FACTORS CAUSING UNCERTAINTIES IN SPATIAL DATA MINING

Hanning YUAN[a] Shuliang WANG[b]

[a]School of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China
[b] International School of Software, Wuhan University, Wuhan 430079, China
E-mail: hnyuanslwang@yahoo.com,

Commission IV, WG IV/3

KEY WORDS: Factors, Uncertainties, Spatial data mining

ABSTRACT:

Spatial data mining is to extract the unknown knowledge from a large-amount of existing spatial data repositories areas (Ester et al., 2000). The spatial data are to represent the spatial existence of an object in the infinitely complex world. They may be incomplete, noisy, fuzzy, random, and practical because the computerized entities are different from what they are in the real spatiotemporal space, i.e., observed data different from true data. For it works with the spatial database as a surrogate for the real entities in the spatial world, spatial data mining is unable to avoid the uncertainties. If the uncertainties are made appropriate use of, it may be able to avoid the mistaken knowledge discovered from the mistaken spatial data. The uncertainty parameters, such as, supportable level, confident level and interesting level, may further decrease the complexity of spatial data mining. Otherwise, it is unable to discover suitable knowledge from spatial databases via taking the place of both certainties and uncertainties with only certainties. Based on the unsuitable even mistaken knowledge, the spatial decision may be made incorrectly. The uncertainties mainly arise from the complexity of the real world, the limitation of human recognition, the weakness of computerized machine, or the shortcomings of techniques and methods. Their current constraints might further propagate even enlarge the uncertainty during the mining process.

1. OBJECTIVE REALITY

The world is an infinitely complex system that is large, changeable, nonlinear, and multi-parameter, about 80% information of which is spatial-referenced (Wang, 2002). In the spatial world, there are more inexact entities with indeterminacy or inhomogeneity than the exact ones. The spatial entity in the world includes historical information, current status, and future trend. At any moment, it receives the information from other entity, and it also eradiates its own information. The information of different entities may be overlapped, mixed, or deformed. Two entities of the same classification may eradiate different spectrum information, while two entities that eradiate the same spectrum information may belong to different classifications. As a result, it is confused to correctly classify the pixels with the same gray degrees in the boundary area where two different classifications overlap. In the real world, the information cannot be incarnated if it is not sensed by the observation of a certain instrument. Remote sensing captures spatial data via detecting the spectrum with sensors. Traditionally, it was presumed that the spatial world stored in spatial database was crisply defined, precisely described and accurately measured in computerized databases (Burrough, Frank, 1996). For instance, an object model assumes that the spatial entities may be precisely described via points with exactly known coordinates, lines linking a series of crisply known points, and areas bounded by sharply defined lines. However, these cases seldom happened in the real world, and in many cases, there do not exist the pure points, lines, and polygons with geometric definitions (Wang, Shi, 2002).

Some true spatial values are even inexact or inaccessible. The true values of spatial data are the actual characteristics of the spatial entity reality. Some true spatial values exist but are impossible to obtain. One is unobservable for they are spatial data with long history, the other is impractical to observe because they are too complex, difficult or expensive for human to get in the constraint contexts of current cognition, instruments and techniques, times and capitals. As to some spatial values, there are further no true values at all in the real world. Some spatial entities have no sharp boundaries or cannot be precisely determined. Take it for example that the spectrum of the spatial entity makes the image data uncertain. It is a fundamental function to determinate whether or not the spatial element belongs to the predefined entity, and the classification determination is performed on the accessible spatial values that are measured by sensors. The overlapped or mixed pixel of remote sensing images comprehensively reflects the classifications of different but neighbor objects on the ground. The additional but indispensable measurement step will further cause uncertainty because of the limitations in the process. Remote-sensing images of different objects may show the phenomena of spectral uncertainty created by spatial entities. One is that two objects belong to the same type or species but with different spectrums, which cannot be uniform as one spectral curve, but are composed of a series of different spectral curves, and cause a wide distribution. In a generalized category it also includes the multi-angular, multi-temporal and multi-scale effect, e.g., Rocks/Minerals, Vegetation. The other is that two objects belong to different classifications but with the similar or same spectral features in a certain wavelength range, e.g., the camouflage in military. New uncertainties may further be caused during the process of additional but indispensable measurements.

