UNCERTAINTY-BASED SPATIAL DATA MINING

Wenzhong SHI[1], Shuliang WANG[1,2], Deren LI[3], Xinzhou WANG[3]

1 Advanced Research Centre for Spatial Information Technology, Department of Land Surveying & Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong
2 International School of Software, Wuhan University, Wuhan, China, 430079
3 National Laboratory for Information Engineering in Surveying Mapping and Remote Sensing, Campus of Surveying and Mapping, Wuhan University, Wuhan, China, 430079
Fax: +852-2330 2994; E-mail: lswzshi@polyu.edu.hk, lsslwang@polyu.edu.hk

Abstract

Although uncertainties exist in spatial data mining, they have not been paid much attention to. Uncertainty may have an influence on the confidential level, supportable level, and interesting level of spatial data mining. This paper proposes uncertainty-based spatial data mining. First, the concept is given in the integrated contexts of both uncertainty and spatial data mining. The inherent uncertainties that have their own characteristics play an important role in spatial data mining. Second, the external aspects and their internal sources of uncertainty-based spatial data mining are given. Besides the errors, spatial uncertainties further include positional uncertainty, attribute uncertainty, topological uncertainty, inaccuracy, imprecision/inexactitude, inconsistency, incompleteness, repetition, vagueness, noisy, omittance, misinterpretation, misclassification, abnomalities and knowledge uncertainty. Given a mathematical interpretation, the internal sources may be randomness, fuzziness, blunders, chaos, etc. To control and reduce uncertainty in an acceptable degree, one is data acquisition that highlights the information acquired from the process of data collection and data amalgamation, the other is data cognition that emphasizes the knowledge discovered from data extraction process and information generalization. Third, the usable techniques and methods that may possibly cope with the uncertainties in spatial data mining are briefly overviewed. For example, GIS data models, analysis of error propagation, probability theory and mathematical statistics, extended sets. The cloud model integrates the randomness and fuzziness by using the formalization-computerized language, and it is more appropriate when there exist more than one uncertainty at the same time, e.g., randomness and fuzziness. Finally, a case study is given on Baota landslide.

Key Words

Uncertainty, spatial data mining, external aspects, internal sources, cloud model

1 Introduction

The rapid development of the instruments and infrastructures on Geo-Informatics makes spatial data complex, changeable and big, which has been beyond the human ability to analyze and interpret. This bottleneck that human faced with large amounts of spatial data is still short of knowledge cannot be resolved by only a single conventional technique respectively, for instance, cognitive science, mathematics, artificial intelligence, machine learning, pattern recognition, spatial data analysis, or database technique (Li, 1997; Li et al., 2001). Thus it promotes the speedy growth of the novel multidisciplinary field for detecting spatiotemporal patterns across multiple data sets that are accumulating, i.e. spatial data mining, or knowledge discovery from spatial databases. As it can extract more generalized or summarized rules, spatial data mining may enhance human ability to interpret spatial data and generate useable information. For example, data mining may reveal how the variation in climate affects the spatial distribution of land cover in ways that would be extremely difficult to predict with traditional statistical approaches. Now, a growing attention has been paid to spatial data mining, which will significantly expand the use of geospatial data in a variety of scientific or practical areas (Fayyad et al., 1996; Di, 2001; Ester et al., 2000; Miller, Han, 2001).

There are uncertainties in spatial data, and they may directly or indirectly affect the quality of spatial data mining (Han, Kamber, 2001; Wang, 2002). Spatial data of the database is to represent the spatial existence of an object in the infinitely complex world. But there are virtually uncertainties inherent in most of the spatial data capturing and data analyzing due to the limitations or constraints of current instruments, technologies, capitals, and human skills. The uncertainty is the major component of spatial data quality, which is specified as an essential characteristic of data by the Federal Geographic Data Committee's (FGDC) Content Standards for Digital Spatial Metadata (Goodchild, 1995). In the Spatial Data Transfer Standard (SDTS), the data quality is further divided into five fundamental components: positional accuracy, attribute accuracy, logical consistency, lineage, and completeness (Burrough, Frank, 1996). In the sequence, temporal accuracy, thematic accuracy and currency are also added (Shi, Wang, 2001). Moreover, the uncertainty is an essential part of many models of spatial data based decision-making, which has become the subject of a growing volume of research and figured prominently in research agendas, e.g., spatial decision-making support, intelligent GIS (geographical information system), sustainable resources and environments, image interpretation driven by knowledge, robot motion planning, computer aided design, RS (remote sensing), GPS (global positioning system), weather prediction, transportation management, environmental protection geology, agriculture, biology. Because it works with the spatial database as a surrogate for the real entities in the spatial world, and spatial data mining is unable to avoid the uncertainties (Wang, 2002). If the uncertainties hidden in the database have been taken as the input of spatial data mining, and further have not been paid attention to, the resulting discovered output might be the wrong knowledge. In consequence, the wrong knowledge may more easily leads to a mistaken decision-making. Thus it is necessary to deal with uncertainty so as to make the discovered knowledge aware of the level of uncertainty present. Spatial data cleaning, a data preprocessing phase of spatial data mining, has generally focused on the uncertainty aspects of data incompleteness, data inaccuracy, data repetition, data inhomogeneity, data inconsistence, and image deformation (Wang, Wang, Shi, 2002).

