# 15 Spatial Analysis and Inference

## OVERVIEW

- This chapter focuses on five areas: analyses that address concepts of area and centrality, analyses of surfaces, analyses that are oriented to design, and statistical inference.

## LEARNING OBJECTIVES

- **Methods for measuring properties of areas.**
- **Measures that can be used to capture the centrality of geographic phenomena.**
- **Techniques for analyzing surfaces and for determining their hydrologic properties.**
- **Techniques for the support of spatial decisions and the design of landscapes according to specific objectives.**
- **Methods for generalizing from samples, and the problems of applying methods of statistical inference to geographic data.**

## KEY WORDS AND CONCEPTS

Central tendency, centroid, minimum aggregate travel (MAT), standard deviation, clustered, dispersed, and random patterns, normative methods, optimization, p-median and coverage problems, spatial interaction modal, shortest path problem, traveling-salesman problem, heuristics, rook's or queen's case, confidence limits, hypothesis testing, randomization tests, slope, aspect, friction surface

## OUTLINE



# 15.1 The Purpose of Area-Based Analyses

- One way in which humans simplify geography of the Earth's infinite complexity is by ascribing characteristics to entire areas rather than to individual points.

### 15.1.1 Measurement of Area

- Refers to the origins of the Canada GIS in the need to accurate measurements of area and mentions the manual dot-counting and planimeter methods.
- Figure 15.1 gives the algorithm for calculation of the area of a polygon

### 15.1.3 Measurement of Shape

- This section talks mostly about gerrymandering and the need for measures of compact shape

# 15.2 Centrality

- Section reviews numerical summaries

### 15.2.1 Centers

- Lists several measures of *central tendency* including mean, median, and mode
- Introduces *centers* which are the two-dimensional equivalent of the mean and lists some of their properties
- The *centroid* or *mean center* is the most convenient way of summarizing the locations of a set of points
    - Found by taking the weighted average of the *x* and *y* coordinates
- The point of minimum aggregate travel (MAT) is the point that minimizes the total straight-line distance from a set of points

- Discusses the stability of MAT given relocation of a point further away on the same line as an illustration of the point that humans can be very poor at guessing the answers to optimization problems in space
- Notes that all methods in this and the next section are based on plane geometry

### 15.2.2 Dispersion

- The measure of choice for numbers with interval or ratio properties is the standard deviation, or the square root of the mean squared difference from the mean
    - Standard deviation and variance are considered more appropriate measures of dispersion than the range (difference between the highest and lowest numbers) because as averages they are less sensitive to the specific values of the extremes.
    - RMSE is a similar measure of dispersion
- A simple measure of dispersion in two dimensions is the mean distance from the centroid

## 15.3 Analysis of Surfaces

### 15.3.1 Slope and Aspect

- Derivative measures such as slope and aspect are defined
- Discusses non-differentiable surfaces and the effects of grid resolution on the calculation of these measures
    - Notes that the spatial resolution used to calculate slope and aspect should always be specified.
- Discusses that there are several alternative measures of slope, and it is important to know which one is used in a particular software package and application.
    - Slope can be measured as an angle and as rise over run
    - Unfortunately there are two different ways of defining run. Figure 14.16 shows the two options, depending on whether run means the horizontal distance covered between two points, or the diagonal distance (the adjacent or the hypotenuse)
- Also, the number of surrounding points used to calculate slope and aspect will vary, as do the weights (see Box 15.2)

### 15.3.2 Modeling Travel on a Surface

- Finding paths across continuous surface easier to solve on a raster as follows
    - Each cell is assigned a friction value equal to the cost or time associated with moving across the cell (the cost layer)
    - A set of allowable moves is selected: rook's case or queen's case (Figure 15.9)
    - Given a cost layer, and a defined origin and destination, a least cost-path is determined

### 15.3.3 Computing Watersheds and Channels

- Describes how a DEM provides an easy basis for predicting how water will flow through a river catchment

### 15.3.4 Computing Visibility

- Describes the process of calculating a viewshed analysis

## 15.4 Design

- This section looks at the analysis of spatial data with the objective of creating improved designs.

