CHAPTER 21

A Comprehensive Model of Uncertainty in Spatial Data

J. Gottsegen, D. Montello, and M. Goodchild

INTRODUCTION

There is a developing interest in the problem of uncertainty as compared to accuracy or error in spatial data (Unwin, 1995). The notion of uncertainty is broader than error or accuracy and includes these more restrictive concepts. While accuracy is the closeness of measurements or computations to their "true" value or some value agreed to be the "truth" (Unwin, 1995), uncertainty can be considered any aspect of the data that results in less than perfect knowledge about the phenomena being modeled. Thus it is a statement of doubt and distrust in results and is a form of "unknowing" (Thrift, 1985). Researchers of uncertainty in nonspatial data have sometimes used the term "imperfect" data (Motro, 1997).

Uncertainty in data also encompasses data quality described as "fitness for use" by Chrisman (1982). While the terms error and accuracy connote judgments of the appropriateness or usability of data based solely on the comparison of data values to some other set of values, data quality begins to consider the needs of the data user as important in determining the adequacy of data. Issues such as scale, level of aggregation, or classification scheme are critical in a user's assessment of whether a particular set of data is useful for a given task.

This chapter represents an effort to develop a comprehensive view of uncertainty in spatial data. It considers the user a critical component in the definition of uncertainty, and it identifies the processes in the use and creation of data that may contribute to uncertainty. It presents a general model of uncertainty as a framework for this identification and as a possible basis for relating the work being done on the assessment and management of uncertainty in data.

This chapter proceeds by first considering the objectives of defining a conceptual model of uncertainty. It then discusses some of the specific aspects of spatial data development and data use that lead to uncertainty. It then presents the uncertainty model, and identifies where literature in the field fits into the model. It concludes with a consideration of insights gained from the model and the research suggested by the model.

REQUIREMENTS OF A GENERAL MODEL OF UNCERTAINTY

A model of uncertainty should serve several functions. It should expose potential sources of uncertainty and provide ideas for managing or addressing these various sources of uncertainty. This includes identifying sources of uncertainty that cannot be managed. It should also serve as a framework for designing empirical research that can inform efforts resulting from the first function. Finally, it can be used as a benchmark to assess how spatial data researchers are progressing in addressing the topic of uncertainty.

A model of uncertainty needs to address all of the aspects of the use and development of spatial data that constrain a user's knowledge of appropriate uses for the data and the phenomena represented by the data. Such a model must represent the relationship between the data user's knowledge of the world and the data or information maintained by a system (Smets, 1997). It

must also integrate the well-accepted definition of data quality as "fitness for use" and the ideas of accuracy and error that have been the subject of considerable research over the past several years. For example, uncertainty in the use of spatial data is partially inversely related to spatial data quality, but it can be considered a distinct concept. That is, high quality spatial data implies a degree of knowledge (i.e., low degree of uncertainty) about the data, the methods used to create it, and its underlying data model or characteristics. The converse is not necessarily true, however. One may have a great deal of metainformation about data that are of low quality simply because they are inappropriate for a certain use. Similarly, accuracy and error have a component that relates directly to uncertainty. The random component of error is, by definition, impossible to know and measure. At best it can be estimated or inferred based on some hypothesized model of its distribution. Bias is deviation from a benchmark that is known to follow a pattern. Hence if error is composed only of bias, there is no uncertainty in terms of the locations (or other measurements) comprising the data.

The model of uncertainty should also describe aspects of a user's query, retrieval, and decision-making process that result in a loss of information about the appropriateness of a given use for particular data. This is the component of uncertainty that is often neglected in the spatial data accuracy literature. However, if one talks about uncertainty in spatial data or the use of spatial data, one must consider the person who may or may not be certain. Another way to view this is that a data user inherently assumes some risk of "incorrect" results from using data. An acceptable level of uncertainty can be considered the amount of this risk that a decision-maker is willing to accept. Information that reduces uncertainty reduces this risk or makes it identifiable (Stinchcombe, 1990). The ability to assess fitness for use implies not only sufficient knowledge about the data, but also a well-defined conception of the use to which the data is applied and the validity of the constructs underlying the application of the data. For example, are the methods for assessing ecosystem health valid? This includes a detailed understanding of how the data that are queried and used represent or correspond to the phenomena that the user is interested in analyzing.