However, the uncertainties in spatial data mining have not been addressed to the same degree to spatial data mining itself (Di, 2001; Ester et al., 2000; Wang, 2002; Wang et al, 2003). First, although there have been considerable theories and techniques on either spatial data mining (Li et al., 2002) or the uncertainties in spatial data (Shi, Wang, 2001), each effort is focused on its own field. It is strange to find out the integration of spatial data mining and uncertainties when dealing with the elements, measurement, modeling, propagation, and cartographic portrayal in the literatures. Second, many efforts on the uncertainties specialize on the general autocorrelation and are not oriented to discover knowledge from spatial data sets in the uncertainty context, e.g., the spatiotemporal prediction at a particular location at a specific moment in time when the given conditions at a known location and at the present time change. Third, most techniques on the uncertainty may describe some specific situation. The common users without enough background knowledge may have difficulty in making sense of the exact nature of uncertainty that an expert specifies. For example, probability theory (Arthurs, 1965,) pay only attention to randomness, and it is also difficult for the average users to understand the probabilistic density function specified by the experts (Haining, 2003). Fourth, spatial uncertainties are strongly weighted towards precise locations and boundaries, requiring coordinates or area polygons. Some commercial GIS software and data vendors argue that techniques for dealing with uncertainty have no demand in the marketplace or confuse what is otherwise a bullish enthusiasm for the technology (Goodchild, 1995). They are not good at dealing with anything other than absolute time and position. Fifth, besides hard computing, soft computing should be also studied in the context of spatial data mining together with uncertainty. The mechanism of spatial data mining is close to human thinking, and its algorithms are often soft computing. The traditional theories, for example, spatial statistics (Cressie, 1991) belongs to hard computing, which needs a lot of observed sample values to deal with randomness. But the decreasing of uncertainties is unequal to the increasing of spatial data (Wang, 2002). Fuzzy sets care for fuzziness, the algorithms of which belong to soft computing (Zadeh, 1994). But it is difficult to decide the fuzzy membership function. Moreover, the soft computing becomes the hard computing once the fuzzy membership function has been decided. Sixth, new techniques on multi-uncertainty should be further studied because many uncertainties appear at the same time in spatial data mining, e.g., randomness and fuzziness. Then, in order to continue enjoying its success in spatial applications, spatial data mining should think of the uncertainties carefully, and the theories and techniques to deal with the uncertainties may have to be further studied. How to select the existed interesting techniques of spatial data uncertainties and apply them in spatial data mining? How to think of the impacts from spatial data uncertainties when spatial data mining is carried out? How to realize the mutual uncertain transformation between qualitative knowledge and quantitative data? All these questions promote the work of uncertainty-based spatial data mining in this paper.

This paper is to propose the uncertainty-based spatial data mining. The remained sections will be organized as follows. Section 2 is the concepts. Then section 3 presents the external aspects and the internal sources. The usable methods will be presented in section 4. Section 5 gives a case study. Conclusion is finally drawn in section 6.

2 Concepts

Spatial data point to the data that are able to represent the spatial existence of an entity, and there are various kinds, e.g., positional data, attributes, temporal data, images, graphics. So it is difficult to define an uncertainty-based spatial data mining completely. Here, only a describing definition is given.

The uncertainty-based spatial data mining is to extract knowledge from the vast repositories of practical spatial data under the umbrella of uncertainties with the given perspectives and parameters. With different granularities, scales, mining-angles, and uncertain parameters, it discovers the collective attribute distribution of spatial entities via perceiving various variations of spatial data and their combinations in the data space. It is derived from spatial data mining that is a branch of data mining, and the discovered knowledge is also diversity (Table 1).

Table 1. The discovered knowledge of uncertainty-based spatial data mining

Knowledge	Data mining	Spatial data mining	Uncertainty-based spatial data mining
Association rule	Yes	Yes	Yes
Clustering rule	Yes	Yes	Yes
Classification rule	Yes	Yes	Yes
Characteristics rule	Yes	Yes	Yes
Serial rule	Yes	Yes	Yes
Regression rule	Yes	Yes	Yes
Dependent rule	Yes	Yes	Yes
Spatial topological rule		Yes	Yes
Spatial distribution rule		Yes	Yes
Outlier	Yes	Yes	Yes

2.1 Characteristics

The concept is in the integrated contexts of both uncertainty and spatial data mining. It is an uncertain process for spatial data mining to discover the little-amount refined knowledge from the large-amount coarse data. In details, the uncertainties in spatial data mining may exist in spatial data, theories and techniques, mining process, knowledge characteristics, knowledge representation, knowledge interpretation, and so on. At the same time, the manipulations of spatial data mining are more abundant than common data mining on transaction data, for examples, overlaying map layers, buffering spatial entities, overlapping spatial objects, merging / amalgamating polygons, which not only helps people to mine more knowledge, but also increases the chances to produce more uncertainties.

First, the huge amount of objective spatial data may be incomplete, noisy, fuzzy, random and so on. The real world abounds in uncertainty, and any attempt to model any aspect of the entities in the world should incorporate uncertainty. There may be uncertainty in the understanding of entities or in the quality or meaning of the data. The serious uncertainties in spatial data should be identified instead of presenting them as correct. As the spatial data are the objectives of spatial data mining, the uncertainties are brought to spatial data mining along with spatial data at the beginning.

Second, from different perspectives of the same set of data, there are various kinds of knowledge that may be discovered. Either different people apply the same technologies, or the same people apply different technologies may discover different knowledge from the same data sets. Even with different mining-angles, different granularities, and different scales, people may achieve different knowledge (Wang, 2002). Moreover, the unknown knowledge is refined with high abstraction level, small scales, and small granularities, whereas the existing data are coarse with low abstraction level, big scales, and big granularities. At a higher hierarchy, there may be uncertainty about the level of uncertainty prevalent in various aspects of the database.

Third, the mining theories and techniques are able to deal with, manage, control, and make use of some aspects of the uncertainties, e.g., probability theory and mathematical statistics for randomness (Arthurs, 1965), fuzzy sets for fuzziness (Zadeh, 1965), rough sets for incompleteness (Pawlak, 1991), cloud models for the integration of randomness and fuzziness (Li, 1997). But the uncertainty in the model may further result in more uncertainty introduced to entities or the attributes describing them.

Fourth, the resulting knowledge may be hidden, implicit, valid, novel and interesting. Either spatial or non-spatial is unknown in advance, potentially useful, and ultimately understandable, together with three parameters to measure its uncertainties.

Fifth, it is uncertain to represent the discovered knowledge. People often think of decision-making with qualitative concept instead of quantitative data. The discovered knowledge is generalized or summarized when many quantitative data are generalized into few qualitative concepts. And it is uncertain to transform among data and concept.

Finally, the framework may include preprocessing uncertainty, mining uncertainty, resulting uncertainty, and interpreting uncertainty, for example, spatial data cleaning, summarization and generalization, knowledge representation, and knowledge application. Throughout the process of spatial data mining, all the uncertain characteristics of spatial data may be propagated and cumulated. And new uncertainties will further come into being during the process of data mining, knowledge representation, and knowledge interpretation.

