TITLE: SPATIAL DATA ANALYSIS

BYLINE: Michael F. Goodchild, University of California, Santa Barbara, www.geog.ucsb.edu/~good

SYNONYMS: spatial analysis, geographical data analysis, geographical analysis

DEFINITION:
Methods of data analysis perform logical or mathematical manipulations on data in order to test hypotheses, expose anomalies or patterns, or create summaries or views that expose particular traits. Data often refer to specific locations in some space. To qualify as spatial, the locations must be known and must affect the outcome of the analysis. While many spaces might be relevant, including the space of the human brain or the space of the human genome, the history of spatial data analysis is dominated by location in geographic space, in other words location on or near the surface of the Earth. Thus geographical and spatial are often essentially synonymous. More formally, spatial data analysis can be defined as a set of techniques devised for the manipulation of data whose outcomes are not invariant under relocation of the objects of interest in some space. The term *exploratory spatial data analysis* (ESDA) describes an important subset that emphasizes real-time interaction, the creation of multiple views of data, the search for patterns and anomalies, and the generation of new hypotheses as opposed to the formal testing of existing ideas. The term *spatial data mining* describes another important subset that emphasizes the analysis of very large volumes of spatial data.

HISTORICAL BACKGROUND:
Berry and Marble [1] made one of the earliest efforts to assemble a systematic review of methods of spatial data analysis, drawing on a literature that had accumulated for many decades. Their interest was sparked in large part by what later became known as the Quantitative Revolution in Geography, a paradigm shift that originated at the University of Washington in the late 1950s and spread rapidly as the original group of graduate students found faculty positions. Bunge [2] summarized the core concept: that the analysis of patterns of phenomena on the Earth's surface could lead to a set of formal theories about the behavior of human and natural systems, and that the discovery of such theories would put the discipline of geography on a sound scientific footing. Substantial progress was made in the 1960s, particularly in the study of patterns of settlement and economic activity, and in the study of such physical phenomena as meandering rivers and stream channel networks.

Beginning in the 1960s, the development of geographic information systems (GIS) provided a major impetus by creating a simple structure in which methods of spatial data analysis could be implemented. By the 1980s GIS had become a popular and rapidly growing software application, with a flourishing industry and tools to enable spatial data analysis, along with the necessary techniques for data acquisition, editing, and display. Today GIS is often portrayed as an engine for spatial data analysis, and many new techniques have been added to what are now literally thousands of methods. GIS finds application in virtually all disciplines that deal with the surface and near-surface of the

Earth, ranging from ecology and geology to sociology and political science [3]. It is extensively used in logistics, in planning and public decision making, in military and intelligence applications, and in the management of utility networks.

While the use of computers to perform spatial data analysis was already well established in the 1960s, ESDA emerged rather later, when the graphics and interactive capabilities of computers had advanced sufficiently. By the early 1990s researchers were developing novel ways of linking multiple views using the windowing techniques that emerged at that time, and exploiting the high-resolution graphics that became available on standard personal computers. Today, interactive tools inspired by ESDA are widely available in GIS products, and more specialized software is also available (see, for example, GeoDa, http://geoda.uiuc.edu).

Interest in spatial data mining has grown in the past decade, driven in part by the increasing availability of very large volumes of spatial data. For example, it is now routine to capture the location and time of use of credit and debit cards, and to apply sophisticated algorithms in an effort to detect fraudulent use. Heavy use of spatial data analysis is made by intelligence agencies, based on software that can examine telephone and email traffic and detect references to places.

SCIENTIFIC FUNDAMENTALS:
Several approaches have been devised for organizing the thousands of techniques that qualify as spatial data analysis. Perhaps the commonest, represented by several recent textbooks and by the organization of some GIS user interfaces, is based on a taxonomy of spatial data types. Very broadly, one can capture variation within a space using either raster or vector structures; a raster structure is created by dividing the space into discrete, regularly shaped elements and describing the contents of each, while vector structures describe each feature present in the space as either a point, line, area, or volume, with associated attributes.

