Cloud Model-Based Spatial Data Mining

Shuliang Wang[1,2], Deren Li[1], Wenzhong Shi[2], Deyi Li[3], and Xinzhou Wang[1]

[1]National Laboratory for Information Engineering in Surveying Mapping and Remote Sensing
Wuhan University, Wuhan, Hubei, China,430079
[2]Department of Land Surveying & Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[3]Chinese Institute of Electronic System Engineering, No.20, Fuxing Road, Beijing, China, 100840

Abstract

In spatial data mining, we have to deal with uncertainties in data and mining process. The nature of the uncertainties can be, for example, fuzziness and randomness. This paper proposed a cloud model-based data mining method that may simultaneously deal with randomness and fuzziness. First, cloud model is presented, which is described by using three numerical characteristics, Ex, En and He. Furthermore, three visualization methods on cloud model are further proposed, which can be produced by the cloud generators. Second, cloud model-based knowledge discovery is further developed. In cloud model context, spatial data preprocessing pays more attention to data cleaning, transform between qualitative concepts and quantitative data, data reduction, and data discretization. Spatial uncertain reasoning is in the form of linguistic antecedents and linguistic consequences, both of which are implemented by X-conditional and Y-conditional cloud generators. Spatial knowledge is represented with qualitative concepts from large amounts of data, and also the cloud model. Finally, as an example, these methods are applied to mine Baota landslide monitoring database. The experimental results show that the cloud model can not only reduce the task complexity, and improve the operational efficiency, but also enhance the comprehension of the discovered knowledge.

I. INTRODUCTION

With the rapid development of Geo-spatial science and technology, tremendous volume of spatial data has been accumulated (Ester et al., 2000; Li et al., 2001). These accumulated data are in the nature of multivariable, nonlinear, uncertain, and even chaos, which has far exceeded human ability of using the data by conventional techniques, such as data analyzing functions in database management system, learning techniques of machine learning, or three mainstream approaches of symbol-based symbolicism, neurons-based connectionism and sense-feedback-act based behaviorism in artificial intelligence (Miller and Han, 2001; Wang, 2002). Compared with the conventional affair data, spatial data are more complex and with larger data volume. Spatial database stores not only positional data and attribute data, but also topological relationships, thematic layers, spatiotemporal scale, images and graphics. At the same time, the storage structure, query method, data analysis, database manipulation, etc. are all different conventional database. Furthermore, spatial data are accumulating continuously. In order to interpret and make full use of spatial data, the nontrivial knowledge will have to be extracted from them. But people are still short of the knowledge. Therefore, spatial data mining, a branch of data mining in Geo-spatial science, emerges.

Spatial data mining is to extract previously unknown, potentially useful, and ultimately understood rules from spatial data, and it is also named knowledge discovery from spatial databases, or spatial data mining and knowledge discovery (Piatetsky-Shapiro, 1994; Ester et al., 2000; Li et al., 2001). The discovered rules are associations with spatial objects at the cognitive hierarchy, and they may be description and prediction, for example, association rule, clustering rule, classification rule, characteristics rule, serial rule, predictive rule, and outlier. As a computerized simulation of human intelligence, spatial data mining discovers the patterns not only in a granularity world, but also among various granularity worlds. When dealing with different granularity worlds, the discovering manipulations of generalization and summarization are often soft computing with discrete linguistic terms instead of continuous data (Han and Kamber, 2001). Being an interdisciplinary discipline, spatial data mining is linked with cognitive science, data mining, artificial intelligence, machine learning, spatial analysis, mathematics, and so on. So many theories and methods can be employed in spatial data mining, e.g., probability theory, evidence theory, spatial statistics, fuzzy sets, rough sets, neural network, genetic algorithms, decision tree, visualization, online analytical process and data warehousing, rules induction, generalization and characterization of spatial objects, summarization and contrast data characteristics, classification and prediction, clustering and outlier analysis, similarity analysis in spatial databases (Wang, 2002). The discovered knowledge is supposed to be applied in many spatial aspects, e.g., judgment and decision-making support, intelligent GIS (geographical information science), knowledge driven interpretation and analysis of remote sensing images, knowledge based pattern recognition, knowledge engineering, integration of spatial techniques, etc. Now, a growing attention has been paid to spatial data mining (Miller and Han, 2002; Li et al., 2002).

