14 Spatial Data Analysis

OVERVIEW

- This chapter is the first in a set of three dealing with geographic analysis and modeling methods.
- The chapter begins with a review of the relevant terms, and an outlines the major topics covered in the three chapters
- Examines methods constructed around the concepts of location, distance, and area

LEARNING OBJECTIVES

- **Definitions of spatial data analysis and tests to determine whether a method is spatial.**
- **Techniques for detecting relationships between the various properties of places and for preparing data for such tests.**
- **Methods to examine distance effects, in the creation of clusters, hotspots, and anomalies.**
- **The applications of convolution in GIS, including density estimation and the characterization of neighborhoods.**

KEY WORDS AND CONCEPTS

Spatial analysis, inductive, deductive, normative, queries, measurements, transformations, , algorithm, metric, buffer, point in polygon, polygon overlay, spurious polygons, coastline weave,

tolerance, Thiessen polygons, Inverse-distance weighting (IDW), Kriging, semivariograms, density estimation

OUTLINE

CHAPTER SUMMARY

14.1 Introduction: what is spatial analysis?

- The techniques covered in these three chapters are generally termed spatial rather than geographic, because they can be applied to data arrayed in any space, not only geographic space.
- Spatial analysis the crux of GIS because it includes all of the transformations, manipulations, and methods that can be applied to geographic data to add value to them, to support decisions, and to reveal patterns and anomalies that are not immediately obvious
    - Spatial analysis is the process by which we turn raw data into useful information,
- The term analytical cartography is sometimes used to refer to methods of analysis that can be applied to maps to make them more useful and informative
- In this and the next chapter the authors look first at some definitions and basic concepts of spatial analysis.
- Chapter 16 is devoted to spatial modeling, a loosely defined term that covers a variety of more advanced and more complex techniques, and includes the use of GIS to analyze and simulate dynamic processes, in addition to analyzing static patterns.
- The human eye and brain are also very sophisticated processors of geographic data and excellent detectors of patterns and anomalies in maps and images.
    - So the approach taken here is to regard spatial analysis as spread out along a continuum of sophistication, ranging from the simplest types that occur very

quickly and intuitively when the eye and brain look at a map, to the types that require complex software and sophisticated mathematical understanding.

- Spatial analysis is a set of methods whose results change when the locations of the objects being analyzed, or the frame used to analyze them, changes.

14.1.1 Examples

- John Snow map of cholera
- Openshaw's technique which generates a large number of circles, of random sizes, and throws them randomly over the map. The computer generates and places the circles, and then analyzes their contents, by dividing the number of cases found in the circle by the size of the population at risk. If the ratio is anomalously high, the circle is drawn (Figure 14.4).
- Spatial analysis can be
    - *inductive*, to examine empirical evidence in the search for patterns that might support new theories or general principles, in this case with regard to disease causation.
    - *deductive*, focusing on the testing of known theories or principles against data
    - *normative*, using spatial analysis to develop or prescribe new or better designs

14.2.1 Analysis of Attribute Tables

- Likely that the kinds of factors responsible for the occurrence of an observed phenomenon are contained within the attribute tables of a GIS
    - One way to examine this suspicion is to plot one variable against the other as a scatterplot.
    - Allow us to examine in detail the dependence of one variable on one or more independent variables.
- Regression analysis focuses on finding the simplest relationship indicated by the data.
    - Multiple regression extends this principle to consider the effects of multiple independent variables
- Relationships between variables can vary across space, which is an issue termed spatial heterogeneity
    - Geographers have developed a set of techniques that recognize such heterogeneity explicitly:
    - The example of Geographically Weighted Regression is given

14.2.2 Spatial Joins

- One of the most powerful features of a GIS is the ability to join tables based on common geographic location.