The uncertainty is more popular in macro-world (e.g., astro-space) and micro-world (e.g., the space that electron, proton are moving), both of which are moving at a high speed (Duncan, 1994). The length of moving objects, and the distance between two objects, all have contractility. The contractility changes

with the velocity or the distance. The time is relative, and different inertia system has different time coordinates. Furthermore, the time of an inertia system may find that the time of another inertia system opposite is slower. Because of time expansion, the observed time is always bigger than the true time.

2. SUBJECTIVE COGNITION

Compared with the complexity of the real world, the ability of human cognition is very limited at present. In order to guide their lives in the context of needs and wishes, people try to make sense of the real world around themselves in terms that they can understand and manipulate (Figure 2). First, they understand the spatial entity in the world, define and generalize the spatial entity, observe the spatial entity, relate the observations to an established conceptual data model, represent the spatial data in formal term, and store the data in the computerized machine. Some entities are perceived rather than the real entities, e.g., subjective neighbourhoods. Then, they edit, retrieve and analyze the spatial entities on the basis of the stored data with the computerized GIS. When cartographers perform generalization operations, such as aggregating, amalgamating and merging on features, other category features may be grouped into a certain feature.

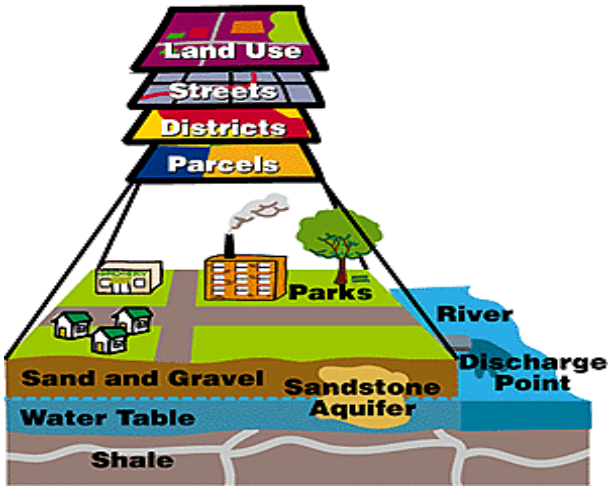

Figure 2 objective cognitions

As the entity is complex and changeable in the real world, people have to select the most important spatial aspects to approximately approach the reality entity. All spatial data are acquired with the aids of some theories, techniques and methods that specify implicitly or explicitly the required level of abstraction and generalisation (Miller, Han, 2001). So the depicted data are less than the total data about the spatial entity, and only an essential part of the real variation is described. The desired level is closely related to the spatial nominal concept of perceived reality, and it is defined by database specification of human cognition. In fact, the desired level is also the constraints from the current limitation of people cognition, and the spatial database composed of the captured data is only an abstracted representation. In consequence, the computerized entities may lose some aspects of the real entities, which make some uncertainties go along with spatial databases. Take the imagery data in remote sensing for example, the incomplete definitions of soil and forest may result in the vagueness about exactly what is their boundary in the ground.

3. APPROXIMATE TECHNIQUES

The observer cannot perceive the spatial of uncertain spatial entity directly, but only after they have been filtered by the uncertainty theories (Zimmermann, 2001). Based on the human cognition, the spatial entities in the real world are mapped to the crisp spatial objects in the computerized database via the given techniques for formal modeling, reasoning and computing (Figure 3). And the stored spatial objects are digitally represented with spatial data and their spatial relationships in a spatial database (Shi, Wang, 2002). Because the entity is indeterminate while the techniques are often deterministic (Burrough, Frank, 1996), the traditional techniques are often problematic when they are used to handle the spatial uncertainty.

