ABOUT DIFFERENT KINDS OF UNCERTAINTY IN COLLECTIONS OF SPATIAL DATA

Vincent B. Robinson
Department of Geology & Geography
Hunter College - CUNY
695 Park Avenue
New York, NY 10021

Andrew U. Frank
Department of Civil Engineering
University of Maine at Orono
Orono, ME 04469

ABSTRACT

We differentiate between two broad types of uncertainty. Type I uncertainty deals with our inability to measure or predict a characteristic or event with uncertainty, where the characteristic or event is inherently exact. In type II uncertainty, there is a situation of intrinsic ambiguity regarding the concept to represented. The many sources of the two types of uncertainty are listed. Methods of handling differing kinds of uncertainties contained in collections of spatial data are suggested. Due to recent advances and its ability to represent and process the vagueness of natural language concepts, fuzzy logic is identified as major area for future research in managing uncertainty in spatial information systems.

INTRODUCTION

Quality of data and information is of great concern to designers and users of all spatial information systems. This paper first defines the 'correctness' of data in a formal way and then attempts to classify different types of uncertainty and indicates methods to deal with them. It is felt that the methodical discussion of terminology may be useful for the current discussion of standards for Geographic Information Systems (Moellering, 1982).

Data accuracy is thought of as the accordance between 'reality' and the information stored in a spatial database using a fixed interpretation (mapping). Inexact data may stem from inaccurate measurements or from differing interpretation of the same data. This leads to the identification of two broad types of uncertainty commonly associated with spatial information systems. The first type of uncertainty is related to our inability to measure or predict a characteristic or event with uncertainty, where the characteristic or event is inherently exact. The second type of uncertainty is derived from the intrinsic ambiguity regarding the concept to represented. It is hoped, that from this presentation a better understanding of the nature of uncertainty will result and lead to development of methods to handle some of the manifest problems.

FUNDAMENTAL HOMOMORPHISM OF INFORMATION SYSTEMS

Data is collected and placed in a database so that that human users can find answers to certain types of questions they need to fullfill their functions in an organisation. It is assumed, that using the database to find the necessary information is simpler, faster and less expensive than if every user has to collect the information directly from reality. In doing so the user relies on the implicit assumption that the answers they retrieve from the database are essentially the same they would find, if they wen out and collected the data themselves.

In order to further formalize this notion, we use the mathematical concept of a homomorphism. Given two sets of objects, original and model, and for each operations on the original objects or model objects. A homomorphism is then a mapping between original and model, such that the result of operations on the original object correspond to the result of performing the corresponding operations on the corresponding model object and vice-a-versa. For an information system (ie. a model of reality) this is to says that information received from the information system (ie. a model operation on model objects) corresponds to information gathered from corresponding operations on reality (Frank, 1982).

SPATIAL INFORMATION SYSTEMS

We use the term **spatial information system** (SIS) to include all sorts of information systems that contain and process data with respect to location in space, especially over the surface of the earth (Frank, 1980). A spatial information system may include databases commonly described as a geographic information system, land information system, geographic expert systems, etc. We assume that the following discussion applies to all such systems insofar as they store and treat data with respect to location in 'reality'. Some of the concepts are even more general and may be applied to other, non-spatial information systems. Each information element in a spatial information system consists of two pieces, namely the description of the nonspatial properties of the object and its location and extension.

In the flow of data-to-information common to spatial information systems the measurement process and the subjective assessment/classification process are attempts to gather data relevant to the purpose of the information system. That is to say that data gathered from reality is filtered as a function of the task domain of the SIS. Furthermore, since SIS's tend to be more general purpose than most information systems, definition of the task domain is itself often vague.

ON THE NATURE OF UNCERTAINTY

The distinction between exact concepts and inexact concepts has important implications in how we view the uncertainty contained in a collection of spatial data. Figure 1 illustrates how the two fundamental elements of information in an SIS may combine to define four possible states of uncertainty. We find areas that are clearly determined, precisely measured, and properties of which are measured with exactness. On the other hand, we have areas, that are very imprecisely determined, eg. the Lowlands of South Carolina. Although we may have at our disposal the means to exactly define and measure the concept of population as an attribute of a spatial entity, because of our inability to exactly define the location and extension of the Lowlands of South Carolina we may not be certain about the total population reported to be residing in the lowlands.