14.2.3 The Point-in-Polygon Operation

- The point in polygon operation is used to determine whether a point lies inside or outside a polygon.
- Occurs when point-like events must be compared to properties of the surrounding environment

14.2.4 Polygon Overlay

- Polygon overlay is similar to the point-in-polygon operation in the sense that two sets of objects are involved
- Exists in two form depending whether discrete or continuous perspective is taken:
    - The complexity of computing a polygon overlay was one of the greatest barriers to the development of vector GIS
    - From the discrete-object perspective, the task is to determine whether two area objects overlap, to determine the area of overlap, and to define the area formed by the overlap as one or more new area objects
    - The continuous-field version of polygon overlay does this by first computing a new dataset in which the region is partitioned into smaller areas that have uniform characteristics on both variables.
- One of the issues that must be tackled by a practically useful algorithm is known as the spurious polygon or coastline weave problem.
    - Although the same boundary line may be represented in both source datasets, its representations will almost certainly not be the same.
    - The most common method for dealing with this is the specification of a tolerance.
        - If two lines fall within this distance of each other, the GIS will treat them as a single line and not create slivers.

14.2.5 Raster Analysis

- Overlay in raster is much simpler – the attributes of each cell are combined according to a set of rules

14.3 Analysis Based on Distance

- The ability to calculate and manipulate distances underlies many forms of spatial analysis
    - based on the concept that the separation of features or events on the Earth's surface can tell us something useful

14.3.1 Measuring Distance and Length

- A metric is a rule for the determination of distance between points in a space.
- Pythagorean or straight-line metric is explained with an equation and diagram (Fig 14.12)
    - Notes this metric does not work for latitude and longitude, must use the spherical metric provided in Section 5.7 to calculate great circles
- Distance along a route represented by a polyline is often calculated by summing the lengths of each segment of the polyline
- Because there is a general tendency for polylines to short-cut corners, the length of a polyline tends to be shorter than the length of the object it represents.
- Length of a 3 dimensional line measured off its planimetric representation will also be shorter than its true length

14.3.2 Buffering

- Builds a new object or objects by identifying all areas that are within a certain specified distance of the original objects
- In raster, buffers can be spread outwards from objects to create friction surfaces

14.3.3 Cluster Detection

- Points patterns can be identified as clustered, dispersed, or random
- Kinds of processes responsible for point patterns are:
    - First-order processes involve points being located independently
    - Second-order processes involve interaction between points
- Briefly introduces the K function as an example of a descriptive statistic of pattern, and explains a simple example

14.3.4 Dependence at a Distance

- The Moran statistic (introduced in Section 4.5) is a global measure that distinguishes between positively (clustered) and negatively (dispersed) autocorrelated patterns.
- Local measures of clustering can be used to identify hot spots

14.3.5 Density Estimation

- Convolution is described as the attenuating effect of distance on a function
- Potential functions have many uses in spatial analysis and are intended to measure influence at a distance.
- Kernel function is a central idea in density estimation
    - In density estimation, each point is replaced by its kernel function and the various kernel functions are added to obtain an aggregate surface, or continuous field of density.

14.3.6 Spatial Interpolation

- Spatial interpolation is a process of intelligent guesswork, in which the investigator (and the GIS) attempt to make a reasonable estimate of the value of a continuous field at places where the field has not actually been measured.
- Notes that the one principle that underlies all spatial interpolation is the Tobler Law

14.3.6.1 Thiessen polygons

- To estimate value at any point take the value measured at the closest point. This leads to a map in which the value is constant within polygons surrounding each point.

14.4.4.2 Inverse-distance weighting

- IDW is the method that is most often used by GIS analysts.
- It estimates unknown measurements as weighted averages over the known measurements at nearby points, giving the greatest weight to the nearest points.
- The mathematical notation is given
- There are various ways of defining weights, but the option most often employed is to compute them as the inverse squares of distances
- IDW is an exact method of interpolation because its interpolated results honor the data points exactly

- Cautions that IDW uses weights that are never negative, values are always calculated between the limits of the measured values, and it tends to regress to the mean outside the area of the data points (Figure 14.25)

14.3.6.3 Kriging

- The basic idea is to discover something about the general properties of the surface, as revealed by the measured values, and then to apply these properties in estimating the missing parts of the surface
- Provides a good but general explanation of the development of the semivariogram, isotropic and anisotropic semivariograms, range, sill, and nugget.
- Notes that a non-zero nugget occurs when there is substantial error in the measuring instrument
- To make estimates using Kriging, need to reduce the semivariogram to a mathematical function, usually by selecting one from a set of standard functional forms and fitting that form to the observed data points in the semivariogram

