6 Uncertainty

OVERVIEW

- Uncertainty in geographic representation arises because almost all representations of the world are incomplete.
- This chapter identifies many of the sources of geographic uncertainty and the ways in which they operate in GIS-based representations.
- Uncertainty arises from the way that GIS users conceive of the world, how they measure and represent it, and how they analyze their representations of it.
- This chapter investigates a number of conceptual issues in the creation and management of uncertainty, before reviewing the ways in which it may be measured using statistical and other methods.
- The propagation of uncertainty through geographical analysis is considered.

LEARNING OBJECTIVES

- **Understand the concept of uncertainty, and the ways in which it arises from imperfect representation of geographic phenomena;**
- **Be aware of the uncertainties introduced in the three stages (conception, measurement and representation, and analysis) of database creation and use;**
- **Understand the concepts of vagueness and ambiguity, and the uncertainties arising from the definition of key GIS attributes;**
- **Understand how and why scale of geographic measurement and analysis can both create and propagate uncertainty**

KEY WORDS AND CONCEPTS

Uncertainty, error, ambiguity, vagueness, data quality, fuzzy membership, accuracy, precision, RMSE, Gaussian or Normal distribution, error propagation, aggregation, simulation, ecological fallacy, concatenation, conflation, induction, deduction, MAUP

OUTLINE

CHAPTER SUMMARY

This chapter is very dense in places. Many students will need extra mentoring to make it through the thorough descriptions of the effect of error and the implications of various reported error measurements. Alternatively, instructors of introductory courses may choose to use the 'U1, U2, U3' schema in order to provide a selective overview of the principal issues.

6.1 Introduction

- This chapter uses the term *uncertainty* as an umbrella term to describe the problems that arise out of our incomplete representations of the world.
- Various terms are used to describe differences between the real world and how it appears in a GIS
- The established scientific notion of measurement *error* focuses on differences between observers or between measuring instruments.
- *Ambiguity* and *vagueness* identify further considerations which need to be taken into account in assessing the *quality* of a GIS representation.
- The US Federal Geographic Data Committee's various standards list five components of quality: attribute accuracy, positional accuracy, logical consistency, completeness, and lineage.

- Uncertainty may thus be defined as a measure of the user's understanding of the difference between the contents of a dataset, and the real phenomena that the data are believed to represent.
- In GIS, the term uncertainty has come to be used as the catch-all term to describe situations in which the digital representation is simply incomplete, and as a measure of the general quality of the representation.
- The chapter structures the discussion of uncertainty through a consideration of the chain of events in which *conception* prescribes *measurement and representation*, which in turn prescribes *analysis*. This is summarized in Figure 6.1.

6.2 U1: Uncertainty in the conception of geographic phenomena

A characteristic that sets geographic information science apart from most every other science is that it is only rarely founded upon *natural* units of analysis.

6.2.2 Conceptions of attributes: Vagueness and ambiguity

6.2.2.1 Vagueness

- Given the lack of natural units of analysis, we often transform point-like events into area objects. This leads to two important questions
    - Is the defining boundary of a zone crisp and well-defined?
    - Is our assignment of a particular label to a given zone robust and defensible?
- Thus, uncertainty can exist both in the positions of the boundaries of a zone and in its attributes.
- The questions have statistical implications, cartographic implications, and cognitive implications.
- Box 6.1 introduces school catchments as functional zones

6.2.2.2 Ambiguity

- Many linguistic terms used to convey geographic information are inherently ambiguous.
- Many objects are assigned different labels by different national or cultural groups, and such groups perceive space differently.
- Object names and the topological relations between them may thus be inherently *ambiguous*.
- GIS cannot present a value-neutral view of the world, yet it can provide a formal framework for the reconciliation of different worldviews
- Ambiguity is introduced when imperfect *indicators* of phenomena are used instead of the phenomena themselves.

- *Direct* indicators are deemed to bear a clear correspondence with a mapped phenomenon.
- *Indirect* indicators are used when the best available measure is a perceived surrogate link with the phenomenon of interest.
- Conception of the linkage between any indicator and the phenomenon of interest is subjective, hence ambiguous.
- Our ability to generalize about spatial distributions is constrained by the different taxonomies that are conceived and used by data-collecting organizations within our overall study area.
- How may mismatches between the categories of different classification schema be reconciled?
- The process of reconciling the semantics of different classification schema is an inherently *ambiguous* procedure

Applications Box 6.2 Vagueness, ambiguity, and the geographies of family names

- Provides an interesting discussion of how surnames can be used as indicators of regional identity and diversity