		Attribute of Spatial Entity	
		Exact	Inexact
Locational Definition of Spatial Entity	Exact	No Uncertainty	Uncertain Attribute
	Inexact	Uncertain Location	Uncertain Attribute Uncertain Location

Figure 1. Combinations of Attribute and Locational Exactness/ Inexactness Possible in a Spatial Information System

To elaborate, lets consider two spatial phenomena. First, a cadastral land information system where property boundaries define the boundary of a 'crisp' spatial set. That is, a portion of the earth surface can either belong to 'property A' or 'property B' but not both. However, in a landuse/landcover data base there are several situations where 'crisp' boundaries are not in reality detectable. Using classical Boolean logic, Robinove (1981) provides a review of the difficulty in defining 'exact' boundaries. In effect, he argues for modifying the existing system of representation so that vague boundaries, such as ecotones, are be more accurately represented as the vague, or fuzzy, spatial phenomena that they are. In the former it is a question of how accurate an 'exact' concept is measured, while in the latter it is how to 'exactly' represent an intrinsically 'inexact' concept.

Classes of Uncertainty

We can identify with the procedures of data gathering different imperfections. The user must live with such differences and with the resulting uncertainty. In the traditional logic system that underlies the current genre of SIS's, if perfect accuracy were attained there would be :

1. No imprecision in determining the location or extension of a certain phenomenon.

2. No error in measuring the essential characteristics that define the phenomenon measured.

3. No uncertainty regarding the relative location or label of the phenomenon.

4. No inaccuracy in the assessment of the phenomenon with respect to the interpretation of concepts.

5. No differences between subjects in interpreting the concepts used.

6. No changes in reality that are not immediately reflected in the data stored.

Normally, perfect accuracy can not be achieved for the following reasons:

1. Objects to be measured are often only vaguely defined.

2. Measurements are inherently imprecise (but with additional measurement expensed, nearly arbitrarily levels of precisions can be achieved).

3. In measuring, gross errors not of a statistical nature may slip in.

4. Schemes for classification, or 'labelling', are always imprecise and lead to different attributions depending on subjective judgements.

5. Attributes encoded on a ordinal scale (e.g. dense, medium, sparse), function largely as approximate qualifiers of labels.

6. The subjective and context-sensitive interpretation of 'facts' influences the encoding of facts and affects data during the use of a data collection.

7. Large differences between the intended use of a data collection and the actual use may lead to subtantial differences in the definition of terms and categories, thus leading to semantic error.

8. Facts as represented in the data collection usually represent a past state of reality.

We differentiate between two types of uncertainty. Type I uncertainty deals with our inability to measure or predict a characteristic or event with uncertainty, where the characteristic or event is inherently exact. Error propagation resulting from the distribution of measurement/observation error is an example of Type I uncertainty. The mapping here is from an exact characteristic to an exact representation of an exact concept. Another kind of measuring error may give rise to Type I uncertainty. In the course of a measuring process gross errors not of a statistical nature may slip in and be encoded in the collection of spatial data. The use of certain measuring and computational methods (Baarda, 1973) may limit the possible effects of undetected gross errors on the results (reliability). The underlying assumption for these two aspects of uncertainty in measurements is that the phenomenon being measured does in fact exist without imprecision, and that the sources of our uncertainty are found in the measurement process.

In type II uncertainty, there is a situation of intrinsic ambiguity regarding the concept to represented. The use of 'prototypes' , ie.

examples of 'pure' cases, is a common tool in an attempt to minimize uncertainty. What is interesting about many examples of this type of uncertainty is their relation to natural language concepts. It is common for spatial data to be collected in a database format for purposes of representing the type and distribution of land use or land cover over a portion of the earth's surface. In addition, the use of Landsat data to gather such representations provides an excellent example of how we move from 'objective' information to 'subjective' labels.

To illustrate that the problems mentioned above are real, consider the problem of optimizing the labeling of image classes. Aronoff's (1984) method like others begins from the proposition that a 'location' on a map must belong to one class. That assumption is couched in the proposition that 'map error' is a yes or no matter. In his logic framework there is no such thing as a degree of error. The existence of ambiguity is virtually ignored. However, the existence of a single ambiguous point illustrates that 'pure' classes are a construct that maintained for the sake of conceptual convenience and tradition. It is also significant that the basis for 'verification' is the interpretation of data by a human analyst. Thus, classification error, in this case, is itself subject to the imprecision with which humans manipulate linguistic concepts such as 'forest', 'residential', etc.