14.4 Conclusion

Essay topics

1. What is meant by the process of density estimation, and why is it the logical twin of spatial interpolation?
2. Why does the interpolation of real land-surface heights challenge standard methods of automated interpolation?
3. Statistically-minded medical epidemiologists were very critical of the original Geographical Analysis Machine (Figure 14.4) analysis. Outline the cases for and against the approach it took.
4. Evaluate the view that, although ‘grounded in good theoretical principles’, in practice Kriging is as arbitrary as any other automated contouring approach.
5. Differentiate between first and second order effects in geospatial data and provide illustrative examples of each. Why is it almost always virtually impossible to differentiate between them in practical analysis?
6. Attempt a review and classification of approaches to spatial point pattern analysis.
7. What are the key problems in the application of statistical hypothesis testing to geospatial data?

8. Evaluate the effectiveness of the Moran's I statistic for measuring global spatial autocorrelation paying particular attention to the input data, ease of computation, sampling distributions, and interpretation.
9. Why is polygon overlay such a central operation in a vector GIS environment, and what are the essential steps in a good algorithm to accomplish it?
10. Compare and contrast Kriging with inverse distance weighting as methods for spatial interpolation from control point samples.
11. How and why does geospatial analysis using geographical data differ from the 'spatial analysis' conducted by other sciences? Which is the more difficult and why?

MULTIPLE CHOICE QUESTIONS (MCQ)

1. In spatial analysis in a GIS environment, which of the following is the most important resource?
   a. A really powerful computer
   b. Good functionality in the software
   c. An intelligent user
   d. High quality data

2. If we estimate the length of a real world feature using its representation as a polyline in a GIS, is the result almost certain to be (a) shorter or (b) longer than the real world feature?

3. For each of the techniques listed below, state whether or not it being a sensible thing to do depends on the truth of the Tobler 'First law' of geography:

Technique	**Law? Y/N**
Spatial queries	
Buffering	
Interpolation	
Density estimation	

4. For which type of data is kernel density estimation appropriate?
   a. A point pattern

b. A sample of heights
c. Values aggregated over areas

5. Tick the true answer or answers to the following five statements. 'Inverse distance weighting and Kriging are similar interpolators because they both:

a. specify a spatial structure in the data
b. are objective
c. compute a distance-weighted sum of neighboring data values
d. provide estimates of the error at every point in the field
e. have a good theoretical basis'.

6. A metric is a ……?

ACTIVITIES

1. Deconstructing John Snow: 2004 marked the 150th anniversary of John Snow's celebrated work on cholera in London, and Biographical Box 14.1 presents the standard story. His work has been presented as pioneering spatial epidemiology and disease mapping, geovisualization, and even GIS 'overlay'. Create a poster that de-bunks the myth by examining the timing of his intervention, the relationship to a priori theory, and the detail of the map itself. Possible sources of information can be found in:

    a. Brody, H et al, (2000) Map-making and myth making in Broad Street: the London cholera epidemic, 1854 *The Lancet*, 356(9223) 64-68
    b. www.jsi.com
    c. www.ph.ucla.edu/epi/snow.html, which has numerous useful materials

Reporting results by means of a poster is much more challenging than at first sight it appears. Students will need direction on how to do this effectively and have access to appropriate resources. Some advice can be found at www2.glos.ac.uk/gdn/abstracts/a141.htm.

Nowadays, the same objectives can be addressed by production of either a PowerPoint presentation or a Web page, with the supreme advantage that these can easily be archived and made available to subsequent classes as examples of good and bad practice.

2. The data file below, taken from Davis (2002, Figure 5.66, page 373), has three columns for the (x, y, z) co-ordinates of a sample of survey control points of some topographic (relief) data:

x	**y**	**Height**
0.3	6.1	870
1.4	6.2	793
2.4	6.1	755
3.6	6.2	690
5.7	6.2	800
1.6	5.2	800
2.9	5.1	730
3.4	5.3	728
3.4	5.7	710
4.8	5.6	780
5.3	5	804
6.2	5.2	855
0.2	4.3	830
0.9	4.2	813
2.3	4.8	762
2.5	4.5	765
3	4.5	740
3.5	4.5	765
x	**y**	**Height**
4.1	4.6	760
4.9	4.2	790
6.3	4.3	820
0.9	3.2	855
1.7	3.8	812
2.4	3.8	773
3.7	3.5	812
4.5	3.2	827
5.2	3.2	805
6.3	3.4	840
0.3	2.4	890
2	2.7	820
3.8	2.3	873
6.3	2.2	875
0.6	1.7	873
1.5	1.8	865
2.1	1.8	841
2.1	1.1	862
x	**y**	**Height**
3.1	1.1	908
4.5	1.8	855
5.5	1.7	850
5.7	1	882
6.2	1	910
0.4	0.5	940
1.4	0.6	915
1.4	0.1	890
2.1	0.7	880
2.3	0.3	870
3.1	0	880
4.1	0.8	960
5.4	0.4	890
6	0.1	860
5.7	3	830
3.6	6	705

a. Import these data into your GIS and apply a selection of interpolation techniques (such as IDW and Kriging) to produce alternative visualizations;
b. Select the result that you think best reconstructs the entire field from the sample survey point data and in less than 250 words, explain why you think this is 'best';
c. Visualize this 'best' result using alternative display techniques.

Initially, you might attempt to isoline these data by hand (see Chapter 4).

3. Using a standard road atlas for your country, select any five towns. To complete the exercise the atlas must include a table/matrix of the road distances between towns and your five should be among the listed places.

   a) Locate your five places on a simple grid placed over the map and read off their (x, y) co-ordinates.
   b) Use the standard formula for the straight line distance given in Section 14.3.1 to compute the ten unique distances between these places and assemble these into a matrix.
   c) O'Sullivan and Unwin (2010, Chapter 5) show how this type of matrix can be used to develop almost all of the standard point pattern statistics, as well as 'adjacency' matrices that implement differing conceptions of 'next to'. Using a suitable threshold distance, convert your matrix into a 0/1 binary matrix of adjacencies.
   d) Examine each of the measures you have created in relation to the three properties that any metric distance should have. Simply stated these are (1) that the distance between points must be a positive number unless the points are the same, in which case the distance will be zero; (2) that the distance between two points is independent of which way round it is measured; and (3) the *triangle inequality*, which states that it must always be at least as far to travel between two points via a third point rather than to travel directly.
   e) Now assemble a different matrix, of the observed road travel distances as given in the source atlas and compare this with your 'straight line' values. There will be a general similarity, but what reasons are there for the differences? How do these distances measure up to the necessary properties?
   f) Use the road distance matrix to compute for each row (place) a relative distance, defined as the road distance from that place divided by the average of all the distances from that place (i.e. the row total divided by the number of places). Do the metric properties now hold?

g) Finally, a nearness statistic can be computed as 1/ (the relative distance + 1) and again summarized in matrix form. Comment on what it shows.

The second half of this experiment is detailed by Worboys, M.F., (1996) Metrics and topologies for geographic space , in Advances in Geographic Information Systems Research II: Proceedings of the International Symposium on Spatial Data Handling, Delft, Kraak, M.J. and Molenaar, M. (eds.), Taylor & Francis: London, pages 365-376, and at www.spatial.maine.edu/~worboys/mywebpapers/sdh1996.pdf

4. Investigating 'naïve' notions of distance: Section 14.3.1 discusses the metric distances used in almost all spatial analysis with GIS. As we saw in Activity 5, Chapter 4, this is a notion held mostly by spatially aware professionals (SAPs) and it may not always represent the ideas that individuals have about distance. A 2001 paper by Michael Worboys (Biographical Box 10.3) "Nearness relations in environmental space", *International Journal of Geographical Information Science*, 15(7):633-651 outlines a simple experiment that investigates the idea of distance measured as nearness or proximity. In the paper he presents an experiment with human subjects concerning the vague spatial relation 'near' in environmental space. Three approaches to experimental analysis are presented and discussed: nearness neighborhoods as regions with broad boundaries, fuzzy nearness and distance measures, and four-valued logic. The text of the paper is available at www.spatial.maine.edu/~worboys/mywebpapers/ijgis2001.pdf. Using a class of volunteers and the area around your location, repeat the experiment, and compare your findings to those of Worboys. To what extent can GIS be modified to incorporate this, and similar measures of 'distance'?