First, most of the traditional tools are crisp, deterministic, precise, and dichotomous character. In dual logic, a proposition is "true or false" and nothing in between, in set theory, an element either belongs to a set or not rather than "more or less", and in optimization, a solution is either feasible or not (Zimmermann, 2001). They have implicitly assumed that the spatial entities are determinate or homogeneous, which is often not true in the real world (Figure 3). The acquired data via mathematical techniques are not as well as the real attributes in the world. For example, probability theory and fuzzy sets both integrate set theory and predication equation, and map the uncertainty to a numerical value in the interval [0, 1] in order to abstractly approach the spatial entity in the real world (Wang, Shi, Wang, 2003).

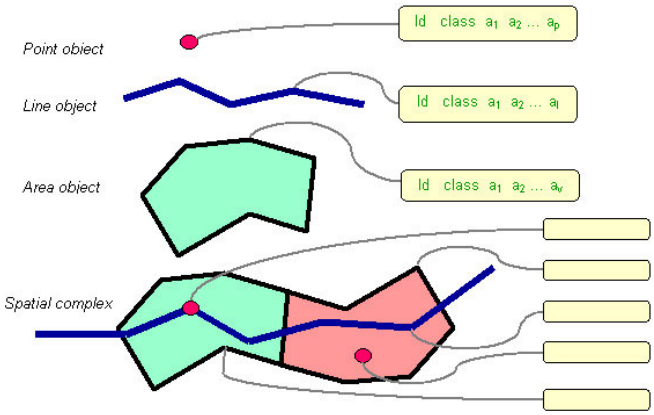

Figure 3. Approximate techniques

Second, no one solution can handle the complex interaction of different types of uncertainty. Each solution is capable of assessing just some aspects of uncertainty. The existing method on uncertainty often models a specific type of uncertainty under the specific type of circumstances, e.g., the theory of fuzzy sets can only model the fuzzy uncertainty.

Third, some techniques are incomprehensible to most common users without the background-associated knowledge, even some decision-makers. They may be unaware of, even misuse of the accuracy descriptors such as reliability diagrams and position error estimation on the basis of probability theory and

mathematical statistics (Arthurs, 1965). But it is one important role of reasoning under uncertainty to assist in decision-making. Fourth, a prerequisite to analyze spatial uncertainty is the availability of prior information about the uncertainty in data sources and how the uncertainty affects the outcome of GIS manipulations. This information may be known either exactly as a range with upper and lower bounds around some mean value; stochastically possessing a probability distribution function; or possibilistically belonging to a fuzzy set. However, factual prior information on the uncertainty is scarce, some are difficult, expensive, or even impossible to obtain.

4. COMPUTERIZED MACHINE

Spatial data in the computerized machine uncertainly reflects about the real world via binary digits in the form of zeros and ones when they are used to describe, acquire, store, manipulate and analyze spatial entities in the context of human needs (Goodchild, 1995). Some of the uncertainty may come from the computerized machine, e.g., physical modeling, logical modeling, data encoding, data manipulation, data analysis, algorithms optimization, computerized machine precision, output. And it is a discrepancy between the encoded and actual value of a particular spatial.

Any imaginable measuring device records its measurement only with a finite precision, even if the device is designed and used perfectly. Given the precision of a measuring device, the outcome may be lack of the infinite accuracy in the output instruments, e.g., monitor, printer. In order to record a measurement with infinite precision, the instrument would require an output capable of displaying an infinite number of digits. By using more accurate measuring devices, uncertainty in measurements can often be made as small as needed for a particular purpose, and the accuracy will become greater and greater. However, it only approaches but never reaches an absolute accuracy. Thus there is no real measurement with infinitely precision, instead of a value with a degree of uncertainty. During the process of machine-based computing and analysis, e.g., GIS buffering, layer overlapping and data mining, these uncertainties are accumulated and propagated. And the computerized machine may further produce new uncertainties.