2.2 Uncertainty parameters

There are three threshold parameters to measure the uncertainty in spatial data mining, i.e., supportable level (support), confidential level (confidence), and interesting level (interest) (Li et al., 2001). The supportable level of a rule carries the statistical significance of the spatial entities in the rule, the confidential level portrays the strength of the rule, and the interesting level describes how people are interested in the rule in their spatial decision-making. The virtue of a rule is characterized by its uncertainty parameters. The supportable level and the confidential level are also named prevalence, predictability in MineSet (Brunk, Kelly, Kohavi, 1997).

The supportable level and the confidential level are often found in spatial association rules. An association rule with large supports and high confidences is desired because they can be applied to many data, and hold with high probabilities. Take "P ⇒ R (s%, c%)" for example. Here P and R are sets of spatial and non-spatial predicates, s% and c% are the supportable level and the confidential level. The support level s% of a pattern P in a set of spatial objects S is the probability that a member of S satisfies pattern P, and it is the support of all entities associated with the rule. The confidence c% level of the rule P ⇒ R is the probability that the pattern R occurs if the pattern P occurs. The rule means that it is s% evidence to conclude that there is a correlation between P and R, and the degree of correlation between P and R is c%. That is, s% of the entities in the database contains $P\cup R$, and c% follow the clause that R occurs if the pattern P occurs.

The uncertainty parameters are the threshold indices, and they play an important role in decreasing the complexity of spatial data mining. There might exist thousands of spatial rules in case of very large database. The mining calculation involves repeated scanning of database and computing effort, which may become very complex as the number of spatial entities in the combination grows. Although the hardware and software of calculation are getting cheaper, it is still expensive to calculate the larger number of data combinations. Besides the algorithms, e.g., Apriori, Sampling, Dynamic Item count, Partitioning, Parallelism, the uncertainty parameters are also used to reduce the database activity via filtering the discovery of infrequent, uninteresting, or unhelpful rules (Wang, 2002). Once the frequent subsets in a database are determined, the rule can be mined via simpler algorithms.

2.3 Spatial data cleaning

Spatial data cleaning is an essential in uncertainty-based spatial data mining. It is the process of improving spatial data quality. In a narrow sense, spatial data cleaning includes understanding the semantic fields and their relationships in databases, checking and affirming the completeness and consistence of acquired data, determining cleaning regulations in the real task context, eliminating error data, removing redundant data, filling lost data with a certain technique, handling noisy data， resolving data conflicts, revising data, improving accuracy, correcting the radiate and geometric deformation on graphics and images, and bettering whole usability of spatial data (Wang, Wang, Shi, 2002).

Spatial data cleaning is not a simple processing of turning the records into the right records, and it also analyzes and recombines spatial data. It pays more attention to the content inconsistency than the form conflicts among the multi-sources spatial data. The methods on spatial data cleaning are tightly related to the exact task of spatial data mining, and its basic methods are data merging or data purging. Based on a certain objective, spatial data recombination extracts the spatial data from the separate sources, then put into the target spatial database, which may not only save the storage and computation, but also accelerate the speed, accuracy and validity. Under the umbrella of the techniques, spatial data mining can be classified into three types, data migration that gives simple migration regulations, data scrubbing that makes use of specific field knowledge, and data auditing that makes data clean with statistical analysis.

2.4 Advantages and benefits

It is known that the uncertainty is unavoidable in spatial data sets, and it can never be eliminated completely, even as a theoretical idea. Moreover, the decreasing of spatial data uncertainty is unequal to the increasing of data. The limitation of mathematical model and technology may further propagate even enlarge the uncertainty during the process of GIS analysis, which more easily leads to mistaken decision making.

Simultaneously, on the spatial reality world, the mathematical hypothesis should be in the context that the uncertainty is unavoidable, and the data acquired from the reality world are often incomplete. It is unable to well study an entity via taking the place of both certainties and uncertainties with only certainties. 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.

If the uncertainties are made good and right use of, it may be able to avoid the mistaken knowledge discovered from the mistaken spatial data. Otherwise, based on the mistaken knowledge, the spatial decision may be made incorrectly. To improve spatial data quality in the context of spatial data mining, it includes understanding the semantic fields and their relationships in databases, checking and affirming the completeness and consistence of acquired data, determining cleaning regulations in the real task context, eliminating error data, removing redundant data, filling lost data with a certain technique, handling noisy data, resolving data conflicts, revising data, improving accuracy, correcting the radiate and geometric deformation on graphics and images, and bettering whole usability of spatial data.

3 Aspects and sources

The uncertainty mainly arises from the complexity of the real world, the limitation of human recognition, the weakness of computerized machine, and the shortcomings of techniques and methods. In details, they may include instruments, environments, observers, projection algorithms, slicing and dicing, coordinate system, image resolutions, spectral properties, temporal changes, etc. At the same time, their current limitations might further propagate even enlarge the uncertainty during the process of spatial data mining. The external aspects and their internal sources of uncertainty-based spatial data mining are shown in Table 2. All of them are affected by the scale, granularity and sampling in spatial data mining. Some may further be treated as the factors in assessing the success of spatial data mining (Wang, 2002).

Table 2. External aspects and internal sources of uncertainty-based spatial data mining

External aspects	Internal sources			
	Randomness	Fuzziness	Blunders	Chaos
Error	Yes	Yes	Yes	Yes
Positional uncertainty	Yes	Yes	Yes	Yes
Attribute uncertainty	Yes	Yes	Yes	Yes
Topological uncertainty		Yes	Yes	
Inaccuracy	Yes	Yes	Yes	Yes
Imprecision				Yes
Inconsistency	Yes	Yes	Yes	Yes
Incompleteness		Yes	Yes	
Repetition	Yes		Yes	Yes
Vagueness	Yes	Yes		
Lineage				Yes

Temporal uncertainty	Yes	Yes	Yes	Yes
Knowledge uncertainty	Yes	Yes	Yes	

3.1 External aspects

There are various external aspects of uncertainties in spatial data mining. Compared with the error, the uncertainties in spatial data mining are more complex and common. Besides the errors, spatial uncertainties further include positional uncertainty, attribute uncertainty, topological uncertainty, inaccuracy, imprecision/inexactitude, inconsistency, incompleteness, repetition, vagueness, noisy, omittance, misinterpretation, misclassification, abnomalities and other possibilities (Table 2).