Both of the components described above result from the fact that spatial data and access to them entail formalizations and abstractions. A user must distill her interest in ecosystem health to a set of structured relationships between specific measurements. Additionally, to retrieve data to analyze system health, she must produce a set of discrete valued conditions. Of course, spatial data is also a discrete representation of a model of given phenomena.

The model must also include a consideration of the representation of uncertainty. While considerable research in GIS has concentrated on various techniques of representing error in spatial data, most of it has proceeded based on the assumption that such representation is valuable. It has not developed any consistent guiding principles about how and why such representation would be useful to the user of spatial data. The efficacy of the representation depends on the process by which a data user defines the query or information request, evaluates the appropriateness of the data for a particular use, and uses the data. The interpretation of a representation depends on the user's map schema (MacEachren, 1995). Certain types of representations may facilitate specific decisions to a greater degree, and the user may engage different decision-making methods or heuristics in response to the portrayals of error. In addition, data representations themselves introduce potential uncertainty.

SPECIFIC COMPONENTS OF UNCERTAINTY

Given the comments above, uncertainty in spatial information and its use has several aspects:

1. Uncertainty includes the degree to which the data representation differs from the world or some higher quality representation of it. This may be identifiable or measurable or not. This encompasses the traditional definition of error. The deviation of a representation from reality may be due to (a) measurement error (bias and random), (b) insufficient knowledge to measure a precisely defined concept (e.g., map precisely bounded categories), (c) concepts that cannot be precisely defined, and (d) precision limitations of the measurement or storage device.
2. Uncertainty also relates to the compatibility and consistency between the formally specified need of the user (i.e., data query or information retrieval request) and the data representation. This includes (a) lexical (naming) differences, (b) semantic differences (different associations or meanings for same terms), (c) differences in classification, and

(d) geometric and topological characteristics of the data (e.g., scale, resolution).

3. Uncertainty also depends on the match between the formal specification of a user's query or information request, her subsequent use of the data, and the phenomena of interest. This notion is similar to the idea of construct validity in social sciences (Rosenthal and Rosnow, 1991). That is, do the data and relationships between them specified by the user adequately represent the phenomenon of interest. In some cases where the phenomenon of interest is complex, such as ecosystem modeling, it is extremely difficult to identify and specify the important relationships.

In a sense, uncertainty in data use is the degree to which the formalized structure of data is incompatible with the concepts that a data user is trying to analyze. Together the three aspects listed above influence the applicability of data for a certain use. Other considerations can add to the uncertainty of decisions made with the data, for example, the robustness of the application in which the data is used. However, these do not pertain to uncertainty in spatial data per se.

THE MODEL

The following model describes the different components mentioned above in detail. It identifies several steps in the formalization of user's conceptions of an issue into a query and in the formalization of a conception of a phenomenon into a digital spatial data set. Understanding these steps allows us to see where information is lost or changed in this process. This loss of or change in the information results in the mismatches described above.

The general structure of the model is shown in Figure 21.1. It has two sides, one representing the data user and one representing the data development component and an arrow indicating data representation. The data user and development components begin with a person's conceptualization of the world and proceed to a formal specification that is entered into a computer either as a query or a data set. The database system or query processor matches the two specifications and returns the query results. Each of these components involves a sequence of transformations from less to more formal specifications, and each transformation can entail a potential loss, alteration, or creation of information. The details of the transformations for the data user and development components are shown in Figures 21.2 and 21.3, respectively. The result of these transformations is that a perfect match of formal specifications may not represent a close match to conceptualizations of a phenomenon.

