4 The Nature of Geographic Data

OVERVIEW

- Elaborates on the *spatial is special* theme
- Focuses on how phenomena vary across space and the general nature of geographic variation
- Describes the main principles that govern scientific sampling, how spatial variation is formalized and measured as spatial autocorrelation, and outlines the concept of fractals.

LEARNING OBJECTIVES

- **How Tobler's First Law of Geography is formalized through the concept of spatial autocorrelation;**
- **The relationship between scale and the level of geographic detail in a representation;**
- **The principles of building representations around geographic samples;**
- **How the properties of smoothness and continuous variation can be used to characterize geographic variation;**
- **How fractals can be used to measure and simulate surface roughness.**

KEYWORDS AND CONCEPTS

Time and space; spatial autocorrelation and the Tobler Law; scale; representation; types of spatial objects; fractals and self-similarity; spatial sampling; distance decay; induction and deduction; isopleth maps; choropleth maps; adjacency; regression; generalization.

OUTLINE

CHAPTER SUMMARY

4.1 Introduction

Reviews the governing principles of the development of representations already covered, and adds three more related to the *nature of spatial variation*:

- that proximity effects are key to understanding spatial variation, and to joining up incomplete representations of unique places;
- that issues of geographic scale and level of detail are key to building appropriate representations of the world;
- that different measures of the world co-vary, and understanding the nature of co-variation can help us to predict

4.2 The fundamental problem revisited

- Distinguishes between *controlled* variation, which oscillates around a steady state, and *uncontrolled* variation.
- Some applications address controlled variation, such as utility management
- Others address uncontrolled, such as those studying longer term processes
- Introduces the concept of *time series* and acknowledges the concept of *temporal autocorrelation*.
- “Our behavior in space often reflects past patterns of behavior”, thus it is one-dimensional, need only look in the past
- *Spatial heterogeneity* is the tendency of geographic places and regions to be different from each other.
- Occurs in both form and process

- The principles addressed in this chapter help answer questions about what to leave in and what to take out of digital representations
- Scale and spatial structure help determine how to sample reality and weight sample observations to build representations

Technical Box 4.1 Types of spatial objects

- Classifies geographic objects by their topological dimension
- Points, lines, area objects, volume objects, time (as the 4th dimension)
- Classification of spatial phenomena into object types is dependent fundamentally upon scale

4.3 Spatial autocorrelation and scale

- Spatial autocorrelation measures attempt to deal simultaneously with similarities in the location of spatial objects and their attributes
- Brief discussion about how *neighboring* might be defined, more later in this chapter.
- Measures of spatial and temporal autocorrelation are *scale dependent*
- The issue of *sampling interval* is of direct importance in the measurement of spatial autocorrelation
- When the pattern of spatial autocorrelation at the coarser scale is replicated at the finer scale, the overall pattern exhibits the property of *self-similarity*.

Technical Box 4.2 The many meanings of scale

- Scale is in the details
- Scale is about extent
- Scale of a map – including reference to *representative fraction* and confusion between large and small scales. This book, thus, uses “coarse” and “fine”.

4.4 Spatial sampling

- *Sample frame* is defined as the universe of eligible elements of interest.
- Might be bounded by the extent of the field of interest or by the combined extent of a set of areal objects
- *Sampling* is the process of selecting points from a continuous field or selecting some objects while discarding others
- Any geographic representation is a kind of sample
- Procedures of *statistical inference* allow us to infer from samples to the population from which they were drawn
- Classical statistics often emphasizes the importance of randomness in sound sample design

- Types of sampling designs include: simple random sampling, spatially systematic sampling (problems if the sampling interval and spatial structure coincide so that the sample frame exhibits *periodicity*), stratified random sampling, periodic random changes in the sampling grid, clustered sampling, sampling along transects
- In circumstances where spatial structure is either weak or is explicitly incorporated through clear definition of subpopulations, standard statistical theory provides a robust framework for inferring the attributes of the population from those of the sample
- However, the existence of spatial autocorrelation fundamentally undermines the inferential framework and invalidates the process of generalizing from samples to populations.