1082-4006/03/0901~2-60 $5.00

However, there exist uncertainties in the objective data sets to be mined, transform between quantitative data and qualitative concept, knowledge discovery, and knowledge reasoning of spatial data mining. First, the uncertainties in spatial data indicate the difference between observed values and true values in spatiotemporal space, in the aspects of position, attribute, temporary, logical relationship and completeness (Shi and Wang, 2002). Second, spatial data mining has to perceive the variation and kinds of combinations, with different granularities to represent attributes of spatial objects' collective distribution in the attribute space when it generalizes and inducts a given set of quantitative data with the same feature category, the process of which is also uncertain. Third, the discovered knowledge is more generalized than the original data. And always the former are qualitative concepts, e.g., natural language, while the latter are quantitative practical data. It is further uncertain for data mining to mutual transform between quantitative data and qualitative concept. Fourth, spatial reasoning in data mining is under the umbrella of uncertainty. The essential issue of knowledge representation bridges the gap between the data and concept. But it is uncertain to represent the discovered knowledge according to human thinking. The uncertainties may directly or indirectly affect the quality of spatial data mining. If the uncertainties are carefully considered, it may be possible to avoid mistaken decision-making based on wrong information (Shi and Wang, 2002). Therefore, it is necessary to consider the uncertainties in spatial data mining, and apply the theories and techniques that deal with the uncertainties well.

In many cases of spatial data mining, the fuzziness and randomness often appear at the same time. Mathematical models and quantitative computation are always indispensable in current methods on the transformation between qualitative concept and quantitative data, e.g. hierarchy analysis, quantitative weighting, experts group marking, and qualitative analysis (Li, 1997). Based on them, a number of methods have been further developed, e.g., probability theory, evidence theory, spatial statistics, and error band are on random uncertainties, while fuzzy set and rough set are on imprecise uncertainties (Shi and Wang, 2002). But all of them are unable to deal with both the fuzziness and randomness. Then, people may have the following questions. How to represent the qualitative concept? How to indicate the fuzziness and randomness? How to realize the mutual transformation between qualitative concept and quantitative data, and indicate the ability of soft reasoning? The cloud model is an alternative to solve these problems in spatial data mining, because it can integrate the fuzziness and randomness in a unified way, and can transform between spatial concepts, i.e. qualitative basic linguistic terms, and data, i.e. quantitative values (Li et al., 1995).

II. CLOUD MODEL

Cloud model (Li et al., 1995) is a model of the uncertainty transition between a linguistic term of a qualitative concept and its quantitative numerical representation. In short, it is a model fro handling uncertainty transition between qualitative concept and quantitative representation. Let X be the set X={x}, as the discourse universe, and T a term associated with X. The membership degree of x in X to the term T, $C_T(x) \in [0,1]$, is a random number with a stable tendency. The cloud of T is a mapping from the discourse universe U to the unit interval [0,1], i.e.,

$C_T(x)$:X→[0,1] $x \in X\ x \to C_T(x)$

With the cloud model, the mapping from the discourse universe to the interval is a one-point to multi-point transition, i.e. a piece of cloud while not a membership curve. As well, the degree that any cloud drop represents the qualitative concept can be specified. A piece of cloud is made up of many cloud drops, visible shape in a whole, but fuzzy in detail, which is similar to the natural cloud in the sky. Any one of the cloud drops is a mapping in the discourse universe from qualitative concept, i.e. a specified realization with uncertain factors.

Several kinds of cloud models have been developed to match different demands, such as basic clouds, floating clouds, synthesized clouds, resolved clouds, geometric clouds, etc. (Li, 1997; Di, 1999; Wang, 2002). Basic clouds are the element clouds, e.g., normal cloud based on normal distribution. They may be directly generated by a set of data. Excluding the basic clouds, the other clouds, which are constructed by given clouds, are called virtual clouds. Floating cloud mechanism is used to generate default clouds in the blank areas of the universe by other given clouds. If we consider a universe as a linguistic variable and we want to represent linguistic terms by clouds, the only indispensable work is to specify key clouds at the key positions. Other clouds can be automatically generated by the floating cloud construction method. A synthesized cloud is used to synthesize linguistic terms into a generalized one. If we use the mechanism of synthesized cloud construction recursively from low concept levels to high concept levels, we can get concept hierarchies for linguistic variables, which are very important in data mining and knowledge discovery. The algorithms of floating and synthesized cloud generation were presented in. Resolved cloud method is used to decompose large concept to several small ones.