5. AMALGAMATING HETEROGENEITY

The spatial uncertainty becomes even more complex when merging different kinds of spatial data, often from different sources and of different reliabilities (Hunter, 1996). Moreover, there often exist more than one uncertainty at the same time during the process of uncertainty-based spatial data mining. For example, both randomness and fuzziness are often included in spatial entities. In order to create a best possible database, spatial data users would like to see the matching and amalgamation of heterogenous data, i.e., some kind of average, or combination of elements from more than one source. But a common spatial database may conventionally support an exact local application without considering the global application. If these various local databases are integrated together in the global context, the conflicts among various spatial databases may also cause unpredicted uncertainties, e.g., inconsistency across multiple databases. Thus besides the abovementioned uncertainties from the real world, human recognition, computerized machine or techniques, some new uncertainties may further appear in spatial data if they are acquired from different sources with heterogenous representations.

In a word, the uncertainty is unavoidable in spatial data sets, and it can never be eliminated completely, even as a theoretical idea. During the process of spatial data mining, spatial uncertainty can propagate even become bigger when several spatial uncertainties are accumulated. The limitations of human recognition, mathematical model and technology may further enlarge the uncertainty, which more easily leads to mistaken decision making. Moreover, the increasing of the amount of spatial data may not result in the decreasing of the spatial uncertainty.

6. CONCLUSIONS

This paper presented the factors causing uncertainties in spatial data mining. They might include the complexity of the real world, the limitation of human recognition, the weakness of computerized machine, or the shortcomings of techniques and methods. In fact, the rational uncertainties (e.g., the uncertainties in natural language) may save people out of the data sea, and only the necessary data are allowed to enter decision-making thinking, then to sublime knowledge. Therefore, uncertainty-based spatial data mining is a potential research project.

ACKNOWLEDGEMENTS

The work described in this paper was supported by the funds from This study is supported by the funds from National Natural Science Foundation of China (70231010), Wuhan University (216-276081), and National High Technology R&D Program (863) (2003AA132080)..

REFERENCES

ARTHURS A. M., 1965, *Probability theory* (London: Dover Publications)

BURROUGH P.A., FRANK A.U.(eds), 1996, *Geographic Objects with Indeterminate Boundaries* (Basingstoke: Taylor and Francis)

DUNCAN, T, 1994, *Advanced Physics* [4th edition](London: John Murray)

ESTER M. et al., 2000, Spatial data mining: databases primitives, algorithms and efficient DBMS support. *Data Mining and Knowledge Discovery*, 4, 193-216

GOODCHILD M.F., 1995, Attribute accuracy. In *Elements of Spatial Data Quality,* edited by GUPTILL S.C. and MORRISON J.L (New York: Elsevier Scientific), pp.139-151

HUNTER A, 1996, Uncertainty in Information Systems (London: The McGraw-Hill Companies)

MILLER H. J., HAN J., 2001, *Geographic Data Mining and Knowledge Discovery* (London and New York: Taylor and Francis)

SHI W.Z., WANG S.L., 2002**,** Further Development of Theories and Methods on Attribute Uncertainty in GIS, *Journal of Remote Sensing*, 6(4): 282-289

WANG S.L., 2002, Data Field and Cloud Model -Based Spatial Data Mining and Knowledge Discovery. *Ph.D. Thesis* (Wuhan: Wuhan University)

WANG X.Z., SHI W.Z., WANG S.L., 2003, *Handling Fuzzy Spatial Information* (Wuhan: Wuhan University Press)

ZIMMERMANN H.J., 2001, *Fuzzy Set Theory ----and Its Applications* (4^{th} edition) (Boston: Kluwer Academic Publishers)