Tomlin [4] and others have systematized the analysis of raster data in schemata described as map algebras, image algebras, or cartographic modeling, and several GIS have adopted these schemata in their user interfaces. In one such schema the analysis of raster data is described as either focal, local, zonal, or global: focal operations are performed independently on the contents of each cell; local operations are performed on a cell and its immediate neighborhood; zonal operations apply to contiguous patches of cells with identical descriptions; and global operations apply to the entire raster.

To date a similarly simple systematization of vector operations has not been achieved. Instead, several textbooks (e.g., [5], [6]) organize methods of vector-based analysis according to the types of features being analyzed, focusing in turn on points, lines, areas, and volumes. For example, techniques for the analysis of sets of points might determine the degree of dispersion of the points; search for anomalous clusters; or find a shortest tour through the points. Some texts also provide descriptions of methods for the analysis of relationships or interactions between features. For example, retailers and traffic engineers commonly use methods of spatial data analysis to predict the numbers of trips

expected between home neighborhood areas and such destination points as shopping centers or places of work.

Longley et al. [3] use a different organizing scheme that is designed to be more strongly related to user motivation, and to overcome some of the ambiguities inherent in an emphasis on data type. Their scheme assigns techniques to six categories, ordered by increasing conceptual sophistication: query and reasoning, measurement, transformation, descriptive summary, optimization, and hypothesis testing.

Query and reasoning functions rely on the presentation of alternative views to the user. For example, a set of data on average income by US state might be presented as a map, as a table, as a histogram, and as a scatterplot in which average income is graphed against another variable such as percent with more than high-school education. The user gains insight by examining the alternative views, by querying specific values, or by selecting data items in one window and seeing them highlighted in the other windows.

Measurement functions represent one of the earliest motivations for GIS. Manual methods for obtaining measures of such properties as area, length, slope, or shape from maps are notoriously inaccurate, tedious, and time-consuming, whereas it is trivial to obtain them from digital representations. Nevertheless digital representations are only approximations or generalizations of real phenomena, and many estimates exhibit representation-related biases.

Transformation functions obtain new objects, or new properties of those objects. They include many key GIS functions, including buffering (the geometric dilation of points, lines, areas, or volumes), overlay (the computation of intersections between objects), and interpolation (the use of data from sample locations to estimate values at locations where no samples were taken). Figure 1 shows an example of buffering, using half-mile circles around points representing the schools of part of Los Angeles. The example was motivated by proposals to ban registered sex offenders from living within a specified distance of a school.

[Figure 1 about here]

Descriptive summaries include the widest range of spatial data analysis techniques. Standard univariate statistics such as the mean, median, mode, standard deviation, and variance have equivalents in multidimensional spaces. Figure 2 shows the two-dimensional equivalents of the mean and standard deviation applied to the black and white populations of Milwaukee, using data by census tract. A suite of summary statistics have been devised for measuring *spatial dependence*, a key property of many spaces based on the observation that measurements of many properties taken close together tend to be more similar than measurements taken far apart. The fields of *spatial statistics* and *geostatistics* are both based on this property, and provide ways of addressing it explicitly. *Spatial heterogeneity*, or the tendency for the properties of spaces to vary widely from one area to another, is also the subject of many forms of descriptive summary. The rapidly evolving field of *local* or *place-based* summaries (e.g., [7]) addresses the spatial

heterogeneity property directly, arguing that it is more important to determine how the results of spatial data analysis vary from one area to another than to attempt to extract single, global results. Another suite of summary statistics addresses the *fragmentation* of landscapes, with particularly strong applications in ecology.

[Figure 2 about here]

Methods of *optimization* focus on the design of spatial pattern rather than on its analysis. They include methods for optimum location of points (e.g., retail stores, schools), lines (e.g. roads, pipelines), and areas (e.g., political voting districts), as well as on the design of optimum routes (e.g., for delivery vehicles or school buses).