Data User

The process of choosing which specific set of data is required for a given use involves a continual refinement from a rough identification of a problem or question to a formal specification of the specific data elements of interest. The rough identification of the problem is often simply the recognition that a problem or question is important. The formal specification is often a query in a specific query language.

Step 1

The user of spatial data often begins with a perceived problem or question to be addressed. This may be in the form of a query (e.g., "I wonder where wetlands are in this county?"), or it may be a more complex need, such as assessing the impacts of land-use change on riparian habitats. In either case, the problem or question begins as informal perception of a need.

Step 2

After perceiving a need for spatial data, the user must formulate a conceptualization of the problem in more detail and with greater structure. This formulation begins to specify the parameters of the decision and therefore the necessary aspects of data used for it. For example, in our case of the hypothetical land-use change assessment, the potential data user will start identifying the possible types and magnitudes of change that may be detrimental. She will also identify the spatial characteristics of these changes that are important (e.g., distance from sensitive habitat, pattern of changes, intervening land uses between the habitat and land-use change, etc.).

Step 3

From the formulation of the problem, the data user then identifies the type of data and the aspects of them that she is interested in (e.g., attributes, classification, spatial and attribute resolution, etc.). This eventually results in an informal expression of a data query. That

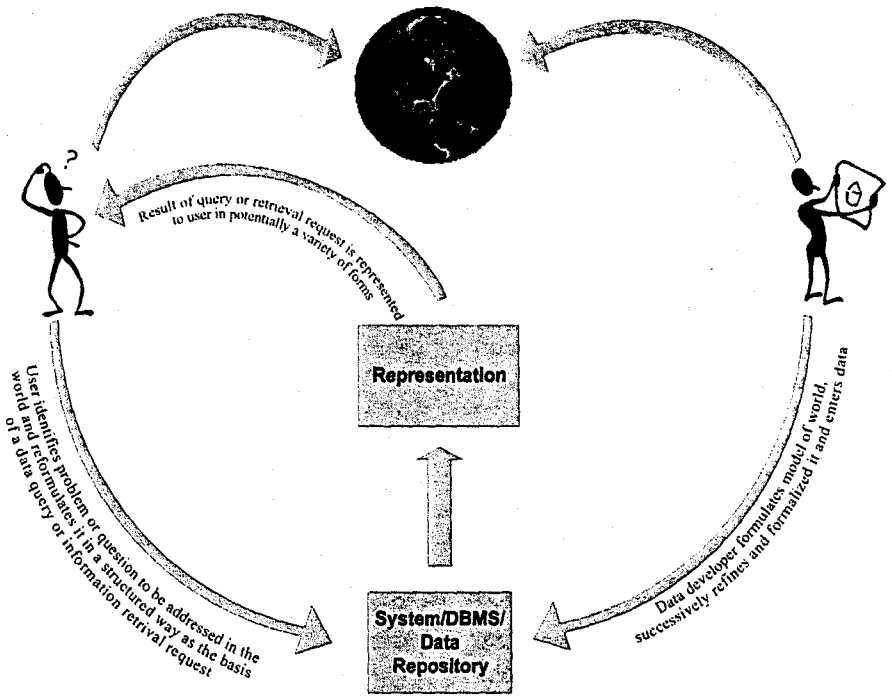

Figure 21.1. General model structure.