Error is everything introduced by limited means of taking measurements. It is mainly classified into systematic error, random error, and blunder error, which can be improved by applying more accurate measurement methods and more sensitive instruments. Alternatively, error can be viewed as a form of inherent uncertainty in some abstracted characteristics of the real world. The theorem of error propagation is a classical method to deal with the data error.

Position uncertainty (or geometric uncertainty, or spatial uncertainty) is the difference between the apparent locations of the feature as recorded in a database under the umbrella of the selected system and its true location in the real world. And it is a fundamental aspect of spatial data mining concerning specific locations on the earth, which is uncertain in human cognition.

Attribute uncertainty (or thematic uncertainty) in spatial data is the spatiotemporal differences between known attributes and attributes to know in terms of the spatial entity (Shi, Wang, 2001). Attribute uncertainty and position uncertainty are both tightly associated with each other, for example, the indeterminate boundary of different classifications (Burrough, Frank, 1996).

Topological uncertainty describes the spatial relationships that are associated with spatial entities having indeterminate or vague boundaries, i.e., disjoint, touch/adjacent, overlap, equal, cover/intersect, covered/ intersected by, contain, contained by. In data mining applications, one must not only be aware of uncertainty, but also exploit it in an effort to discover relationships in data that might not have been discovered otherwise, e.g., association rules describing spatial objects associated with other objects, and generalized attributes for spatial data (Koperski, Han, 1995).

Inaccurate data are the data different from the true value, dated without updating, data from inaccurate calculation, error type data, incorrect data, misunderstanding data, strange form data, or encrypted data.

Imprecise data are due to a finite representation of spatial entities. And they may be from technical instruments, mathematical models, or human sense organs. For example, the regular tessellation used in raster pixels, where the element of the tessellation is the smallest unit that represents space.

Inconsistent data arise when several versions of the same object exist, due either to different time snapshots, or datasets of different sources, or different abstraction levels (Shi, 1994). And they may be grouped into two classifications, conflicts in the context and conflicts out of the context. Spatial inconsistency consists in the data-source interior and among different sources.

Incompleteness is the uncertainty caused by the reasoning with inadequate information. It is related to totally or partly missing data on the records, or missing the attributes of the record. For example, a dataset is obtained from digitizing paper maps while pieces of lines are gone. The part of a document may also be torn, damaged or otherwise illegible but still mostly usable. The completeness can be assessed relative to the database specification. It is noted that rich data are not equivalent to complete data. Sometimes they may possibly be incomplete.

Repetitive data are that there are more than one repeated data on the same spatial entity in a database, or different databases. The repetitive data may be the records, the attributes of the record, the topological relationships, etc.

Vagueness may come from the spatial entity itself, mathematical modeling, or human cognition. It is an intrinsic property of many spatial features that do not have crisp or well-defined boundaries really.

Lineage describes the systematic uncertainties at any stage of data life, e.g., source observations, acquisition methods, form transformations, data deviations, assumptions and criteria.

Temporal uncertainty is the uncertainty of either date or time on the data and the discovered knowledge, together with the effectiveness for a given valid period of data mining. An exact date may be vaguely specified by only the month, interval between two points, year, or even duration instead of a day. For example, it is uncertain that the distinct date of "17/10/2003" is specified by "October", "between 16/10/2003 and 18/10/2003, "2003", "from 2000 to 2005". Simultaneously, a certain time may be unclearly expressed, e.g., "9:00 a.m." is given in "in the morning". The spatiotemporal uncertainty may be caused by the lost historical data on the references to a place or object, e.g., dated or ephemeral buildings.

3.2 Knowledge uncertainty

In the context of the uncertainties, it plays an essentially important role in spatial data mining to properly represent the knowledge discovered from database because there are uncertainties hidden in the knowledge. When roll-up or drill-down is carried out during the process of spatial data mining, the represented knowledge and the objective data should be transformed back and forth as human being are thinking. The cell of human thinking is a linguistic atom that is the minimum linguistic term, and the linguistic term is the basic unit of natural language. The linguistic atom is corresponding to the most elementary concept.

Fundamentally, the natural language serves to describe complicated concept with most elementary ones, and their various combinations. With the natural language, human beings could observe and analyze the same spatial entities on variant levels of granularities, and further in the different worlds of different granularities simultaneously. So spatial qualitative concept is an alternative to represent the knowledge because the discovered rules are often associated with spatial entities at the cognitive concept hierarchy, and the natural language certainly becomes the best way to represent spatial knowledge. In this context, it is the basis of spatial data mining to search for the qualitative concept described by the natural language to generalize a given set of quantitative datum with the same feature category, and it weights more to describe the quantitative concept with linguistic terms than with precise math equations. That is, spatial qualitative concept can be made by a set of spatial data, and it is more direct and understood than spatial quantitative data. The more abstract the knowledge to be discovered, the greater the advantage.

However, the concept may be either certain or uncertain. The extension of certain concept is precise and stable, while that of uncertain concept is imprecise and changeable. Human describes the uncertain concept via natural language. At the same time, the uncertain concept mainly concerns with fuzziness and randomness. There is a gap to be bridged between the rigidity of computerized spatial data and the uncertainty of spatial qualitative concept. In details, it is difficult to carry out the uncertainties of spatial transition between qualitative concept and quantitative data, especially when more than one uncertainties appear together at the same time, e.g., both the fuzziness and randomness.

3.3 Internal sources

The abovementioned external aspects of uncertainties describe the difference between the observed values of an entity and their true values in a spatiotemporal space. Given a mathematical interpretation, the internal sources may be randomness, fuzziness, blunders, chaos, etc.

Randomness is the uncertainty included in a case with a clear definition, but not always happening every time, or the instabilities of the membership that an element belongs to a qualitative concept. The essence of randomness is that there are a lot of unknown factors to affect it, and the effective impact of each factor is not decisive. So the randomness will become less when people understand and think of those factors. Randomness may be measured via the probability that the case happens. Its mathematical tools are probability theory and mathematical statistics, and their extensions or developments, e.g., spatial statistics, evidence theory, vector of probabilities, “S” band.