4.5 Distance decay

- Discusses the attenuating effect of distance and the need to make to make an informed judgment about an appropriate *interpolation* function and how to *weight* adjacent observations.
- Explains and illustrates the structure of the distance decay equation: *b* as a parameter that affects the rate at which the weight w_{ij} declines with distance
- Discusses linear, negative power, and negative exponential distance decay equations and graphs
- Notes that with these equations, the effects of effects of distance are presumed to be regular, continuous, and *isotropic* (uniform in every direction)
- Several paragraphs discuss how these simple equations may not reflect reality

Technical Box 4.3 Isopleth and choropleth maps

- Explains how isopleth and choropleth maps are constructed and displayed
- Choropleth maps describe properties of non-overlapping areas.
- In this section is the important discussion about spatially extensive and intensive variables:
- Spatially extensive variables are true only of entire areas, such as total population, or total number of children under 5 years of age.
- Spatially intensive variables could potentially be true of every part of an area, if the area were homogeneous – examples include densities, rates, or proportions.
- The figure caption explains that spatially extensive variables should be converted to spatially intensive form.

4.6 Measuring distance effects as spatial autocorrelation

- *Induction* reasons from data to build up understanding, while *deduction* begins with theory and principle as a basis for looking at data.
- Knowledge of the actual or likely nature of spatial autocorrelation can be used *deductively* in order to help build a spatial representation of the world.
- The *measurement* of spatial autocorrelation is a more i*nductive* approach to developing an understanding of the nature of a geographic dataset.
- If the phenomenon is conceived as a field, then spatial autocorrelation measures the smoothness of the field using data from the sample points, lines, or areas that represent the field.
- If the phenomena of interest are conceived as discrete objects, then spatial autocorrelation measures how the attribute values are distributed among the objects, distinguishing between arrangements that are clustered, random, and locally contrasting.
- Figure 4.11 shows examples of each of the four object types, with associated attributes, chosen to represent situations in which a scientist might wish to measure spatial autocorrelation.

Technical Box 4.4 Measuring similarity between neighbors

- Outlines the concept of the weights matrix, w_{ij}
- Measures of spatial autocorrelation compare a set of locational similarities w_{ij} with a corresponding set of attribute similarities c_{ij} , combining them into a single index in the form of a cross-product.
- Explains ways to measure similarity of attributes
- Briefly outlines the Moran Index

4.8 Taming geographic monsters

- Expands the discussion in Technical Box 4.6 into *fractal geometry*
- Fractals can be thought of as geometric objects that are between Euclidean dimensions
- Ascertaining the fractal dimension of an object involves identifying the scaling relation between its length or extent and the yardstick (or level of detail) that is used to measure it.
- Illustrates the log-log relationship between length and step length

Technical Box 4.5 The Strange Story of the Lengths of Geographic Objects

- Introduces the method of measuring a line by counting the number of steps of a given length
- Using these measurements to determine fractal dimension is covered in the regular text following this box

ESSAY TOPICS

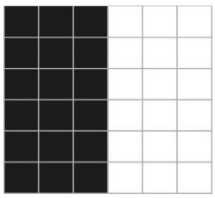
1. Why, and under what circumstances, do GI scientists sample?
2. Chapter 4 is mostly about how geography is represented in a digital computing environment, but what does 'representation' actually mean and what are the limits on what it can achieve?
3. In what ways is the advent of mass data storage in online environments (e.g. 'the Cloud') having an effect on representation?
4. If it is the case that many geographic objects have fractal characteristics, what are the consequences for how we might represent them in a GIS?
5. Section 4.2 develops the idea of spatial heterogeneity. How can this be seen to be in conflict with the Tobler Law (Section 3.1)?
6. What are some of the determinants of how phenomena are represented on maps?
7. Compare and contrast the representation problems given by data about 'where' with those about 'when'.
8. An argument has been made that how we represent geography in GIS can be improved by understanding human perception and cognition. In what ways do you agree or disagree with this idea?
9. Define what in science is meant by the words 'induction' and 'deduction' and give a reasoned account of why most geographic knowledge has been derived by induction.