Numerical characteristics

The cloud model has three numerical characteristics, Expected value (Ex), Entropy (En) and Hyper-Entropy (He), which integrates the fuzziness and randomness of spatial concepts in a unified way. In the discourse universe, Ex is the position corresponding to the center of the cloud gravity, whose elements are fully compatible with the spatial linguistic concept; En is a measure of the concept coverage, i.e. a measure of the spatial fuzziness, which indicates how many elements could be accepted to the spatial linguistic concept; and He is a measure of the dispersion on the cloud drops, which can also be considered as the entropy of En. Figure 1 shows the linguistic term "displacement is 9 millimeters (mm) around" with three numerical characteristics, i.e., Ex=9, En=0.5, and He=0.02.

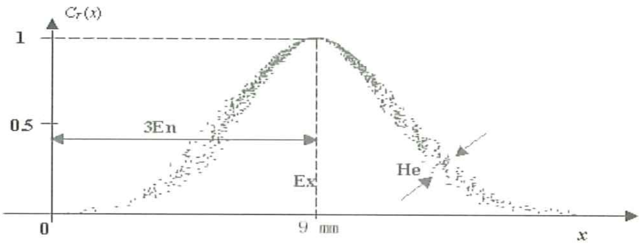


Figure 1. Three numerical characteristics of "displacement is 9 millimeters around."

Given three numerical characteristics Ex, En and He, the cloud generator can produce as many drops of the cloud as you would like. In the extreme case, {Ex, 0, 0}, where both the entropy and hyper entropy equal to zero, denotes the concept of a deterministic datum, and the greater the number of cloud drops, the more deterministic the concept.

Visualization methods

In the discourse space, there are three kinds of visualization methods to illustrate the cloud graph including all the cloud drops.

(a) Dot with gray degree. One dot specifies the position of one cloud drop, and its gray degree indicates how certain the cloud drop represents the concept;

(b) Circle (or ball) with scale. One cloud drop is depicted by one circle, which represents the position of cloud drop, and the scale of the circle indicates how certain the cloud drop represents the concept; and

(c) N+1 dimensions. N dimensions specify the positions of the cloud drop, while another dimension axis to denote the significance of the certain degree of the cloud drop representing the concept.

Here is an example on an elementary concept " the neighborhood of the coordinate origin in two-dimension plane", which is used to explain the transition between the deterministic data drops, i.e. cloud drops, and the spatial concept. Let the expected value of the concept be Ex={0,0}, entropy be En={0.1,0.1}, and hyper entropy be He ={0.01,0.01} respectively. With the above methods, three kinds of cloud graphs (Figure 2(a), Figure 2(b) and Figure 2(c)), each of which has 1000 cloud drops to represent this concept, can be produced by a forward cloud generator. Seen from these three Figures, the bigger the distance between the cloud drop and the coordinate origin, the smaller the degree of certainty. Hence, the cloud model well integrates the fuzziness and randomness of linguistic concepts in a unified way.

Cloud generators

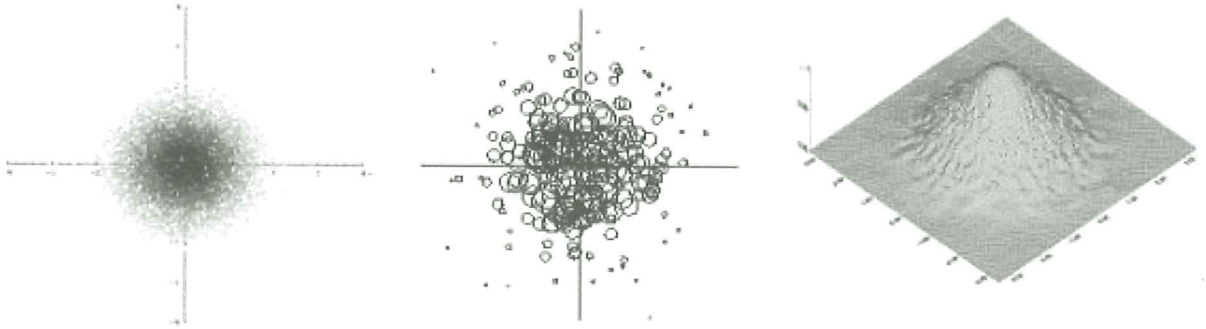
The above three visualization methods are all implemented with the forward cloud generator in the context of the given {Ex, En, He}. Despite of the uncertainty in the algorithm, the positions of cloud drops produced each time are deterministic. Each cloud drop produced by the cloud generator is plotted deterministically according to the position. On the other hand, it is an elementary issue in spatial data mining that spatial concept is always constructed from the given spatial data, and spatial data mining aims to discover spatial knowledge represented by a cloud from the database. That is, the backward cloud generator is also necessary. It can be used to perform the transition from data to linguistic terms, and may mine the integrity {Ex, En, He} of cloud drops specified by many precise data points. Because it is common and useful to represent spatial linguistic atoms (Li et al., 2001), the normal compatibility cloud will be taken as an example to study the forward and backward cloud generators in the following.