6.2.3 Fuzzy approaches to attribute classification

- *Frequentist* approaches to assigning values to areas are based on the notion that the probability of a given outcome can be defined as the proportion of times the outcome occurs in some real or imagined experiment, when the number of tests is very large.
- However, in the geographic situation, there is only one field with precisely these characteristics, and one observer
- The *subjectivist* conception of probability represents a judgment about relative likelihood of a single occurrence and is best illustrated through the concept of *fuzzy membership*
- One of the major attractions of fuzzy sets is that they appear to let us deal with sets that are not precisely defined, and for which it is impossible to establish membership cleanly.
- Box 6.2 (Fuzziness in classification: description of a soil class) shows a typical extract from the legend of a soil map, including frequent use of terms such as ‘very’, ‘moderate’, ‘about’, ‘typically’, and ‘some’.
    - Figure 6.8 shows an example of mapping classes using fuzzy methods. The final map shows how these can be converted to crisp categories
- Researchers have struggled with the question of whether fuzzy methods are more *accurate*.

- If we are uncertain about which class to choose then it is more accurate to say so, in the form of a fuzzy membership, than to be forced into assigning a class without qualification.

6.3 U2: Further uncertainty in the representation of geographic phenomena

6.3.1 Representation of place / location

- The conceptual models (fields and objects) impose very different filters upon reality, and their usual corresponding representational models (raster and vector) are characterized by different uncertainties
- The vector model requires *a priori* conceptualization of the nature and extent of geographic individuals and the ways in which they nest together into higher-order zones.
    - In the vector model, point-like objects often appear only as aggregate counts for apparently *uniform* zones.
- The raster model defines individual elements as square cells, with boundaries that bear no relationship at all to natural features
- Discusses mapping coastline as a field
- Introduces the concept of *mixel*, a pixel whose area is divided among more than one class

6.3.2 Statistical models of uncertainty in attribute measures

A geographic database is a collection of measurements of phenomena on or near the Earth's surface

6.3.2.1 Nominal case

- Describes the structure and interpretation of the *confusion matrix*
- Provides the equation for the *kappa index*
- Identifies some problems with this method
- Notes that in vector area model cases, error has two forms: misallocation of an area's class and the mislocation of an area's boundary

6.3.2.2. Interval/ratio case

- Here, error is best thought of not as a change of class, but as a change of value such that the observed value x is equal to the true value x plus some distortion δx.

- If the average distortion is zero, with positive and negative errors balanced out, the observed values are said to be *unbiased*
- Distinguishes between *accuracy*, which has to do with the magnitude of the error, and *precision* which is defined in two ways
    - The variability among repeated measurements
    - The number of digits used to report a measurement (see Technical Box 6.4 that summarizes rules that are used to ensure reported measurements do not mislead)
- Discusses in detail the calculation and relevance of RMSE
- Explains the structure and interpretation of the Gaussian (or Normal) probability distribution

6.3.3 Statistical models of uncertainty in location measures

- This section is particularly detailed in its examination of the implications of positional error in our spatial databases. Many students will need extra mentoring to understand this section.
- A two-dimensional measured position (x,y) is subject to errors in both *x* and *y*
- In three dimensions, we expect the RMSEs of *x* and *y* to be the same, but *z* is often subject to errors of quite different magnitude.
- National Map Accuracy Standards often prescribe the positional errors that are allowed in databases.
- The 1947 US National Map Accuracy Standard specified that 95% of errors should fall below 1/30 inch (0.85 mm) for maps at scales of 1:20,000 and finer (more detailed), and 1/50 inch (0.51 mm) for other maps
- A useful rule of thumb is that features on maps are positioned to an accuracy of about 0.5mm. Table 6.2 shows the corresponding distance on the ground at different scales

6.4 U3: Further uncertainty in the analysis of geographic phenomena

6.4.1 Internal and external validation through spatial analysis

- *Internal* validation can be achieved through simulation of different possible outcomes (i.e. error propagation)
- *External* validation can be achieved by merging diverse data sources

6.4.2 Validation through autocorrelation: The spatial structure of errors

- Error propagation measures the impacts of uncertainty in data on the results of GIS operations.
- There are two strategies available for evaluating error propagation
    - Obtain a complete description of error effects based upon known measures of likely error. This is discussed in detail by the use of some examples
    - Simulate the impacts of uncertainty on results which requires the generation of a series of realizations

6.4.3 Validation through investigating the effects of aggregation and scale

- The measurement of geographic individuals is unlikely to be determined with the end point of particular spatial analysis applications in mind.
- As a consequence, we cannot be certain in ascribing even dominant characteristics *of* areas to true individuals or point locations *in* those areas.
- This source of uncertainty is known as the *ecological fallacy* (inappropriate inference from aggregate data about the characteristics of individuals)
- Gives rise to the related *aggregation* or *zonation* problem, in which different combinations of a given number of geographic individuals into coarser-scale areal units can yield widely different results.
- The effects of scale and aggregation are generally known as the *Modifiable Areal Unit Problem* (MAUP).