MULTIPLE CHOICE QUESTIONS (MCQ)

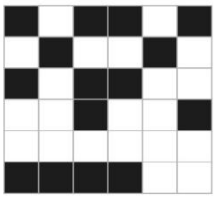
1. To a cartographer, is a 1:250,000 map conventionally referred to as large (A) or small (B) scale?
2. The Tobler Law is formalized by which of the following concepts?
    a. spatial heterogeneity
    b. fractal dimensions
    c. spatial heterogeneity
    d. spatial autocorrelation

3. Study the three 6 by 6 ‘chess boards’:

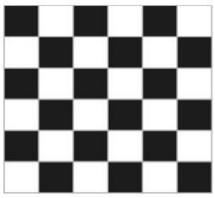
(a)



(b)



(c)



   a. Which of these illustrates positive spatial autocorrelation?
   b. Which of these illustrates negative spatial autocorrelation?

4. Using the so-called ‘Rook’s Case’ definition of adjacency in which only links along rows and down columns are allowed, how many black to black joins (BB) are there in (b). Is it 27, 0 or 6?
5. Using the so-called ‘Queens Case’, in which diagonal links are also allowed, how many black to black joins (BB) are there in (b)? Is it 47, 14 or 25?
6. A more than expected number of black to white joins indicates positive spatial autocorrelation. TRUE or FALSE?
7. Of the following alternatives ranges, in which range is the typical fractal dimension of sinuous lines used to map coastlines such as that of Norway likely to lie?
   a. Below 1.0
   b. Between 1.5 and 1.9
   c. Between 1.1 and 1.5
   d. Above 2.0
8. For each of the following real geographic features, write down the length dimension of the database object that would best represent them in a GIS developed for mapping at the 1:50,000 scale. (Answer L0, L1, L2 or L3)

a. Public building
b. Highway
c. Housing
d. Woodland
e. Surface relief
f. Census tract
g. Railway

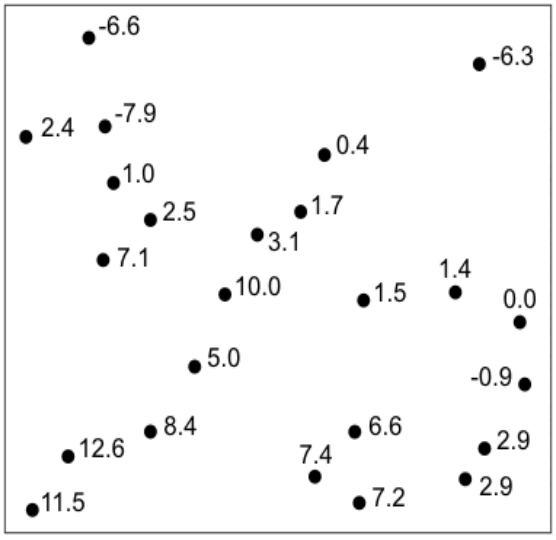
9. Choropleth maps should only be used be used to display …?…… spatial variables.
    a. extensive
    b. nominal
    c. ordinal
    d. intensive
10. Induction can be defined as …..?
    a. beginning with theory and principles as a basis for looking at data
    b. a way of measuring spatial autocorrelation
    c. reasoning from data to build up understanding
11. For each of the listed datasets, state which type of map would be most appropriate for its display (select isopleth, choropleth or neither of these):
    a. Earth surface relief
    b. % of senior citizens in the population
    c. height of the 500mb atmospheric pressure surface
    d. land use
    e. population density in people/km2
12. Which of the following is not one of the many meanings of the word ‘scale’?
    a. extent
    b. fraction
    c. detail

CLASS AND INDIVIDUAL ACTIVITIES

1. Figure 4.4 presents seven possible point sampling schemes. Figure 3.13 showed a 1:24,000 USGS topographic map on which the wooded area is shown as a light green wash. Use at least three of the sampling designs (random, grid and cluster are suggested) estimate the actual wooded area. Each point sample can be recorded as ‘wooded/not wooded’ and the percentage of the total map area estimated in each case by the proportion of ‘hits’ on the woodland. Initially use a low number of points, but then double this and repeat the exercise. A simple way to generate a random

point sample is to lay transparent graph paper over the map and then use tables of random numbers for both the *x* and *y* co-ordinates of each point.