### 15.4.1 Point Location

- Explains Hakimi's theorem that states that for the problem of minimizing distance, the only locations that have to be considered are the nodes
- Location-allocation problems which involve where to locate and how to allocate demand to central facilities include
    - The *p*-median problem which seeks optimum locations for any number *p* of central facilities such that the sum of the distances between each weight and the nearest facility is minimized.
    - The coverage problem which seeks to minimize the furthest distance traveled
- Models that make predictions of how consumers will choose among available options are spatial interaction models

### 15.4.2 Routing Problems

- Routing and scheduling involves decisions about the optimum tracks followed by vehicles

    - The shortest path problem finds the path through the network between a defined origin and destination that minimizes distance or some other measure based on distance
    - The traveling-salesman problem (TSP) is to select the best tour out of all possible orderings of places to visit, in order to minimize the distance (or other measure) traveled
    - The orienteering problem is similar to TSP but the objective is to maximize the rewards associated with visiting a selection of the stops while minimizing the total distance traveled
- In these problems, solution methods usually use heuristics which are algorithms designed to work quickly and to come close to the optimum answer

## 15.5 Hypothesis Testing

- Much work in statistics is *inferential* which uses information obtained from samples to make general conclusions about a larger population, on the assumption that the sample came from that population
- Very briefly introduces the concepts *confidence limits* and *hypothesis-testing*

### 15.5.1 Hypothesis tests on geographic data

- Although inferential tests are standard practice in much of science, they are very problematic for geographic data
- Many inferential tests propose the existence of a population from which samples are obtained independently. However,
    - A geographic dataset is often all there is of a given area – it is the population
    - If we could regard a geographic dataset as a sample, the pervasiveness of spatial dependence means samples would not be independent
    - The Earth's surface is heterogeneous, making it difficult to take samples that are truly representative of any large region
- Before using any inferential tests on geographic data, must ask
    - Can I conceive of a larger *population* about which I want to make inferences?
    - Are my data acceptable as a *random* and *independent* sample of that population?
        - If the answer to either of these questions is no, then inferential tests are not appropriate

- R*andomization* tests which simulate a large number of random arrangements of the data offer a good alternative
- Also, new versions of inferential tests that cope effectively with spatial dependence and spatial heterogeneity are now being developed.

## 15.6 Conclusion

### ESSAY TOPICS

1. Parks and other conservation areas have geometric shapes that can be measured by comparing park perimeter length to park area, using the methods reviewed in this chapter. Discuss the implications of shape for park management in the context of a) wildlife ecology and b) neighborhood security.
2. Use tracings of a number of country borders to compute at least one common measure of shape such as the circularity index $S = P/3.54\sqrt{A}$ (in which P=perimeter and A= area) detailed in the text. Other indices are mentioned by O'Sullivan and Unwin (2002, 177-179). What can be concluded from this exercise about the difficulty of measuring area, the dependence of derived measures (such as shape) on such basics, and the discriminatory power of the indices used?
3. You have been hired to suggest optimum locations for five schools in a new town that is being developed. Suggest how you might approach the problem using location-allocation methods.
4. Besides being the basis for useful measures, fractals also provide interesting ways of simulating geographic phenomena and patterns. Browse the Web for sites that offer fractal simulation software, or investigate one of many commercially available packages. What other uses of fractals in GIS can you imagine?
5. 'Shape is an extremely difficult property to measure, or even to define in a precise manner' (Davis, 2002, page 355). Why?

## MULTIPLE CHOICE QUESTIONS (MCQ)

1. Give short definitions of each of the following ‘centers’:
   a) Centroid …….
   b) Mean center ……
   c) Median centrer ……
   d) MAT point ……
2. Rank the following GIS operations in order of their probable computational complexity from 1 (easiest) to 4 (hardest):

| **Operation** | **Rank** |
|---|---|
| *p*-median problem in location-allocation modeling | |
| Finding the point of minimum aggregate travel | |
| Traveling salesman solution for 10 places | |
| Mean center of a point pattern | |

3. Which statement best completes the sentence? “An heuristic is a computational device for…:
   a) Finding correct solutions
   b) Finding solutions that might be correct but we can’t be sure
   c) Speeding up an otherwise lengthy computation
   d) Solving location-allocation problems
4. Write down the two basic approaches to statistical hypothesis testing:
   a) ……..
   b) ……
5. What are the two necessary requirements for a valid sample drawn from a population?
   a) ……..
   b) ……..
6. What percentage of the Earth’s surface that is land has an antipodal point (the point that would be reached by drilling a straight line through the Earth’s center and out to the other side) that is also land?
   a) 25
   b) 50
   c) 2
   d) 18?

## ACTIVITIES

1. Mean centre and standard distance. Davis, J.C. (2002) *Statistics and Data Analysis in Geology*, Wiley: NY, Figure 5.107 page 444, has a map of two bush species, creosote bush and brittlebush, in a 100 x 100m area in southern Arizona. The associated data file is available from the website at www.wiley.com/college/davis, under the file name THERMAL.TXT (the full reference is www3.interscience.wiley.com:8100/legacy/college/davis/0471172758/datafiles/ascii/thermal.txt). There are 190 records and the data has (*x*, *y*) co-ordinates in the range from 0 to 90, with the bush type indicated by a '1' or a '2' in the third column, as in the first 10 items:

| *x* Axis | *y* Axis | Type |
|---|---|---|
| 21.97 | 6.56 | 1 |
| 25.59 | 8.41 | 1 |
| 36.9 | 7.79 | 1 |
| 42.01 | 4.1 | 1 |
| 17.91 | 10.05 | 1 |
| 16.42 | 10.87 | 1 |
| 12.37 | 10.66 | 1 |
| 13.44 | 12.3 | 2 |
| 15.14 | 12.71 | 1 |
| 18.98 | 12.1 | 1 |
| Etc | ... | ... |

a) Visit the website and download the full data file. The data can be imported into a spreadsheet for the analysis;
b) Map the distributions as a dot map;
c) Compute the mean center ('centroid') and standard distance of the two distributions;
d) Do these differ significantly?