5. Study the paper: Gatrell, A.C., Bailey, T.C., Diggle, P.J. and B.S. Rowlingson (1996) Spatial point pattern analysis and its application in geographical epidemiology. Transactions, Institute of British Geographers, NS 21: 256-274. This deals with three example studies of the use of spatial statistical analysis in medical geography.

    a. A study of 325 post-coded cases of childhood leukemia in west central Lancashire, 1954–92. Unlike studies by others of the entire northern region, or of the area of West Cumbria immediately to the north of this area, using the K(d) and D(d) functions, they fail to detect any evidence of purely spatial clustering;

b. A study of Burkitt's lymphoma in Uganda using 174 cases that shows some evidence of space-time clustering as evaluated using the space-time difference function D(d, t);
c. A study of lung cancer in Chorley and South Ribble (Lancashire) using a so-called raised incidence model designed to test whether or not there is a significant effect in distance from a waste incinerator. This suggests that there may well be an effect that would repay detailed medical scrutiny.

In each case summarize the problem, available data, reasons for the choice of technique and result of the analysis.

6. Visit www.csiss.org and download the GeoDa software and associated data and tutorial files. Use it to illustrate the creation of Thiessen polygons from a point data set such as the locations of juvenile crimes set taken from Bailey, T. and Gatrell, A. (1995) *Interactive Spatial Data Analysis*, Wiley: NY, page 95.

7. The concepts involved in ordinary and universal Kriging are not easily explained, and there is no substitute for a structured activity that covers at least some of the ground. The entire process is best introduced in three distinct phases (variogram cloud, semi-variogram and model, and estimate computation) each treated on its merits, rather than as a single 'button click' in a GIS. Use some suitable point-valued data, such as those provided in Activity 3 above, to compute a variogram 'cloud' and then summarize this as an experimental semi-variogram by dividing the distance axis into 'bins' and computing a series of means. Your GIS may be able to do this (check), but if not download and use one or other of the readily available packages for geostatistical analysis such as GSLIB (www.gslib.com), GS+ (www.geostatistics.com) or VarioWin (http://www-sst.unil.ch/research/variowin/). Failing this, a few simple lines of code are all that is needed to compute the values and send them to a file, which can then be entered into any spreadsheet or statistical analysis program for further analysis and visualization. Using the data from Activity 3, O'Sullivan and Unwin (2002, pages 45-49 and 265-273) illustrate a typical analysis. Modeling the semi-variogram by fitting an appropriate model is likely to be computationally more difficult, but again is worthwhile doing as a distinct step. Finally, computation of the Kriging estimates and their mapping will certainly need computer power (see O'Sullivan and Unwin, pages 274-281).

8. A phrase that is often used is that 'people who play with sharp tools often get cut'. Examine the defaults in any GIS known to you for a selection of the methods introduced

in this chapter, list them and then indicate whether or not you think that a) a beginner, b) an average user and c) a knowledgeable user would 'get cut' using them.

FURTHER READING

Cliff A D, Ord J K 1973 *Spatial Autocorrelation*. London: Pion.

Bailey T C, Gatrell A C 1995 *Interactive Spatial Data Analysis*. Harlow, UK: Longman Scientific and Technical.

Burrough P A, McDonnell R A 1998 *Principles of Geographical Information Systems*. New York: Oxford University Press.

De Smith M J, Goodchild M F, Longley P A 2009. *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*. Third Edition. Winchelsea: Winchelsea Press. www.spatialanalysisonline.com.

Isaaks E H, Srivastava R M 1989 *Applied Geostatistics*. New York: Oxford University Press.

O'Sullivan D, Unwin D J 2010 *Geographic Information Analysis*. Hoboken, NJ: Wiley.

Silverman B W 1986 *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

RELATED READING

Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds) 2005 *Geographical Information Systems: Principles, Techniques, Management and Applications* (abridged edition). Hoboken, NJ: Wiley.

16. Spatial statistics, A Getis
17. Interactive techniques and exploratory spatial data analysis, L Anselin
19. Spatial analysis: retrospect and prospect, M M Fischer

Maguire D J, Goodchild M F, Rhind D W (eds) 1991 *Geographical Information Systems: Principles and Applications*. Harlow, UK: Longman (text available online at www.wiley.co.uk/gis/volumes.html).