Finally, methods of *hypothesis testing* address the process of inference, by which analysts reason from the analysis of a sample to conclusions about the larger world represented by the sample. Such methods are well known in statistical analysis, encompassing many well-known statistical tests, significance levels, and the formulation of null hypotheses. Unfortunately the application of such methods to spatial data is confounded by two major issues. First, it is rare for a sample of objects to be representative of any larger and well-defined set; instead, spatial data analysis is commonly applied to all of the objects that exist in a given study area. Second, it is also rare for objects distributed in space to be selected *independently* from any larger set; instead, the property of spatial dependence virtually ensures that nearby objects will have some degree of similarity.

KEY APPLICATIONS

As noted earlier, techniques of spatial data analysis can be applied to virtually all spaces, and all phenomena distributed within such spaces. Nevertheless, the vast majority of applications are found in geographic space, that is, the space defined by the surface and near-surface of the Earth, at spatial resolutions ranging from sub-meter to global.

Many important applications have derived from the need to understand the mechanisms of disease, and particularly its transmission within human populations. The work of Dr. John Snow on cholera [8] is often cited as the seminal example, but today methods of spatial data analysis are routinely used to scan data on such diseases as cancer, searching for anomalous clusters and thus for potential causal mechanisms. Spatial data analysis has been central to the study of outbreaks of new diseases such as West Nile virus and SARS.

Spatial data analysis has also been central to the study of landscape change, and related phenomena of urban sprawl, deforestation, desertification, and habitat destruction. Such analyses are often based on snapshots of landscape obtained from Earth-orbiting satellites, and can form the basis for sophisticated models of landscape change that can be used to investigate the future effects of management alternatives.

Transportation applications are also particularly rich. Methods of spatial data analysis are routinely used to model traffic patterns, and to evaluate planning options, including new

roads, mass transit, and congestion pricing. The possibility of real-time tracking of vehicles using GPS has recently given this field new impetus.

FUTURE DIRECTIONS

The insights that can be obtained from spatial data analysis are limited by its essentially cross-sectional nature – the need to draw inferences from snapshots obtained at one point in time. It is difficult, for example, to ascribe cause when no information is available about change through time. Thus there is great interest in the development of an improved suite of methods for *spatiotemporal* data analysis. In the past the lack of suitable data has been a major impediment, but today vast new sources are becoming available as the result of developments in satellite remote sensing, GPS tracking, and Internet-based data sharing.

GIS owes much of its original stimulus to the paper map, which is of necessity flat. At global scales, analysis based on flattened or *projected* views of the Earth's surface can be misleading, and there is therefore strong interest in developing methods of spatial data analysis for the Earth's curved surface. This interest has been stimulated in part by the recent emergence of *virtual globes*, including Google Earth.

CROSS REFERENCES: Geographic Information System, Spatial Data Types, Spatial Operations and Map Operations

RECOMMENDED READING

1. Berry B. J. L., Marble D. F. (1968): Spatial Analysis: A Reader in Statistical Geography. Prentice Hall.
2. Bunge, W. (1966): Theoretical Geography. Gleerup.
3. Longley P. A., Goodchild M. F., Maguire D. J., Rhind D. W. (2005): Geographic Information Systems and Science. Wiley.
4. Tomlin C. D. (1990): Geographic Information Systems and Cartographic Modeling. Prentice Hall.
5. Haining R. P. (2003): Spatial Data Analysis: Theory and Practice. Cambridge University Press.
6. Bailey T. C., Gatrell A. C. (1995): Interactive Spatial Data Analysis. Longman.
7. Fotheringham A. S., Brunsdon C., Charlton M. (2002): Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley.
8. Johnson S. (2006): The Ghost Map: The Story of London's Most Terrifying Epidemic and How It Changed Science, Cities, and the Modern World. Riverhead.
9. O'Sullivan D., Unwin D. J. (2003): Geographic Information Analysis. Wiley.

FIGURE CAPTIONS

1. The buffer operation. Half-mile buffers have been drawn around points representing the locations of schools in an area of central Los Angeles. Such buffers are often required by legislation; this example was motivated by a proposal to ban registered sex offenders from living within a prescribed distance of schools.
2. Two-dimensional equivalents of the mean and standard deviation. The larger ellipse shows the dispersion of the white population of Milwaukee around its centroid; the

smaller ellipse shows the greater concentration of the city's black population. The map shows percent black, using 1990 data by census tract.