Fuzziness is the uncertainty included in the case that has happened in the opposed and incomplete world, but cannot be defined exactly. That is, the boundary of a qualitative concept or classification is so vague that an element of the data set will not be uniquely assigned to one subset. In fact, fuzzy uncertainty comes from the macro simplification, without getting to the bottom of a factor, or the factor unknown to people. Fuzziness is measured by the fuzzy membership value in the context of fuzzy sets (Zadeh, 1965). The possibilistic approach of uncertainty offered by fuzzy sets forms a useful complement to the measures of probability theory.

Chaos is the uncertainty in a complex large system composed of many cell system (Awrejcewicz, 1989). A single cell may be a simple certain. But the system state shows uncertain when lots of cells are coupled in a complex system. That is, chaos is the complex activities when simple rules are assembled together in a nonlinear certain system, and it has properties of both randomness and certainty. Chaos theory believes that a spatial entity is in an imbalance because it may be impacted by various factors. The imbalance leads to the uncertainty. In geo-spatial science, it shows spatiotemporal asymmetry and instability (Awrejcewicz, 1989). At present, the complex problems with chaos are resolved mainly via distinguishing and extracting the essential rules from chaos system, analyzing and predicting the motion characteristics of the system, or producing artificial chaos on purpose.

Blunders are caused by the mistakes, careless, laziness, tire, inappropriate operation, disturbance, omittance, misinterpretation, misclassification, abnomalities and so on from human being occasionally. The detection and control of blunders may be referenced in (Li, 1988).

3.4 Management and control

The strategies for managing uncertainties in data mining may develop formal, rigorous models of uncertainty, understand how uncertainty propagates through spatial processing and decision making, communicate uncertainty to different levels of users in more meaningful ways, design techniques to assess the fitness for use of geographic information and reducing uncertainty to manageable levels for any given application, learn how to make decisions when uncertainty is present in geographic information, i.e. being able to absorb uncertainty and cope with it in our everyday lives. In applying the strategy, consideration is initially given to: the type of application, the nature of the decision to be made, low risk versus high risk, non-controversial versus controversial, non-political versus political, the degree to which system outputs are utilised within the decision making process (Shi, Goodchild, Fisher, 2002).

There are two main technical directions to control and reduce uncertainty in an acceptable degree. One is data acquisition that highlights the information acquired from the process of data collection and data amalgamation, the other is data cognition that emphasizes the knowledge discovered from data extraction process and information generalization. In the past, the direction of data acquisition has achieved many results, e.g., new instruments, new sensors, data amalgamation system, database technology, computerized network. These results have bettered data acquisition and data allocation in some extent. At the same time, new algorithms on object track and object capture have further ameliorated the quality of produced information and knowledge. The direction of data cognition stresses the design of cognitive process, which reduces the uncertainty by providing decision-maker with more knowledge. An interdisciplinary subject, spatial data mining and knowledge discovery, will play an important role in this direction (Wang, 2002).

4 Usable techniques and methods

It is known that the uncertainties in spatial data are inherent, and may be propagated from the beginning to the end. Now the uncertainty has been studied more in accordance with spatial data quality, the traditional computational solutions of which may address spatial parameterization and forecast uncertainty but are largely mathematical simultaneously.

4.1 GIS data models

It found that many problems of discrete variable and continuous variable did not show isotropy, and semantic symbols had various influences on the results (Fisher, 1991; Mcmaster, 1996). So GIS data model that mainly includes an exact object model and a continuous field model were used, and cartographic convention further enhanced them (Burrough, Frank, 1996).

The object model might discuss the position uncertainty of discrete objects via statistical simulation (Shi, 1994). However, the object model might lose details in one or more dimensions when the computerized GIS handled with the spatial entity. For instance, a bus stop becomes a point without size or shape. In many cases, some attribute values of the spatial entity are inexact or inaccessible. The abovementioned facts make it indiscernible to associate a spatial element (e.g. pixel) to a given entity in attribute classification. In order to improve the exact object model, the continuous field model was further given to study the data uncertainties of continuous objects (Zhang, Goodchild, 2002). But the field model could not take the place of the object model. Hence, when they are used to describe a spatial entity with spatial data, the object model is to discrete entity with vector data, and the field model is to continuous entity with raster data (Goodchild, 1995; Burrough, Frank, 1996). The object model and the field model often compensate each other when depicting the spatial distribution of uncertainties.

4.2 Analysis of error propagation

The uncertainties in spatial data were first studied as errors of observed data (Mikhail, Ackermann, 1976), highlighting the measurement and handling ways of positional errors. Traditionally, the analysis of error propagation often gives the prior hypothesis that the error of input information is known, and then discusses the error of output information according to the theorem of error propagation. There are three alternative methods to analyze the propagation of attribute uncertainty, i.e., Taylor series method, Monte Carlo method and sensitivity analysis. Taylor series method is to approximate the function by a linear function that is locally a good approximation of the function. Monte Carlo method uses an entirely different approach to analyze the propagation of error through the GIS operation. But it is very difficult for both of the two methods to determine the relationships between input errors and output errors.

When it is necessary to assess the outcome quality of data mining while little prior knowledge of errors is known, the sensitivity analysis may be used to mainly study how the imposed perturbations (variations) of the input uncertainty influences the output knowledge, by adding simulated theoretical variables to disturb input information in spatial data mining. The referenced data are the input data without any disturbed variables. And different disturbed variables of input will get different analyzed results of output, which includes different errors, e.g., attribute, position, map deletion, polygon, confidence region. Lodwick et al. (1990) identified differential measures on raster data and map overlapping, associated with extrapolation, classification, differential scales or weights, and resolution. Bonin (1998) studied how the uncertainty was propagated in vector GIS, he (2000) further proposed a noise probability model to estimate attribute uncertainty on the basis of three parameters, i.e., deficit ratio, excess ratio, and confusion ratio. Mishra et al. (1999) studied the mistaken classification by fastening the errors on the true data.

