spatial Analysis. London: Methuen.