Like many other landuse/landcover studies objective data from Landsat is subjected to unsupervised classifications to obtain image classes, then the results are subjectively assigned landcover, or resource, class labels by a human interpreter. The interpreter typically uses aerial photography to accomplish this task (e.g. see Pettinger, 1983). Thus, this is an inherently subjective task in which the interpreter is attempting to match objectively derived image classes with linguistic concepts that are represented by linguistic prototypes in the mind of the interpreter. It is not surprising then that there is variation in interpretation of the very same data, ie imagery, among interpreters. This particularly bothersome when the result is stored in a database because at this point an inherently imprecise concept is stored as an exact representation. Furthermore, a particularly questionable assumption of this procedure is that somehow the interpretation of the ground data is the benchmark against which to measure accuracy.

To illustrate the point that a concept such as a landcover class is inherently vague, let us consider the problem described by Aronoff (1984) when developing a data base to be used to map areas of Douglas Fir and areas of White Fir. In Aronoff's (1984) study it is reported that a full third of all pixels verified as Douglas Fir were classified as White Fir ! Surely there is a reason for this very large discrepancy between objective methods and the subjective ones used to 'verify' the results of objective measurements. As an aside it is interesting as well that a full 41 those pixels verified as White Fir were classified as Douglas Fir by objective means. Further contemplation of this situation would lead us to question the 'purity' of the respective fir stands. For example, exactly, not approximately, but **exactly** when does an area become White Fir rather than Douglas Fir ? When does this occur ? What is it classified if in the pixel there are 40% Douglas Fir, 40% White Fir, 10% Ponderosa Pine, and 10% Red Fir. The natural ambiguity of vegetational communities, as defined by human interpretation is one of the issues that led Robinove (1981) to suggest that we begin mapping landuse without 'crisp' boundaries. Let us carry this problem further. When moving from one vegetational community to another, there is usually not a clear boundary. This has been known for centuries, yet we still map landcover as if boundary between forest and grassland can be exactly determined as a 'crisp' boundary.

Treatment of Uncertainty

Statistical Variations. Measurements with their associated statistical errors can be adjusted using statistical theory. Propagation of errors when combining several measurement values to compute a new value is straightforward, according to the law of error propagation:

$$m^2 = \left(\sum (f.i)^2 \; (m.i)^2 \right)^{\frac{1}{2}} \qquad (1)$$

where m.i is the statistical error on term i and f.i is the influence of this term, ie. the total differential of the function for this term at an approximate location. It is obvious from this formula, that error propagation increases the error, never decreases it.

Ambiguity. Uncertainty arising from ambiguity or subjectivity of the encoding method can be treated with Fuzzy Logic. In this formal system, exact information is viewed as the special case where inexactness has been reduced to zero. Buckles and Petry (1982, 1983) have presented a fuzzy representation of data for relational data bases that satisfies two of problems most often encountered in collections of land use data. The need for a single land use type to associated with a tuple has in the past been due to restrictions placed on data base management by the underlying logic. Using a fuzzy representation of data for relational databases (Buckles and Petry, 1982), Robinson and Strahler (1984) have shown how one can represent seemingly conflicting landuse/landcover classifications in a fuzzy data base. Thus, preserving explicitly the uncertainty inherent in landuse/landcover labeling.

Repeated interactions with a spatial data base often results when a user searches for an accurate representation of an approximate spatial concept using a system that can not represent, much less retrieve, an approximate concept. The user ends up being uncertain about the how well the retrieved data represents the approximate concept, this leads to further interaction with the information system in an attempt to lower that level of uncertainty below some level and in the end may make decisions based on a low confidence in retrieved information. Many of these additional steps would be eliminated if the system were capable of representing and retrieving such approximate information. Robinson and Strahler (1984) have shown how fuzzy representations of data can be incorporated into a landuse/landcover data base management system. Using the results of the work by Buckles and Petry (1982), they show how fuzzy logic can be used to retrieve landcover data stored as linguistic variables.

Another rather subjective class of data found in collections of spatial data are attributes encoded on a ordinal scale, for example - low, medium, high. Terms such as high, low, medium can be treated as linguistic hedges within the context of the above discussion of linguistic information such as landuse/landcover types. Retrieval of land information in the form of ordinal labels leads to a large number of interative operations when an exact system is used to represent and retrieve such linguistic hedges.