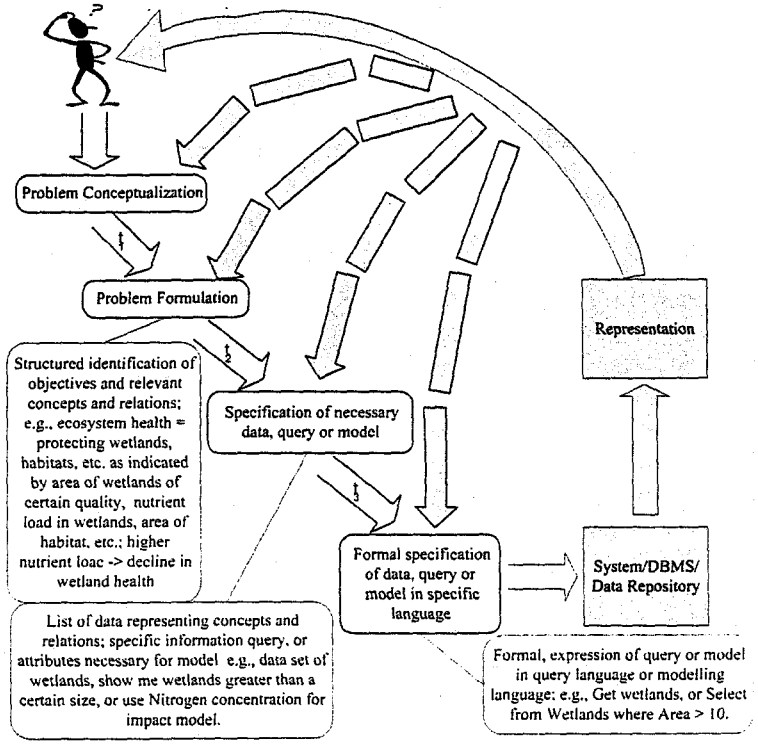

Figure 21.2. Data user component.

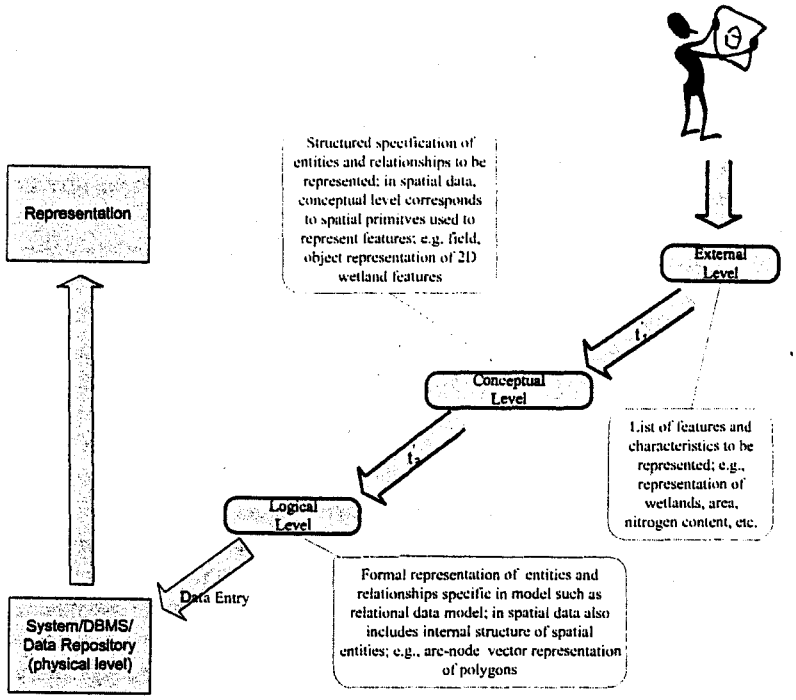

Figure 21.3. Data development component.

is, it is an informal specification of the information that the user requires from the database.

Step 4

The informal specification of the required information leads to a formal query of a database. This can be a query of a data repository where the desired result is a data set, or it can be a query of an individual data set where the desired result is a set of features that match a given condition. Of course, this query is most often expressed in a formal query language.

Data Development

The steps in this component of the model correspond to the data modeling process outlined by many authors (Peuquet, 1984; Ullman, 1988). The process of developing a data set entails defining a data model on which the data set is based. This begins as a fairly informal decision about the features to be represented in the data set and moves to a rigorous, formal specification of the way features and their relationships are represented and stored.