The input of the forward normal cloud generator is three numerical characteristics of a linguistic term, (Ex, En, He), and the number of cloud-drops to be generated, *N*, while the output is the quantitative positions of *N* cloud drops in the data space and the certain degree that each cloud-drop can represent the linguistic term. The algorithm in details is:

[1] Produce a normally distributed random number En' with mean En and standard deviation He;

[2] Produce a normally distributed random number x with mean Ex and standard deviation En';

[3] Calculate $y = e^{-\frac{(x-Ex)^2}{2(En')^2}}$;



(a) Dot with gray degree (b) Circle with scale (c) N+1 dimensions

Figure 2. Three kinds of visualization methods to illustrate the cloud graph

[4] Drop (x_i, y_i) is a cloud-drop in the universe of discourse; and

[5] Repeat step 1-4 until *N* cloud-drops are generated.

Simultaneously, the input of the backward normal cloud generator is the quantitative positions of N cloud-drops, x_i (i=1, …, *N*), and the certainty degree that each cloud-drop can represent a linguistic term, y_i(i=1, …, *N*), while the output is the three numerical characteristics, Ex, En, He, of the linguistic term represented by the *N* cloud-drops. The algorithm in details is:

[1] Calculate the mean value of x_i (i=1, …, *N*),

$$Ex = \frac{1}{N}\sum_{i=1}^{N} x_i \;;$$

[2] For each pair of (x_i,y_i), calculate

$$En_i = \sqrt{-\frac{(x_i - Ex)}{2\ln y_i}} \;;$$

[3] Calculate the mean value of En_i (i=1,…,*N*),

$$En = \frac{1}{N}\sum_{i=1}^{N} En_i \;;$$

[4] Calculate the standard deviation of En_i,

$$He = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(En_i - En)^2} \;.$$

With the given algorithms of forward and backward cloud generators, it is easy to build the mapping relationship inseparably and interdependently between qualitative concept and quantitative data. The cloud model improves the weakness of rigid specification and too much certainty, which comes into conflict with the human recognition process, appeared in commonly used transition models. Moreover, it performs the interchangeable transition between qualitative concept and quantitative data through the use of strict mathematic functions, the preservation of the uncertainty in transition makes cloud model well meet the need of real life situation. Obviously, the cloud model is not a simple combination of probability methods and fuzzy methods.

III. KNOWLEDGE DISCOVERY

Cloud model may be used in data preprocessing, transform between quality and quantity, uncertainty reasoning, knowledge representation, etc.

Spatial data preprocessing

Spatial data are polluted by incompleteness, inaccuracy, error, repetition, inconsistence, fuzziness, randomness, heterogeneous data and so on. It has more chance to discover useful and interesting knowledge from preprocessed data than from dirty data. And the knowledge from trusted data is also trusted. Therefore the objective data to be mined have to be preprocessed in order to improve the discovery performance, e.g., data cleaning, data transformation, data reduction, and data discretization (Wang and Shi, 2002).

Spatial data cleaning is to make dirty data cleaner. When there are tuples with incomplete attribute values (null or unknown), people have tried to ignore the incomplete tuples, or fill in the missing value with a global constant or a mean of all the existing attribute values. Considering both fuzziness and randomness, cloud model may fill in the missing values of an attribute with the most probable value, i.e., one of its numerical characteristics, Expected value (Ex), and further show its confidential level with its other two numerical characteristics, En and He at the same time. Spatial data transform is to change an inappropriate data form into an appropriate form for specific mining algorithms, i.e., normalization, smoothing, aggregation and conceptualization. Thereinto, the transform between qualitative concepts and their quantitative expressions play an important role. As described in section 2, cloud model is good at the mutual transform between qualitative concepts and their corresponding quantitative data. It bridges the gap between qualitative concept space and quantitative data space, and the conceptual space represents different concepts in the same characteristic category. Although the representations of the data set are changed, the final generated knowledge should be the same as the data without any transform.