However, they are all strictly mathematical. This may cause such incomprehensibility problems to the lay users without the background-associated knowledge that they may be unaware of, even misuse of the accurate descriptors, e.g., reliability diagrams and position error estimation. And it is also quite difficult to explain to them. Moreover, based on the man-made statistical simulation, the sensitivity analysis on spatial uncertainty imports theoretical errors in the context of probability theory and mathematical statistics. It needs a lot of data. And some issues of the sensitivity analysis may be studied further, for example, theoretical error of spatial uncertainty estimation, index to measure the uncertainty of parameter, measurement of how sensitive an attribute uncertain is, etc.

4.3 Probability theory and mathematical statistics

Probability theory is the classical mathematical theory on randomness (Arthurs, 1965). Probability theory and mathematical statistics study randomness via considering the stochastic probability that the case happens. And the probability is an indicator of the frequency or likelihood that an element is in a class. With probability theory and mathematical statistics, some theories and techniques, for example, spatial statistics, error band, epsilon band, “S” band, evidence theory, etc., were further put forward and applied (Shi, Wang, 2001, 2002).

Based on the classical crisp sets, spatial statistics studied the stochastic uncertainty (Cressie, 1991). For the error (confusion) matrix on the result of remote sensing image classification could not show the spatial distribution of uncertainty, the vector of probability was proposed (Shi, 1994). Evidence theory was an extension of probability theory (Shafer, 1976), and it could model the uncertainty of the mining process for image databases and other databases better than traditional probabilistic models.

However, the entities described with crisp sets-based methods have distinct boundary of attributes, which was not consistent to the reality world with uncertainties.

4.4 Extended sets

The crisp sets were extended to fuzzy sets (Zadeh, 1965), rough sets (Pawlak, 1991), geo-rough space (Wang, 2002), cloud model (Li, 1997), and so on.

Fuzzy sets characterize the fuzziness via the fuzzy membership value that an element belongs to a concept. And the fuzzy membership deals with the similarity of an element to a class (Zadeh, 1965). Fuzzy sets approach can be extended to spatial data mining, e.g., representing the uncertainty in the spatial relationships used in the spatial association rules mining. The problem with a fuzzy system is it is difficult to deal with too many features, membership functions, and rules. Fuzzy sets and probability theory are both valid approaches to the uncertainty. They 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. The fuzzy membership makes much more sense than the probability when describing how young a man is, while the probability makes much more sense than the fuzzy membership when predicting the outcome of a kid birth. However, neither of them can handle randomness and fuzziness at the same time.

Rough sets specify the uncertainty from incompleteness via a pair of upper approximation and lower approximation, and may identify cause-effect relationships in databases as a form of data mining (Pawlak, 1991). In the given universe of discourse, rough sets are incompleteness-based reasoning in the form of decision-making table. The lower approximation is the set of spatial elements that surely belong to the spatial entity, while the upper approximation is the set of spatial elements that possibly belong to it. The difference that the upper approximation minus the lower approximation leaves is the uncertain boundary. The uncertainty may be managed via incorporating rough sets into the underlying data model and through rough querying of crisp data. Because all uncertainties are generally considered in the boundary set, it is unable to decide whether the element in it belongs to the spatial entity or not.

4.5 Cloud model

In the real spatial world, there often exists more than one uncertainty at the same time, which has to be handled during the process of uncertainty-based spatial data mining. For example, both randomness and fuzziness are often included in spatial entities. In the uncertainty-based spatial data mining, abandon both of the uncertainties via traditional crisp mathematics? Consider only randomness without fuzziness, like probability theory and mathematical statistics? Consider only fuzziness without randomness as fuzzy sets? Think of randomness and fuzziness generally in an indeterminate boundary set? In fact, the complexity of a system is a rough inverse ratio of the precision to reach when the system is studied. If people emphasize particularly on the precision, they may be in hot water (Huang, 1997). Given some conditions, the certainties and uncertainties can be transformed one another. The precise entity in the macro-world may become uncertain in the micro-world. An inexact entity at a certain extent has arrived at some precision if the cognition when it is enough to match the decision-making.

Human natural language with an indeterminate boundary unifies the fuzziness and randomness. Being the carrier of thinking, the natural language represents the intelligence of human thinking and actions with various uncertainties. Although it is difficult to give an exact definition on a piece of natural language, and various people may understand the same natural language with different random meanings, the people can still intercommunicate with each other by using the natural language, e.g., transition between qualitative concept and quantitative data. Similar to the characteristics of the natural language, the cloud model may be an alternative to study spatial data mining in the contexts of randomness and fuzziness

The cloud model integrates the randomness and fuzziness by using the formalization-computerized language in a unified way, in which the advantages of soft computing in the natural language are absorbed (Li, 1997). The essential unit is the concept cloud composed of cloud drops, and the thinking is the precision considering both randomness and fuzziness. It depicts a qualitative concept with 3 numerical characteristics, i.e., Expected value (Ex), Entropy (En) and Hyper-Entropy (He). When lots of cloud drops form a piece of cloud on a concept, the uncertainty is also shown. The cloud model changes human qualitative experience and cognition into the rules of linguistic terms instead of an exact mathematical model. Soft computing-based cloud rule is consistent to real data distribution and human thinking, and hard computing is the special case. The mapping between qualitative concept and quantitative data are implemented with forward cloud generators and backward cloud generators via mathematical methods at any time. Its formal way to transform between quality and quantity may interpret the uncertain reasoning mechanism when more than one qualitative reasoning rules are activated at the same time. Moreover, the cloud model can automatically generate the concept hierarchy, which may improve the discovery efficiency of knowledge in different hierarchies, for the climb and jump of concepts is the basis of knowledge discovery at different hierarchies. So the cloud model may overcome the shortcomings of GIS data models, the difficulties of error propagation, the hard-computing deficiency of probability theory and mathematical statistics, the inherent shortage of membership function in fuzzy sets, the limitation of boundary set in rough sets and so on.