6.4.4 Validation with reference to external sources: Data integration and shared lineage

- *Concatenation* is used to describe the integration of two or more different data sources, such that the contents of each are accessible in the product.
- *Conflation* attempts to replace two or more versions of the same information with a single version that reflects the pooling, or weighted averaging, of the sources.
- Yet such different datasets are likely to have been collected at a range of different scales and for a range of areal units
- Established procedures of statistical inference can only be used to reason from representative samples to the populations from which they were drawn.

6.4.5 Internal and external validation; induction and deduction

- This section provides several areas of caution that need to be considered
    - The Modifiable Areal Unit Problem can be investigated through simulation of large numbers of alternative zoning schemes.

- However, zone design experiments are merely playing with the MAUP, and most of the new sources of external validation are unlikely to sustain full scientific scrutiny, particularly if they were assembled through non-rigorous survey designs.
- In measuring the distribution of all possible zonally averaged outcomes, there is no tenable analogy with the established procedures of statistical inference and its concepts of precision and error.
- The way forward seems to be to complement our new-found abilities to customize zoning schemes in GIS with external validation of data and clearer application-centered thinking about the likely degree of within-zone heterogeneity that is concealed in our aggregated data.
- Notes that within the socio-economic realm, the act of defining zones can also be self-validating if the allocation of individuals affects the interventions they receive

6.5 Consolidation

- Briefly lists the key points made
- Gives some rules for how to live with uncertainty
    - Acknowledge that uncertainty is inevitable
    - Assemble all that is known about the quality of data and use this to assess whether the data are fit for use
    - Gain some impression of the impacts of input uncertainty on outputs
    - Rely on multiple sources of data
    - Be honest and informative in report the results of GIS analysis

ESSAY TOPICS

1. Why are error and uncertainty in the results of a GIS-based analysis not the same thing?
2. Review the ways by which continuous fields can be represented in a GIS.
3. Giving specific examples, explain what is meant by the term ‘ecological fallacy’, how it arises and why it can lead to false conclusions.
4. In assembling objects such as trees into area objects such as a ‘forest’ what are the major characteristics that would be required of these areas? (see Section 6.2.2.1)
5. Outline and contrast the available methods for evaluating error in ‘field’ data with those for object-based representations.
6. Explain why soil classes are archetypal examples of fuzzy objects with uncertain boundaries.

7. Outline a justification for the assertion that 'Modifiable areal units aren't just a technical issue for GI science: they have profound implications for society'.
8. According to at least one GIS authority combinations of maps "... unravel more questions about data quality and boundary mismatch than they solve" (Leung Y and Leung K S (1993) An intelligent expert system shell for knowledge based geographical information systems 1. The tools. *International Journal of Geographic Information Systems*, 7(3): 189-99, page 193). Why?

MULTIPLE CHOICE QUESTIONS (MCQ)

1. Arrange the following notions into the order in which they would commonly be addressed by a GIS analyst:
    a. Representation
    b. Analysis
    c. The real world
    d. Conception
    e. Measurement
2. For each of the definitions given state whether it is of 'error', 'accuracy' or 'precision':

Definition	**Concept?**
Difference between observers or between instruments	
Difference between reality and our representation of reality	
Number of decimal digits in a measurement	

3. Section 6.2.3 contrasts uncertainty about clearly defined objects, handled by probability theory with 'fuzziness' in the objects themselves. Say whether each of the following statements is neither uncertain nor fuzzy, probabilistic, fuzzy or both:

Statement	**Status**
Joe Lobley is about 65	
I think maybe that Joe Lobley is about 65	
Joe Lobley is 65	
I think maybe that Joe Lobley is 65	

4. Which of the following are NOT ways of representing a field variable?

a. Triangulated network
b. Contours
c. Planar enforcement
d. Elevation matrix
e. Pixel

5. A digital elevation matrix set has a claimed RMSE of 2.0m. If the matrix is 100 by 100, how many of the height values might be expected to be in error by over this amount?
    a. 6,800
    b. 3,200
    c. 500