These types of measure are best used in applications such as that shown in Figure 15.3 for the mean center of US population from 1790 onwards, where there is a comparison over time or between different types of object.

2. Point pattern analysis: The values below give coordinates for the distribution of 65 Japanese pine tree seedlings in a small 23m by 23m area:

| | | |
|---|---|---|
| 0.09 0.09 | 0.42 0.49 | 0.62 0.97 |
| 0.59 0.02 | 0.37 0.68 | |
| 0.86 0.13 | 0.76 0.66 | |
| 0.42 0.22 | 0.97 0.86 | |
| 0.02 0.41 | 0.29 0.84 | |
| 0.08 0.59 | 0.58 0.83 | |
| 0.31 0.53 | 0.39 0.96 | |
| 0.94 0.58 | 0.39 0.18 | |
| 0.59 0.67 | 0.73 0.13 | |
| 0.94 0.78 | 0.02 0.18 | |
| 0.17 0.95 | 0.73 0.23 | |
| 0.39 0.79 | 0.52 0.42 | |
| 0.36 0.97 | 0.12 0.66 | |
| 0.29 0.02 | 0.52 0.52 | |
| 0.65 0.16 | 0.47 0.67 | |
| 0.89 0.08 | 0.73 0.73 | |
| 0.48 0.13 | 0.12 0.84 | |
| 0.03 0.44 | 0.32 0.83 | |
| 0.08 0.63 | 0.69 0.93 | |
| 0.32 0.52 | 0.43 0.96 | |
| 0.34 0.68 | 0.48 0.03 | |
| 0.66 0.68 | 0.79 0.03 | |
| 0.98 0.79 | 0.11 0.31 | |
| 0.21 0.79 | 0.89 0.23 | |
| 0.52 0.93 | 0.64 0.43 | |
| 0.36 0.96 | 0.17 0.58 | |
| 0.38 0.03 | 0.91 0.52 | |
| 0.67 0.13 | 0.52 0.67 | |
| 0.98 0.02 | 0.89 0.74 | |
| 0.62 0.21 | 0.11 0.94 | |
| 0.07 0.42 | 0.35 0.86 | |
| 0.12 0.63 | 0.77 0.93 | |

These data are used by Diggle, P.J.(1991) *Statistical Analysis of Spatial Point Patterns* (London: Academic) and have been analysed many times (see pages 128-9).

Map them as a dot map and overlay a suitable grid, using it to perform a quadrat census of the sort illustrated by O'Sullivan and Unwin (2010, pages 127-130). Note how the counts might change if a different grid was used, and that an alternative approach would be to generate the proportions by a true sampling process in which use was made of a sequence of, say, 100 randomly located quadrats. Our count for these data gave results as follows:

| **Number of events in quadrat, k** | **Number of quadrats, x** |
|---|---|
| 0 | 49 |
| 1 | 39 |
| 2 | 10 |
| 3 | 2 |

Use the variance/mean ratio test to assess the probability that these events are a realization of the independent random process (O'Sullivan and Unwin, 2010, pages 142-143 provide the details).

3. Choose an example of some planar-enforced area objects such as, for example, the lower 48 states of the USA, or the counties in a state. In doing this, you should find data in which the basic 'geography' of the co-ordinates and topology have already been constructed for you in a format suitable for ArcGIS or similar GIS format. The US Bureau of Census website at www.census.gov/ has such data readily available.
    a. Select some variables that describe characteristics of the people resident in this area. It makes sense to choose maybe 3-4 variables that together help understanding;
    b. Visualize these data using maps and other statistical graphics. It is most probable that you will choose to use the choropleth technique, but this should not rule out other approaches;

Compute and use an appropriate global measure of spatial autocorrelation to verify that the apparent patterns are globally significantly different from independent random

process/complete spatial randomness process (see O'Sullivan and Unwin, 2010, pages 142-143).

   c. Write a brief (less than 500 words) account of what the study shows about the spatial variation in social/economic conditions across the region.