Now the cloud model has been applied in many fields. It was extended to discover the predictable rules with different time granularities (Yang, Li, 1998), classification rules on improving remotely sensed images (Di, 2001), Boolean association rule in the attribute concept (Du, Li, 2000), serial rules on periodical change and current trends with a pan-concept tree via cloud transform (Jiang, Li, Fan, 2000), characteristics rules and predicable rules on the movements of landslide, clustering rules on the datasets together with data fields (Wang, 2002; Wang et al., 2003), description and analysis of GIS attribute uncertainty (Shi, Wang, 2001, 2002). By using a cloud model-based qualitative control mechanism, Li (1999) carried out the intelligent control of the dynamic balance of a headstand pendulum. A new interpretation for the 24 solar-terms in lunar calendar was given in the cloud model by mapping the uncertain conditions and compendia environments to the uncertainly distributing cloud drops (Li, 1997, 2000).

5 Case study

The case study is on Baota landslide that locates in Yunyang, and in the region of Three Gorge on Yangtze River. When monitoring the movement of a landslide, it is unable to monitor all the points on the landslide. People often select the typical points and monitor them. And the movement rules of the landslide are discovered from the monitoring database on the typical points. Contrast to the whole data on the landslide movement, the monitoring data stored in the database are much more incomplete. The external aspect of incompleteness may come from the internal sources of randomness and fuzziness (Figure 1).

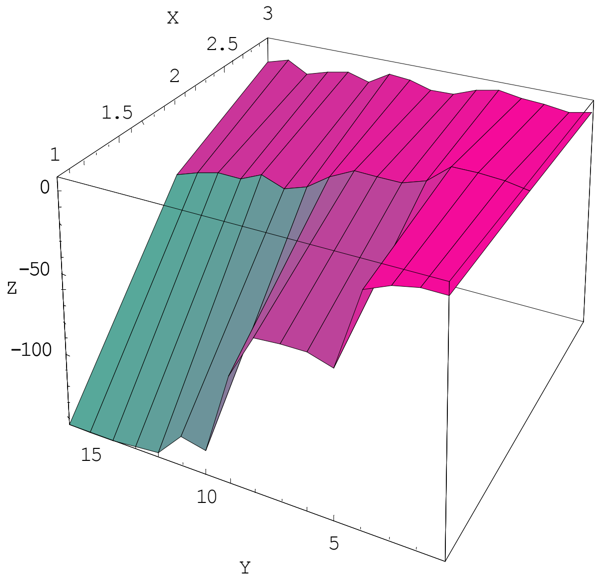

Figure 1. The external incompleteness from the internal randomness and fuzziness

The properties of *dx*, *dy,* and *dh*, are the measurements of displacements in X direction, Y direction and H direction of the landslide-monitoring points. From the observed landslide-monitoring values, the backward cloud generator can mine Ex, En and He of the linguistic term indicating the level of landslide displacement, i.e. gain the concept with the backward cloud generator. Then, with the three gained characteristics, the forward cloud generator can reproduce as many deterministic cloud-drops as you would like, i.e. produce synthetic values with the forward cloud generator (Li, 1997). Allthough there are differences between the synthetic landslide-monitoring values and the observed ones, their collective distribution is consistent. The synthetic landslide-monitoring values can also be taken as the landslide-monitoring values in the context of the three characteristics from the observed ones.

Figure 2 is the cloud-based knowledge on Baota landslide monitoring in X direction, which is the focus vertical direction of Yangtze River. In Figure 2, the symbol of "+" is the original position of monitoring point without movement, different rules are represented via different pieces of cloud, and the level of color in each piece of cloud denotes the discovered rules of a monitoring point. "BT11, …, BT34" are the serial numbers of Baota landslide monitoring point. Figure 2 indicates that all landslide monitoring points move to the direction of Yangtze River, i.e., south, or the negative direction of X axle. The displacements of the back part of Baota landslide are bigger than those of the front part in respect of Yangtze River, and the biggest exceptions are the displacements of monitoring point BT21. Furthermore, when Baota landslide was investigated (Wang, 2002), it was found out that the landslide had moved to Yangtze River, and a small size landslide had taken place near BT21. It matches the discovered spatial knowledge.

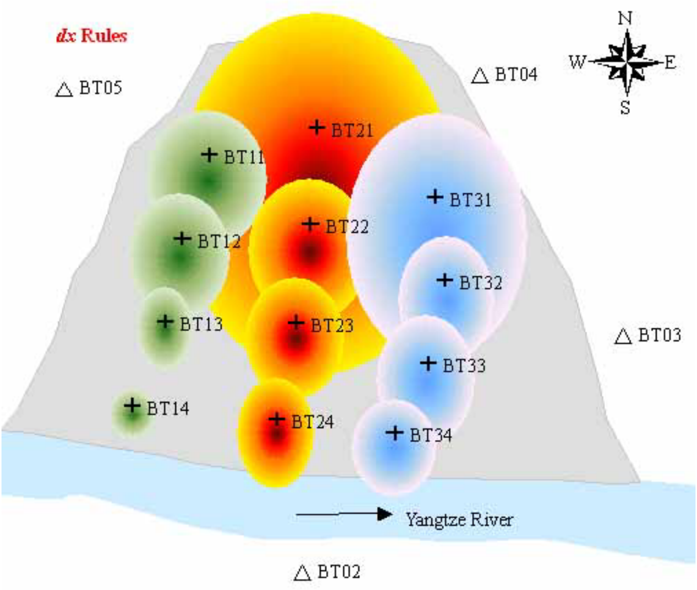

Figure 2. Spatial rules on monitoring points of Baota landslide

6 Conclusions

This paper proposed the uncertainty-based spatial data mining, together with its concepts, aspects, sources, and usable methods.

The uncertainty-based spatial data mining is to extract knowledge from the vast repositories of practical spatial data under the umbrella of uncertainties with the given perspectives and parameters. If the uncertainties are made good and right use of, it may be able to avoid the mistaken knowledge discovered from the mistaken spatial data.

The uncertainty mainly arises from the complexity of the real world, the limitation of human recognition, the weakness of computerized machine, and the shortcomings of techniques and methods. The external aspects of uncertainties may come from the internal sources in a given mathematical interpretation, which decides the selection of usable techniques.

The case study indicated that it was necessary to consider the inherent uncertainties in spatial data mining them. New techniques should be developed to handle the cases when there is more than one uncertainty in spatial data mining at the same time. A practical direction is the problem-oriented data mining with uncertainties.