6. If a variable can be assumed to be normally distributed, with sample mean of 12.34 and best estimate of the standard deviation of 3.44, what percentage of all sample observations can be expected to have values in the range 8.90 to 15.78?
    a. 34
    b. 95
    c. 68
    d. 5

7. Round the following numbers to the specified number of decimal places:

Number	Decimal places	Answer
323.55	0	
25.05	1	
45.7896	2	
34.4999	0	

8. Rate each of the following area objects on a scale of ambiguity that goes from ‘high’ though ‘medium’ to ‘low’:

Object	Ambiguity Rating

Wetland	
Oak forest	
Urban park	
Rain area	
Fog area	
Census tract	
State or County	

CLASS AND INDIVIDUAL ACTIVITIES

1. What tools do GIS designers build into their products to help users deal with uncertainty? Take a look at your favorite GIS from this perspective. Does it allow you to associate metadata about data quality with datasets? Is there any support for propagation of uncertainty? How does it determine the number of significant digits when it prints numbers? What are the pros and cons of including such tools?
2. Visit http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/DLGs and examine the US Geological Survey Digital Line Graph User Guide. Discuss and debate its adequacy in the light of the issues raised in this Chapter.
3. Find out about the five components of data quality used in GIS standards, from the information available at www.fdgc.gov. How are the five components applied in the case of a standard mapping agency data product, such as the US Geological Survey's Digital Orthophoto Quarter–Quadrangle program?
4. Imagine that you are a senior retail analyst for *Safemart*, which is contemplating expansion from its home US state to three others in the Union. Assess the relative merits of your own company's store loyalty card data (which you can assume are similar to those collected by any retail chain with which you are familiar) and of data from the 2000 Census in planning this strategic initiative. Pay particular attention to issues of survey content, the representativeness of population characteristics, and problems of scale and aggregation. Suggest ways in which the two data sources might complement one another in an integrated analysis.
5. The NCGIA Core Curriculum in GI Science has a Unit by Veregin that lists eight components of data quality (spatial accuracy, temporal accuracy, thematic accuracy, spatial resolution, temporal resolution, thematic resolution, consistency and completeness), see www.ncgia.ucsb.edu/education/curricula/giscc/units/u100/u100_f.html. Explore some digital data sets known to you and assess each data set on an ordinal scale good/average/poor on each component. Summarize your results by creating a table

in which the columns represent each of these components and rows represent each dataset.

6. Illustrate the effects of aggregation on basic descriptive statistics using the following 8 by 8 ‘raster’ of cell values:

87	95	72	37	44	24	24	45
40	55	55	38	88	34	24	33
41	30	26	35	38	24	16	12
14	56	37	34	8	18	9	13
49	44	51	67	17	37	45	63
55	25	33	32	59	54	54	76
58	56	37	24	45	58	64	87
70	67	44	34	13	46	67	88

   a. First, compute the mean and standard deviation of the entire set of 64 values. Next, create a 4 by 4 matrix by amalgamating sets of four neighboring cells, replacing the four cell values by their arithmetic mean, and repeat the operation using these 16 values. Finally, repeat again, re-amalgamating to create a 2 by 2 grid. Discuss the answers to the following questions:
   b. Why doesn’t the mean change?
   c. What is the effect on the standard error of the mean and why?
   d. What is the effect on the range of values and the maximum and minimum values?
   e. How do the variance and standard deviation change?
   f. What does this exercise tell you about the nature of geographical data aggregates?
   g. How do you think this effect will influence the results of any analysis using aggregated data?

7. One of the major problems with the idea of ‘quality’ lies in its definition, with at least four possible approaches based on
   a. minimum quality, as with the US National Map Accuracy Standard;
   b. conformance to an explicit metadata description, sometimes called ‘truth in labeling’ or ‘what it says on the can’, as with the US Spatial Data Transfer Standard;
   c. market feedback
   d. fitness for purpose

Organize four teams and ask them to prepare and present a case for the use of each of these approaches for spatial data. This exercise can be given a vaguely competitive edge by having an independent adjudicator.

8. Cohen's Kappa index is a standard measure of between ratings reliability in, for example, educational research. Its use in assessing attribute accuracy in classified satellite imagery is described at the Geographer's Craft website: http://www.colorado.edu/geography/gcraft/notes/manerror/html/attribut.html. It can be introduced either by using a classified image of your locality for which ground truth data are available, or by asking students to perform a classification at the same set of grid points and to compare their estimates of the preponderance of each land use class.