2. Many students assume that it is easy to create a continuous isopleth map (Technical Box 4.3) from a field of point attribute values. This exercise, from O'Sullivan and Unwin (2010, pages 251-253) will show them otherwise. It is also a useful introduction to automated contouring.



The point attributes are 'spot heights' of the average January temperature (^{o}F) in a part of Alberta, Canada; your task is simple: Get a pencil (you may also need an eraser!) and produce a continuous surface representation of these data, by drawing 'contours' of equal temperature (in climatology these are called iso-therms). In threading these through the data, bear in mind three things. First, don't 'join the dots'. Remember that the data are unlikely to be exact and, even with a 0.1^{o} resolution, each isotherm is likely to have substantial spatial width. In many applications it is wildly optimistic to assume that the data are exact. Second, experience suggests that it pays to start the process with a contour value in the middle of the data range and to work up and down from this value. Third, you should try to make the resulting surface of average temperatures as smooth as you can, consistent with it honoring all the data. By honoring the data we mean that the interpolated surface should pass through all the measured data exactly. In practice, this means that there should be no inconsistencies where measured temperatures lie on the 'wrong' side of relevant isotherms.

3. Establishing spatial relationships between zones. Figure 4.12 shows a simplified mosaic of zones, which has been used to generate Table 4.1, a matrix W of

adjacencies in which each element w_{ij} is coded '1' if the zones are adjacent and '0' otherwise. Find a small pattern of zones, such as the Standard Economic Regions of England and Wales and use it to create a similar matrix, noting that it is symmetrical. Next, develop a definition of adjacency that is a ratio-scaled number, such as the shared boundary length (measure this by stepping with dividers). Develop this further for each zone by re-expressing this as a proportion of the total boundary length of that zone, thus creating an asymmetric W matrix. O'Sullivan and Unwin (2010, pages 48-49 and 200-205 develop this further).

4. At only a slightly more advanced level, this exercise can be developed by replacing the adjacency relationship with a distance relation, as shown 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, pp. 365-376. The original paper can be found at: www.spatial.maine.edu/~worboys/mywebpapers/sdh1996.pdf

5. Study Technical Box 4.3 and Figure 4.10 carefully. It should be clear that creating a choropleth map from 'raw' census data involves a series of decisions, any and all of which can affect the appearance of the final map. Create a simple 'workflow' in which you list each step in the process in its correct sequence, identifying at each step the range of choices available to you.

6. The best way to understand work in GIS using the concept of fractal objects and its implications for the measurement of objects that display this property is to compute the fractal dimension of an example, in this case a cartographic line using the Richardson plot. Find a reasonably detailed topographic map at a scale something like 1:24,000 or 1:50,000 and select a length of river as your object of study. Obviously, since we are interested in the sinuosity of linear objects it makes sense to choose a river that shows meandering behavior! Around a 20km length is about right and will not involve you in too much work. Now set a pair of dividers at a large equivalent distance on the ground, say 1km and 'walk' them along the river counting the number of steps. Record the yardstick length and number of segments as in example given in the lesson. Repeat using a halved yardstick, 500m; repeat again and again until the yardstick becomes so short that the experiment is impractical. You should now have a table of paired values. Ideally you should have 5 or 6 such pairs of measures. Plot a graph of the estimated length (on the vertical Y axis,

computed as the number of steps multiplied by the yardstick size) against yardstick size (on X) and comment on its shape. Convert both the numbers of steps and the yardstick lengths to their common logarithms and plot the resulting numbers with the log(number of steps) on the vertical axis and log(yardstick length) on the horizontal. If you have access to a spreadsheet you should be able to do this using that software. Hopefully, the points will fall roughly along a straight line but success isn't guaranteed.

Use the spreadsheet (or a straight edge and a good eye) to fit a linear regression line to your data and so estimate the fractal dimension of your river. What result did you get?