Step 1

The first step in developing a data set is deciding what features and what information about them should be included in the data set. This is often known as the external data model. Thus a data developer may decide that land cover classified in a specific way (e.g., Anderson level I) is important or that streams and stream order numbers are important. In addition, a data developer may also decide that the data set should represent connectivity between stream reaches.

Step 2

The data developer has to make a decision about how the data should be represented in a data set. This is known as the conceptual data model. For example, will land cover be represented as a set of continuous objects or a field of land cover? There are other possibilities, but the field-object differentiation has occupied much of the literature in spatial data modeling (Couclelis, 1992; Goodchild, 1992). We will not dwell on this difference at this point since the reader may peruse one of the papers that treats this difference in detail.

Step 3

The conceptual model is then transformed into a computational or logical model. This specifies the types of data structures that will be used to store the information required. This includes models such as the georelational model used in Arc/Info or others. This logical model determines the input of data into a spatial database. In some cases, it also determines the method of measurement or data collection; however, in other cases it may be partly determined by the data collection method used.

Step 4

As mentioned above, the next step is measurement or data input. The notion of accuracy is the comparison of the results of this process to the world. Thus accuracy must incorporate random and systematic sources of differences between the data and the actual characteristics of the features.

Data Representation

The results of a query may be represented to the user in a variety of forms or modes. Clearly, the most common mode is a cartographic representation or textual representation in the form of a table. This representation is then the basis of either decisions or analyses or of refining or altering the initial request or problem formulation. If the latter is the case, then the user proceeds through the steps outlined in the model again to produce a new query.

Thus the representation of data can influence any of the steps in the model. Representations of metadata may cause the data user to change her formulation of the problem to use the data in different ways (or not use the data at all).

LITERATURE

Considerable attention has been directed to estimating, modeling and representing error in spatial data. Veregin (1989) provides a fairly comprehensive, albeit now dated, survey of the literature in this area. This body of research relates to the data component of our model. That is, it concentrates on the relationship between the values stored in the database and characteristics of the world. In some cases (e.g., Chrisman, 1982), these efforts have started by enumerating the possible sources of error in the data development process. The representation of error research clearly applies to the representation arrow in our model.

The problem of uncertainty in data has also been addressed by computer science and artificial intelligence researchers (cf., Motro and Smets, 1997). In most cases, these efforts have also concentrated on methods for modeling uncertain data and the mathematical techniques for manipulating the representations. Such techniques include interval mathematics, probability theory, Bayesian statistics, fuzzy set theory, and Dempster-Shafer Theory, among others (Bandemer, 1992). These methods represent one particular aspect of data (e.g., imprecision, vagueness, error) in a database. They also focus on the data side of our model.

Literature on decision making and judgment under uncertainty is prevalent in cognitive psychology. Much of it refers to a classic paper by Tversky and Kahneman that described several biases and heuristics people use when making judgments under conditions of uncertainty (Tversky and Kahneman, 1974). Some research has attempted to identify whether certain conditions determine the use of certain heuristics by people (Payne, Bettman, and Johnson, 1993). This research is critical for the data user component of our model, although its exact position in the model remains to be determined. For example, representing uncertainty in data may motivate the use of heuristics or biases in decision making. However, the substantive decision contains uncertainty introduced from other aspects of the decision environment in addition to uncertainty components deriving from data (Jordan and Miller, 1995).

CONCLUSION AND FUTURE RESEARCH

The model described here helps identify aspects of data uncertainty that have not been included in the consideration of data quality or accuracy. We see that there is a gap in our understanding of the result of representing uncertainty because we do not know how a data user might respond to information about uncertainty. These questions depend on the way a user proceeds through the steps of formalizing her request for data. Thus future research is necessary to investigate the types of information that users employ when formulating the problem and how metainformation such as uncertainty information would affect this process. We are proceeding with such research in this project by interviewing users to ascertain their conceptions of uncertainty in data, what kinds of uncertainty they consider in problem solving, and what kinds of uncertainty would change their approach to problem solving. We will in-

tegrate these questions by investigating the changes in their decision making resulting from changes in the representation of uncertainty of a data set.