UNCERTAINTY AND THE EXPLOITATION OF SPATIAL INFORMATION SYSTEMS

Spatial 'facts' can be represented by one of four fundamental types of data - ratio, interval, ordinal, and nominal (see Figure 2). However, it has been noted with interest that "...the overwhelming majority of GIS applications concern some type of discrete phenomena. Topographic feature codes, place names, geocodes, parcel identifiers, land use types, all fall into the same broad group" (Chrisman, 1984: p. 309). The data Chrisman (1984) referred to is in the main nominal data. On the other hand, data such as elevation, spectral, and planimetric measurements are data of the interval/ratio type. Often the nominal data entered into a GIS derived from interval/ratio data. However, a common practice is then to consider only the nominal data during the retrieval process. If interval/ratio data is stored then the query is structured in such a manner that in concept it is a labeling, or classification, process for retrieving 'labelled' data. This appears to be consistent with recent research in forecasting and man-machine studies suggest that human information processing is geared towards the processing of 'qualitative' information rather than 'quantitative' information (Zimmer, 1984).

Figure 2 depicts the relation we suggest exists between the four fundamental data types, objective information, and meaning. Nominal and ordinal data are generally characteristic of the data type desired as output products by the average 'user' of geographic information systems. These are generally expressed in linguistic terms related to either attributes or relations. For example a nominal attribute might be 'residential', while a relation might be 'places Near Bangor, ME'. These terms are meaning-laden, that is

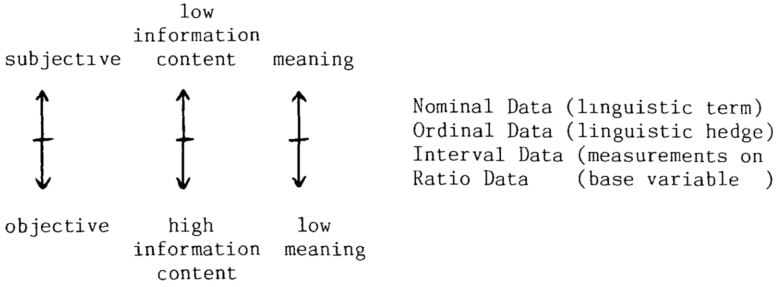
subjective
low information content
meaning
objective
high information content
low meaning
Nominal Data (linguistic term)
Ordinal Data (linguistic hedge)
Interval Data (measurements on
Ratio Data (base variable)

Figure 2. Relationship between information content, meaning, and data types.

to say that they may vary upon accumulated experience of the user and context of query. Ordinal data generally act as linguistic hedges used to place qualifications upon nominal data, often in an attempt to provide a sense of variation in the nominal concept arising out of a measurement process utilizing interval and/or ratio data. Thus, nominal/ordinal data is more meaning-laden to human users than is numerical data of the interval/ratio kind, while interval/ratio data typically contains more 'objective' information.

The above discussion is somewhat related to our previous critique of maps showing crisp boundaries where only fuzzy areas of slow transition exist. It is evident, that maps as products are very helpful sources of information because they present data in a grossly classified way and are therefore easy to use provided the mapmakers and the mapusers concepts agree. One of the main advantages of a map over a aerial photograph, to be classified is its information representation. The map produces less cognitive load on a user, therefore the user is more likely to make correct decisions under pressure (or would you like to use aerial photographs to navigate your automobile around New York City ?). This argument is not true if maps are considered as data collection intended for multiple use. We suggest that the representation of the data should be as near to the source as reasonable feasible and the classification to produce the desired nominal or ordinal data for the user should only be performed when output is produced.

Another subjective source of uncertainty was suggested by Chevallier (1983). The subjective interpretation of facts depending on the background and the goal of the task does not only influence the encoding of facts, but also affects data during the use of a data collection. If multipurpose data collections are built, differences between interpretations during data collection and use of the data must be taken into account.

Large differences between the intended use of a data collection and the actual use may lead to subtantial differences in the definition of terms and categories (semantic error). Discrepancies of this type should not be considered errors in encoding, but are systematic, regular differences in interpretation. Little is known about how such differences can be detected and taken into account. However, whenever they occur the effect may be considerable, especially in influencing decisionmaking process. Thus, much may be learned if the potential of natural language systems capable of representing approximate concepts can be exploited to study semantic error.