7. Organize a debate on the motion that 'This house believes that statistix inferens has no place in the analysis of geographical data'. Preliminary reading for this can be found in:
   a. Gould, P. R. (1970), Is statistix inferens the geographical name for a wild goose?, *Economic Geography* 46: 439-448.
   b. Harvey, D. W. (1966), Geographical processes and the analysis of point patterns, *Transactions of the Institute of British Geographers* 40: 81-95.
   c. Harvey, D. W. (1967), Some methodological problems in the use of Neyman type A and negative binomial distributions for the analysis of point patterns, *Transactions of the Institute of British Geographers* 44: 81-95.
   d. O'Sullivan and Unwin (2002, pages 53-58) develop the argument that any one map can be treated as a realization of a 'random' process and there is similar material in the excellent text: Stewart Fotheringham, Chris Brunsdon and Martin Charlton (2000) *Quantitative Geography: Perspectives on Spatial Analysis*, Sage: London

7. What exactly are multicriteria methods? Examine one or more of the methods in the chapter referenced below, summarizing the issues associated with a) measuring variables to support multiple criteria, b) mixing variables that have been measured on different scales (e.g., dollars and distances), ) and c) finding solutions to problems involving multiple criteria. Eastman J R 1999 'Multicriteria methods.' In Longley P A, Goodchild M F, Maguire D J, Rhind D W (eds) *Geographical Information Systems: Principles,Techniques, Management and Applications* (abridged edition (2005)). Hoboken, NJ: Wiley.

8. Algorithmics is the study of the basic properties of computer algorithms, often by assessment of their computational complexity, as defined by the relationship between their notional time to completion as a function of the inputs. Thus an algorithm that is complete in direct proportion to the number of inputs has linear time order, denoted O(n), whereas computation of a matrix of all possible distances between objects has order $O(n^2)$ and so on. A good introduction to the topic that examines the traveling salesman and other geographical problems is Harel, D.

(2000) *Computers Ltd: What They Really Can't Do*, Oxford: Oxford University Press. Examine each of the problems outlined in the chapter and attempt to assess its computational complexity.

## FURTHER READING


Davis J C 2002 *Statistics and Data Analysis in Geology*, Wiley: Hoboken, NJ, 3rd edition

Fotheringham A S, Brunsdon C, Charlton M 2000 *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: Sage.

Fotheringham A S, O'Kelly M E 1989 Spatial *Interaction Models: Formulations and Applications*. Dordrecht: Kluwer.

Ghosh A, Rushton G (eds) 1987 *Spatial Analysis and Location-Allocation Models*. New York: Van Nostrand Reinhold.

Mardia K V, Jupp P E 2000 *Directional Statistics*. New York: Wiley.


## 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.

35. Multi-criteria evaluation and GIS, J R Eastman

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).

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

27. Knowledge-based approaches in GIS, T R Smith and Ye Jiang, pp. 413-25


## ONLINE RESOURCES

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

Module 1: Basics of Data and Information

Module 4: Transformations and Descriptive Summaries

Module 5: Optimization and Hypothesis Testing

Section 15.2, 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: Centers and dispersion

Section 15.2.1, Module 4: Transformations and Descriptive Summaries

Unit: Centers and dispersion

Sub-unit: Centers

Sub-unit: The Varignon Frame experiment

Section 15.2.2, Module 4: Transformations and Descriptive Summaries

Unit: Centers and dispersion,

Sub-unit: Dispersion

Section 15.2.5, Module 4: Transformations and Descriptive Summaries

Unit: Spatial dependence and fragmentation

Sub-unit: Fragmentation

Section 15.3, Module 1: Basics of Data and Information

Unit: Creating and visualizing information

Sub-unit: Types of spatial analysis

Module 5: Optimization and Hypothesis Testing

Unit: Optimization

Section 15.3.1, Module 5: Optimization and Hypothesis Testing

Unit: Optimization

Sub-unit: Point location

Section 15.3.2, Module 5: Optimization and Hypothesis Testing

Unit: Optimization,

Sub-unit: Routing problems

Section 15.3.3, Module 5: Optimization and Hypothesis Testing

Unit: Optimization,

Sub-unit: Optimum paths

Section 15.4, Module 1: Basics of Data and Information

Unit: Creating and visualizing information,

Sub-unit: Types of spatial analysis

Module 5: Optimization and Hypothesis Testing

Unit: Hypothesis testing

Section 15.4.1, Module 4: Transformations and Descriptive Summaries

Unit: Spatial dependence and fragmentation,

Sub-unit: Spatial dependence

Module 5: Optimization and Hypothesis Testing

Unit: Hypothesis testing,

Sub-unit: Hypothesis tests on geographic data

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

2.14.1. Spatial Decision Support Systems (127), Jacek Malczewski

2.14.6. Artificial Neural Networks for Spatial Data Analysis (188), Suchi Gopal

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

57. Multiple criteria methods

58. Location-allocation on networks

59. Spatial decision support systems

74. Knowledge based techniques