Certain mathematical methods of modeling and manipulating uncertainty imply specific epistemological conditions. For example, if an interval mathematics approach is used, this implies that users of the data consider value intervals the most appropriate representation of uncertain values. Conversely, fuzzy set theory is more compatible with graded categories in perceptions of phenomena. Categories in cognition have been the subject of research in cognitive science for several decades. A common model of categories is the prototype model which is similar to a graded fuzzy category. Thus there is a question about whether a fuzzy representation of uncertain data is consistent with the cognitive categories of a user. The authors are also completing a paper that compares the representation of fuzzy regions to perceptions of regions.

The model is also useful in directing developments in information retrieval systems such as fuzzy queries or natural language interfaces that attempt to provide a more flexible interaction with a data repository. Similar to the needs for representation, the specific nature of an interface (i.e., what kind of interaction it provides) depends on the types of uncertainty that concern users.

REFERENCES

Bandemer, H. General Introduction, in *Modelling Uncertain Data,* H. Bandemer, Ed., Akademie Verlag, Berlin, 1992.

Chrisman, N.R. A Theory of Cartographic Error and Its Measurement in Digital Data Bases, in *Proceedings of AutoCarto 5, Crystal City, VA,* American Society for Photogrammetry and Remote Sensing, 1982.

Couclelis, H. People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS, in *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, Lecture Notes in Computer Science 639,* A.U. Frank, I. Campari, and U. Formentini, Eds., Springer-Verlag, Berlin, 1992.

Goodchild, M. Geographical Data Modeling. *Com. Geosci.,* 18, pp. 401–408, 1992.

Jordan, C.F. and C. Miller. Scientific Uncertainty as Constraint to Environmental Problem Solving: Large Scale Ecosystems, in *Scientific Uncertainty and Environmental Problem Solving,* J. Lemons, Ed., Blackwe Science, Cambridge, 1995.

MacEachren, A. *How Maps Work: Representation, Visualization, and Design,* Guilford Press, New York, 199

Motro, A. Sources of Uncertainty, Imprecision, and Inconsistency in Information Systems, in *Uncertainty Management in Information Systems: From Needs t Solutions,* A. Motro, and P. Smets, Eds., Kluwe Academic Publishers, Boston, 1997.

Motro, A. and P. Smets. *Uncertainty Management in Information Systems: From Needs to Solutions,* Kluwe Academic Publishers, Boston, 1997.

Payne, J.W., J.R. Bettman, and E.J. Johnson. *The Adaptiv Decision Maker,* Cambridge University Press, Cambridge, 1993.

Peuquet, D. A Conceptual Framework and Comparison o Spatial Data Models, *Cartographica,* 21, pp. 66–113 1984.

Rosenthal, R. and R. Rosnow. *Essentials of Behaviora Research: Methods and Data Analysis,* McGraw-Hill New York, 1991.

Smets, P. Imperfect Information: Imprecision and Uncertainty, in *Uncertainty Management in Information Systems: From Needs to Solutions,* A. Motro and P Smets, Eds., Kluwer Academic Publishers, Boston 1997.

Stinchcombe, A.L. *Information and Organizations,* University of California Press, Berkeley, 1990.

Thrift, N. Flies and Germs: A Geography of Knowledge, ir *Social Relations and Spatial Structures,* D. Gregory and J. Urry, Eds., Macmillan, London, 1985.

Tversky, A. and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases, *Science,* 185, pp. 1124–1131, 1974.

Ullman, J. *Principles of Database and Knowledge-Base Systems: Volume 1,* Computer Science Press, Rockville, MD, 1988.

Unwin, D.J. Geographical Information Systems and the Problem of 'Error and Uncertainty,' *Prog. Hum. Geogr.,* 19, pp. 549–558, 1995.

Veregin, H. *Accuracy of Spatial Databases: Annotated Bibliography,* Technical Paper 89-9, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA, 1989.