The regression equation can be used to estimate the fractal dimension. It is of the form:

log[L(s)] = (1-D)log(s) + b

In which Y = log(L, the total estimated length)
b = the 'intercept constant', which isn't of concern
D = the fractal dimension
s = the step length (yardstick)

Note that since the length estimated **L** gets less the longer the length you set for '**s**', the slope of the line, shown by Mandelbrot to be (1-D), is a negative one.

7. Demonstrating a key theorem in statistics, Gold et al (1990, Figure 4.2) has a very old-fashioned computer program to simulate the generation of as many samples of selected size, *n*, as required from a uniform distribution in which every number has an equal chance of occurring. You would probably want to update the code, but the idea is to use it to demonstrate that repeated sampling from this distribution produces a sampling distribution of the mean that is both normal and peaked around the true population value. Computation of the standard deviations of these means will also give a value close to the standard error of the mean of each sample. Finally, the exercise illustrates that the gain in precision of the estimate rises as the square root of *n*. The basic idea is again that of Silk (1979).

8. Imagine you are the director of the market research group of a large supermarket chain, tasked to find a location for a new store in the suburbs of a large city. Prepare

a presentation for your company Board of Directors of the factors you have considered in locating the store and estimating the likely store turnover (See Section 4.4, pages 93-95)

FURTHER READING

Batty M. and Longley P.A. 1994 *Fractal Cities: A Geometry of Form and Function*. London: Academic Press. Available for free download at www.fractalcities.org.

Mandelbrot B.B. 1983 *The Fractal Geometry of Nature*. San Francisco: Freeman.

O'Sullivan D. and Unwin D. 2002 *Geographic Information Analysis.* Hoboken, NY: Wiley

Goodchild M F 1986 *Spatial Autocorrelation*. Concepts and Techniques in Modern Geography 47. Norwich: GeoBooks

Goodchild M F 2001 Models of scale and scales of modelling. In Tate N J, Atkinson P M (eds) *Modelling scale in geographical information science*. Chichester: Wiley

Quattrochi D.A. and Goodchild M.F. (eds) 1996 *Scale in Remote Sensing and GIS*. Boca Raton, Florida: Lewis Publishers.

Silk, J.A. 1979 The use of classroom experiments and the computer to illustrate statistical concepts. *Journal of Geography in Higher Education*, 3:13-25

Tate N.J. and Atkinson P.M. (eds) 2001 *Modelling Scale in Geographical Information Science*. Chichester: Wiley.

Wright D. and Bartlett D. (eds) 2000 *Marine and Coastal Geographical Information Systems*. London: Taylor and Francis.

Wright D. 2002 *Undersea with GIS*. Redlands, CA: ESRI Press.

RELATED READING

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

2. Space, time, geography, H Couclelis

16. Spatial statistics, A Getis

17. Interactive techniques and exploratory spatial data analysis, L Anselin

Maguire D.J., Goodchild M.F. and 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).

9. Concepts of space and geographical data, A C Gatrell, pp. 119-34

30 Generalization of spatial databases, J-C Muller

ONLINE RESOURCES

Fractal Cities - http://www.fractalcities.org/

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 4: Transformations and Descriptive Summaries

Module 5: Optimization and Hypothesis Testing

Module 6: Uncertainty

Section 4.2, Module 1: Basics of Data and Information,

Unit: The nature of geographic data

Sub-unit: Spatial autocorrelation

Section 4.4, Module 1: Basics of Data and Information,

Unit: The nature of geographic data

Sub-unit: Spatial sampling

Module 5: Optimization and Hypothesis Testing,

Unit: Hypothesis testing

Sub-unit: Sampling

Module 6: Uncertainty

Unit: Measuring uncertainty of nominal and ordinal values

Sub-unit: Spatial sampling

Sub-unit: Difficulties in sampling natural areas

Section 4.5, Module 4: Transformations and Descriptive Summaries

Unit: Spatial interpolation and density estimation

Sub-unit: What is spatial interpolation?

Section 4.6, Module 4: Transformations and Descriptive Summaries

Unit: Spatial dependence and fragmentation

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

. 1.6.1. Sampling the World (031)

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

6. Sampling the world

47. Fractals