21. The functionality of GIS, D J Maguire and J Dangermond, pp. 319-35
22. Information integration and GIS, I D H Shepherd, pp. 337-60
23. Cartographic modeling, C D Tomlin, pp. 361-74
24. Spatial data integration, R Flowerdew, pp. 375-87
25. Developing appropriate spatial analysis methods for GIS, S Openshaw, pp. 389-402
26. Spatial decision support systems, P J Densham, pp. 403-12

ONLINE RESOURCES

ESRI Virtual Campus course, *Turning Data into Information* by Paul Longley, Michael Goodchild, David Maguire, and David Rhind (training.esri.com)

Module 1: Basics of Data and Information

Module 3: Query and Measurement

Module 4: Transformations and Descriptive Summaries

Module 1: Basics of Data and Information

Unit: Creating and visualizing information

Sub-unit: Types of spatial analysis

Sub-unit: What is spatial analysis?

Module 3: Query and Measurement, Unit: Querying views of a GIS

Section 14.2, Module 1: Basics of Data and Information

Unit: Creating and visualizing information,

Sub-unit: Types of spatial analysis

Module 3: Query and Measurement

Unit: Querying views of a GIS

Unit: Advanced queries

Module 4: Transformations and Descriptive Summaries

Unit: Histograms, pie charts, and scatterplots

Section 14.3, Module 1: Basics of Data and Information

Unit: Creating and visualizing information,

Sub-unit: Types of spatial analysis

Module 3: Query and Measurement

Unit: Querying for measurements

Section 14.3.1, Module 3: Query and Measurement

Unit: Querying for measurements,

Sub-unit: Distance and length

Section 14.3.2, Module 3: Query and Measurement

Unit: Querying for measurements,

Sub-unit: Shape

Section 14.3.3, Module 3: Query and Measurement

Unit: Querying for measurements,

Sub-unit: Slope and aspect

Section 14.4, Module 1: Basics of Data and Information

Unit: Creating and visualizing information,

Sub-unit: Types of spatial analysis

Module 4: Transformations and Descriptive Summaries

Unit: Buffering, point-in-polygon, and polygon overlay

Section 14.4.1, Module 4: Transformations and Descriptive Summaries

Unit: Buffering, point-in-polygon, and polygon overlay,

Sub-unit: Buffering

Section 14.4.2, Module 4: Transformations and Descriptive Summaries

Unit: Buffering, point-in-polygon, and polygon overlay,

Sub-unit: Point-in-polygon

Section 14.4.3, Module 4: Transformations and Descriptive Summaries

Unit: Buffering, point-in-polygon, and polygon overlay,

Sub-unit: Polygon overlay

Section 14.4.4, Module 4: Transformations and Descriptive Summaries

Unit: Spatial interpolation and density estimation

Section 14.4.4.2, Module 4: Transformations and Descriptive Summaries

Unit: Spatial interpolation and density estimation

Sub-unit: Inverse Distance Weighting

Section 14.4.4.3, Module 4: Transformations and Descriptive Summaries

Unit: Spatial interpolation and density estimation

Sub-unit: Kriging

Section 14.4.5, Module 4: Transformations and Descriptive Summaries

Unit: Spatial interpolation and density estimation

Sub-unit: Calculating density

NCGIA Core Curriculum in GIScience, 2000 (www.ncgia.ucsb.edu/giscc)

2.1.2.1. Simple Algorithms for GIS I: Intersection of Lines (184)

2.1.2.2. Simple Algorithms for GIS II: Operations on Polygons, (185)

2.1.2.3. The Polygon Overlay Operation (186)

2.14.2. Exploratory Spatial Data Analysis (128), *Robert Haining and Stephen Wise*

NCGIA Core Curriculum in GIS, 1990 (www.ncgia.ucsb.edu/pubs/core.html)

5. Raster GIS capabilities

14. Vector GIS capabilities

15. Spatial analysis

32. Simple algorithms I - line intersection

33. Simple algorithms II - polygons

34. Polygon overlay

40. Spatial interpolation I

41. Spatial interpolation II