Acknowledgements

The work described in this paper was fully supported by the funds from the CRC scheme, Research Grant Council of The Hong Kong SAR (Project No. 3_ZB40) and The Hong Kong Polytechnic University (Project No.: 1.34.9709), and partially National High Technology R&D Program (863) (Project No.: 2001AA135081).

References

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

AWREJCEWICZ J., 1989, *Bifurcation and chaos in simple dynamical systems* (Singapore: World Scientific)

BRUNK C., KELLY J. KOHAVI R., 1997, MinSet: An integrated system for data mining. Proceedings of 3rd International Conference on Knowledge discovery and Data Mining, Newport Beach, California, August, pp.135-138

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

BONIN O., 1998, Attribute Uncertainty Propagation in Vector Geographic Information Systems: Sensitivity Analysis. In *Proceedings of the Tenth International Conference on Scientific and Statistical Database Management*, edited by Kristine KELLY (Capri, Italy: IEEE Computer Society), pp.254-259

BONIN O., 2000, New Advances in Error Simulation in Vector Geographical Databases. In Accuracy 200: *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, edited by G.B.M. HEUVELINK, M.J.P.M.LEMMENS (Amsterdam, The Netherlands: University of Amsterdam), 59-65

CRESSIE N., 1991, *Statistics for Spatial Data*. (*New York*: John Wiley and Sons)

DI K.C., 2001, *Spatial Data Mining and Knowledge Discovery* (Wuhan: Wuhan University Press)

Du Y., Li D.Y., 2001, Cloud-based concept division and its application in associated rule mining. *Journal of Software*, 12(2): 196-203

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

FAYYAD U. M. et al., 1996, *Advances in Knowledge Discovery and Data Mining* (Menlopark CA: AAAI/MIT Press)

FISHER P.F., 1991, Modeling Soil map-unit Inclusions by Montecarlo Simulation. *International Journal of Geographical Information Systems*, 5(2), 193-208

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

HAINING R., 2003, *Spatial Data Analysis: Theory and Practice* (Cambridge: Cambridge University Press)

HAN J., KAMBER M., 2001, *Data Mining: Concepts and Techniques* (San Francisco: Academic Press)

JIANG R., LI D.Y., FAN J.H., 2000, Automatic generation algorithms on the pan-concept tree. *Journal of Computer*, 23(5): 470-476

KOPERSKI K. and HAN J., 1995, Discovery of spatial association rules in geographic information databases, *Proceedings of the 4th International Symposium on Large Spatial Databases*, Portland, MN, pp. 47-66

LI D.R., 1988, The Theory on Errors Processing and Reliability--The Development of Modern Photogrammetry (Beijing: Publisging House of Surveying and Mapping)

LI D.R., et al., 2001, On spatial data mining and knowledge discovery (SDMKD), *Geomatics and Information Science of Wuhan University*, 26(6):491-499

LI D.R., et al., 2002, Theories and technologies of spatial data mining and knowledge discovery. *Geomatics and Information Science of Wuhan University*, 27(3): 221-233

LI D.Y., 1997, Knowledge representation in KDD based on linguistic atoms. *Journal of Computer Science and Technology*, 12(6), 481-496

LI D.Y., 1999, Cloud model-based control method of the dynamic balance of a headstand pendulum with three grades. *China Engineering Science*, 1(2):41-46

LI D.Y., 2000, Uncertainties in knowledge representation. *China Engineering Science*, 2(10): 73-79

LODWICK, W.A. et al., 1990, Attribute error and sensitivity analysis of map operations in GIS: suitability analysis. *International Journal of Geographical Information Systems*, 4(4), 413-428

MCMASTER S., 1996, Assessing the impact of data quality on forest management decisions using geographical sensitivity analysis. *GISDATA'96 Summer Institute*

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

MISHRA J.K.et al., 1999, Remotely Sensed Data Based Information System: Considerations for Uncertainty, Errors and Limitations. In *Proceedings of the International Symposium on Spatial Data Quality'99*, edited by SHI W.Z., GOODCHILD M. F. and FISHER P.F.(Hong Kong, China: Department of Land Surveying & Geo-Informatics, The Hong Kong Polytechnic University), pp.605-615

PAWLAK Z., 1991, *Rough sets: theoretical aspects of reasoning about data* (London: Kluwer Academic Publishers)

SHAFER G., 1976, *A Mathematical Theory of Evidence* (Princeton: Princeton University Press)

SHI W.Z., 1994, *Modeling Positional and Thematic Uncertainties in Integration of Remote Sensing and Geographic Information Systems* (Enschede: ITC Publication)

SHI W.Z., WANG S.L., 2001，State of the art of research on the attribute uncertainty in GIS data. *Journal of Image and graphics,* 6[A](9): 918-924

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

SHI W.Z., GOODCHILD M. F., FISHER P., (eds), 2002, *Spatial Data Quality* (London: Taylor & Francis)

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

WANG S.L., WANG X.Z., SHI W.Z., 2002, Spatial Data Cleaning. *Proceedings of the First International Workshop on Data Cleaning and Preprocessing*,edited by Zhang Shichao, Yang Qiang, Zhang Chengqi, Maebashi TERRSA, Maebashi City, Japan, December 9th – 12th,pp.88-98

WANG S.L. et al., 2003, Geo-rough space. *Geo-Spatial Information Science*, 6(1): 11-19

WANG S.L. et al., 2003, A method of spatial data mining dealing with randomness and fuzziness. *Proceedings of the 2nd International Symposium on Spatial Data Quality*, edited by Wenzhong Shi, Michael F Goodchild, Peter F Fisher, Hong Kong, March 19th – 20th, pp.370-383

YANG C.H., LI D.Y.1998，Two dimensional cloud model and its application in prediction. *Journal of Computer*, 21(11): 961-969

ZADEH L.A., 1965, Fuzzy Sets. *Information and Control*. . 8(3): 338-353

ZADEH L.A., 1994, Soft computing and fuzzy logic. *IEEE Software*, 11(6): 48-56

ZHANG J.X., GOODCHILD M.F., 2002, *Uncertainty in Geographical Information* (London: Taylor & Francis)