9. In today's marketplace, it is probable that there will be many possible sources of data for any specific GIS analysis, so that it is important to become what the *Geographer's Craft* (see website below) team call a 'smart shopper'. Imagine you are a GI consultant analyst, tasked to develop a GIS for a specific project. Use the Web to find the data that you would use, then justify your selection. A useful series of questions that a 'smart data shopper' might ask are given at the following website under the heading 'What to Look For and When to Quit': http://www.colorado.edu/geography/gcraft/notes/sources/sources_f.html. The topic of the project and its region of application can be varied by the instructor, but outside of US care needs to be taken to ensure that a sufficiently rich list of data sets is available.
10. A geodemographic classification of the people living in some small zone is one of the most commonly used data source in business use of GIS. These classifications essentially collapse a very large number of variables in a single nominal category that is in some sense held to characterize that zone, such as 'laptops and latte' or 'empty nesters'. Examine at least two of these products and list the sources and likely magnitude of the uncertainties that their use will introduce into any analysis using them.

FURTHER READING

Burrough P.A. and Frank A.U. (eds) 1996 Geographic Objects with Indeterminate Boundaries. London: Taylor and Francis.

This edited collection is the classic book on uncertainty in area objects.

Unwin, D.J. (1995) Geographical information systems and the problem of error and uncertainty. *Progress in Human Geography*, 19(4): 549-558.
Develops the idea of an error sensitive GIS.

Heuvelink G.B.M. 1998 Error Propagation in Environmental Modelling with GIS. London: Taylor and Francis.
It is almost certain that many of the issues raised in this Chapter will only be solved by rigorous use of concepts from geostatistics. The Dutch mathematician-geographer Gerard Heuvelink has been a pioneer in this effort, and his book should be essential reading for all serious GI analysts.

Zhang J.X. and Goodchild M.F. 2002 Uncertainty in Geographical Information. New York: Taylor and Francis.
Very much the standard work in which the idea of uncertainty is developed in considerable detail.

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.

13. Models of uncertainty in spatial data, P F Fisher, pp. 191–205. Much of the research on uncertainty is reviewed in this chapter.
18. Applying geocomputation to the analysis of spatial distributions, S Openshaw and S Alvanides, pp. 267–282. If geostatistical theory is incapable of giving analytical solutions, then simulation of the uncertainty in results may well prove to be the best way of addressing the issue.
40. The future of GIS and spatial analysis, M F Goodchild and P A Longley, pp. 567–80

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).

11. Language issues for GIS, A U Frank and D M Mark, pp. 147-63
12. The error component in spatial data, N R Chrisman, pp. 165-74
24. Spatial data integration, R Flowerdew, pp. 375-87

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 6: Uncertainty

Section 6.1, Module 1: Basics of Data and Information

Unit: Uncertainty

Section 6.2, Module 1: Basics of Data and Information

Unit: Uncertainty,

Sub-unit: Uncertainty in the conception of geographic phenomena

Section 6.2.3, Module 6: Uncertainty

Unit: Uncertainty issues for spatial data

Sub-unit: Fuzzy approaches

Section 6.3, Module 1: Basics of Data and Information

Unit: Uncertainty

Sub-unit: Uncertainty in the measurement of geographic phenomena

Section 6.3.2, Module 6: Uncertainty

Unit: Measuring uncertainty of nominal and ordinal values

Section 6.3.4, Module 6: Uncertainty

Unit: Uncertainty issues for spatial data,

Sub-unit: The spatial structure of errors

Section 6.4, Module 1: Basics of Data and Information

Unit: Uncertainty

Sub-unit: Uncertainty in the analysis of geographic phenomena

Section 6.4.2, Module 6: Uncertainty

Unit: Uncertainty issues for spatial data

Sub-unit: Error propagation

Section 6.4.4, Module 6: Uncertainty

Unit: Measuring uncertainty of interval or ratio values,

Sub-unit: Uncertainty in spatial data

Section 6.4.5, Module 6: Uncertainty

Unit: Uncertainty issues for spatial data,

Sub-unit: Living with uncertainty

Section 6.5, Module 6: Uncertainty

Unit: Uncertainty issues for spatial data

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

2.10. Handling uncertainty (096), ed. *Gary Hunter* (see also *GC notes)*

2.10.1. Managing Uncertainty in GIS (187), *Gary Hunter*

2.10.2. Uncertainty Propagation in GIS (098), *Gerard Heuvelink*

2.10.3. Detecting and Evaluating Errors by Graphical Methods (099), *Kate Beard*

2.10.4. Data Quality Measurement and Assessment (100), *Howard Veregin*

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

45. Accuracy of spatial databases

46. Managing error