CONCLUDING DISCUSSION

Uncertainty in collections of spatial data can arise from a myriad of processes. Neverthless there are two fundamental types of uncertainty present in most collections of spatial data. Type I uncertainty is related to the error associated with the mapping from reality to the information system model of an exact concept. Historically, this is the only type of uncertainty that has been formally acknowledged as present in collections of spatial data. We consider a second type of uncertainty. Type II uncertainty results from a mapping from reality to the information system model of an inexact concept.

It is evident that there is little activity in devising methods of explicitly incorporating uncertainty in the management and retrieval of information in SIS's. The most common prescription it to attach lineage and measures of Type I uncertainty to maps or digital data files, leaving the processing of this information regarding uncertainty to the individual human user. Furthermore, Type II uncertainty has received little, if any, attention.

Evident from our review of different kinds of uncertainty is that the role of natural language (NL) concepts has been virtually ignored. This is a serious omission since (1) much 'data' in SIS's is of a linguistic nature, (2) linguistic rules are the basis for defining many criteria for data collection, (3) users often demand linguistic terms as the output from a SIS, (4) quality of information, including lineage, is most often described using linguistic terms derived from human judgement. Thus, one of the implications of our review is that the role of natural language concepts in SIS's should receive more rigorous, systematic attention. Another is that there is a clear contribution to be made in this area by man-machine studies. Consideration of both NL concepts and methods of explicitly incorporating uncertainty in the processing data in SIS's suggests that fuzzy logic may be of some considerable utility in the domain of SIS's.

Working within the framework of fuzzy data representation and management allows one to explicitly represent uncertainty in such a manner that it is able to become an integral part of spatial information processing functions. Linguistic variables are able to be represented and retrieved in a manner most consistent with their ambiguous nature. It is also clear that this framework has the potential to store differing characterizations of linguistic data on a user-specific basis, thus allowing the conduct of systematic research on variations in meaning according to user, context, and task.

Finally, we hope that this short overview asks more questions than it answers, but provides a flexible framework for additional, indepth investigations. In particular, we feel that present spatial data processing systems should be revisited in light of the two types of uncertainty identified in this paper.

ACKNOWLEDGEMENTS

Partial support was provided by NSF Grant IST-8412406 and a Research Grant from PSC-CUNY. We would also like to acknowledge the support provided by the City University of New York Computing Center, the University of Maine Computing Center, and the BITNET network, without which this collaboration would not have been possible.

REFERENCES

Aronoff, S. 1984, An approach to optimizing labeling of image classes, Photogrammetric Engineering and Remote Sensing, 50 : 719-727.

Baarda,W. 1973, S-Transformation and criterion matrices, *Netherlands Geodetic Commission, New Series*, Vol. 5, No. 1, Delft.

Buckles, B.P. and Petry, F.E. 1982, A fuzzy representation of data for relational databases, *Fuzzy Sets and Systems*, 7 : 213-226.

________ 1983, Information-theoretical characterization of fuzzy relational databases, *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13 : 74-77.

Chevallier, J.J. 1983, *Une approche Systemique des Systemes d'Information du Territoire et de leur integrite*, Ph.D. Thesis, Swiss Federal Institute of Technology, Lausanne.

Chrisman, N. 1984, The role of quality information in the long-term functioning of a geographic information system, *Proceedings* AUTO CARTO-6, 303-312.

Frank, A. 1982, Conceptual framework for land information systems - a first approach, paper presented at meeting of Commission 3 of the FIG, Rome, Italy.

________ 1980, Land Information Systems - An attempt towards a definition. *Nachrichten aus dem Karten- und Vermessungswesen*, Reihe 1, Vol. 81, Inst. f. angew. Geodaesie, Frankfurt FRD 1980]

Moellering, H. 1982, The goals of the National Committee for Digital Cartographic Data Standards, *Proceedings* AUTO-CARTO 5, 547-554.

Robinove, C.J. 1981, The logic of multispectral classification and mapping of land, *Remote Sensing of Environment*, 11: 231-244.

Robinson, V.B. and Strahler, A.H. 1984, Issues in designing geographic information systems under conditions of inexactness, *Proceedings* of 10th International Symposium on Machine Processing of Remotely Sensed Data, 198-204.

Zimmer A.C. 1984, A model for the interpretation of verbal predictions, *Int. Jrnl. Man-Machine Studies*, 20: 121-134.