Spatial data reduction is to reduce the amounts of spatial data in the context that the knowledge from the reduced data is as the same as the original data without reduction, e.g., dimension reduction, data compression, volume reduction of parametric regression and non-parametric histograms, clustering and sampling (Miller and Han, 2001). Dimension reduction, also named attribute reduction, is to search for the minimum set of attributes in the attribute space of tuples with the same feature. Necessary attributes are selected while redundant attributes are eliminated. Dimension reduction has to ensure that the rules of the same feature category are kept unchanged, or the rules from dimension-reduced data are as close as possible to the rules from all original attributes without reduction, e.g., probability distribution of data classes. Data compression reduces the size of spatial data in order to save storage space and mining time. Contextually, whatever the compression is lossless or lossy, mining algorithms can directly manipulate the compressed data without uncompressing it. Otherwise, it is not beneficial for spatial data mining. Cloud model extracts three numerical characteristics from the given data. These numerical characteristics are the rules of probability distribution on the given data. With forward cloud generator, they can produce data as many as you would like. Although the produced data may not exact the given data, both of their rules of probability distribution are consistent to each other. Furthermore, data on low level may be replaced with data-concept hierarchy of high level under the umbrella of the conceptual space generated via cloud model. From the conceptual space, the feature space will be produced to depict complicated spatial objects with multi-properties. The rules are easily uncovered in the context of the feature space. That is, huge amounts of spatial data are replaced by only three parameters with the compression of cloud model, which is an appropriate ratio of spatial data reduction.

Spatial data discretization is to reduce the number of values for a given continuous attribute, via dividing the range of the attribute into intervals. Interval labels are then used to replace actual data values. Some data mining algorithms only accept categorical attributes and cannot handle a range of continuous attribute value, e.g., rough sets. At the same time, the data reduction process to abstract spatial objects may simplify the problem and lessen the data amount greatly when the interest of spatial data mining is changed from the fine granularity world to the coarse one. However, the traditional discretization algorithms are hard partition without considering the human thinking characteristics of spatial data mining. Cloud model-based spatial data discretization is soft partitioning. It reduces the data set, and also generates pan-concept hierarchies automatically. Figure 3 and Figure 4 gives such a pan-concept hierarchical tree on displacement attribute landslide. In the pan-tree structure, a given node may have more than one father nodes, i.e., the nodes“9 millimeters around” has two father nodes “very small” and “smaller”, while in a proper tree structure, the given son node has only one father node.

The data reduction process to abstract spatial objects may simplify the problem and lessen the data amount greatly when the mining interest is changed from the fine granularity world to the coarse one. At coarse granularity, i.e. observe problem at long displacement, the fine difference is overlooked and the commonness, which is more profound than the individuality and help understand problem comprehensively, is gained. Otherwise at fine granularity, i.e. analyze problem closely, more attention is paid on the individuality, which is more vivid than commonness. The higher the abstraction level, the greater the generalization extent, and the smaller the physical knowledge size is, which can be made different combination and condensation according to relevant tasks.

Uncertain spatial reasoning

Logic reasoning is principal spatial data mining. Deduction and induction are two main methods. Deduction is based on a set of most generic, succinct, and universally applicable theory or axiom system, just as classical physics. And the reasoning process of deduction is from generic theory to special cases. However, it is impossible to build “Newton Law” in such computerized science as data mining, cognitive science and artificial intelligence. Induction reasoning is executed from special cases to generic theory, which is the common approach to obtain new knowledge. In spatial data mining, deduction is the process of knowledge application while induction is the process of data mining. Taking advantage of human intelligence, the reasoning of spatial data mining should be the generic theory assembling evidences, similarly to the molecular biology (Wang, 2002), which is uncertain.

Cloud model-based spatial uncertain reasoning is in the form of “If $A_1, A_2, \ldots, A_m$, then $B_1, B_2, \ldots, B_n$”, where $A_1, A_2, \ldots, A_n$ are linguistic antecedents, and $B_1, B_2, \ldots, B_n$ are linguistic consequences. According to the values of m, n, we have the detailed reasoning forms of one-factor and one-rule, multi-factor and one-rule, one-factor and multi-rule, and multi-factor and multi-rule. One X-conditional cloud followed by the other Y-conditional cloud represents a spatial rule, e.g., “If elevation is low, then the road density is high”. Figure 5 is a reasoning generator of one-factor and one-rule that is “If A, then B”, and its output cloud with one input.

In Figure 5(a), CG_A is the X-conditional cloud generator for linguistic term A, and CG_B is the Y conditional cloud generator for linguistic term B. Given a certain input x_A, CG_A generates a

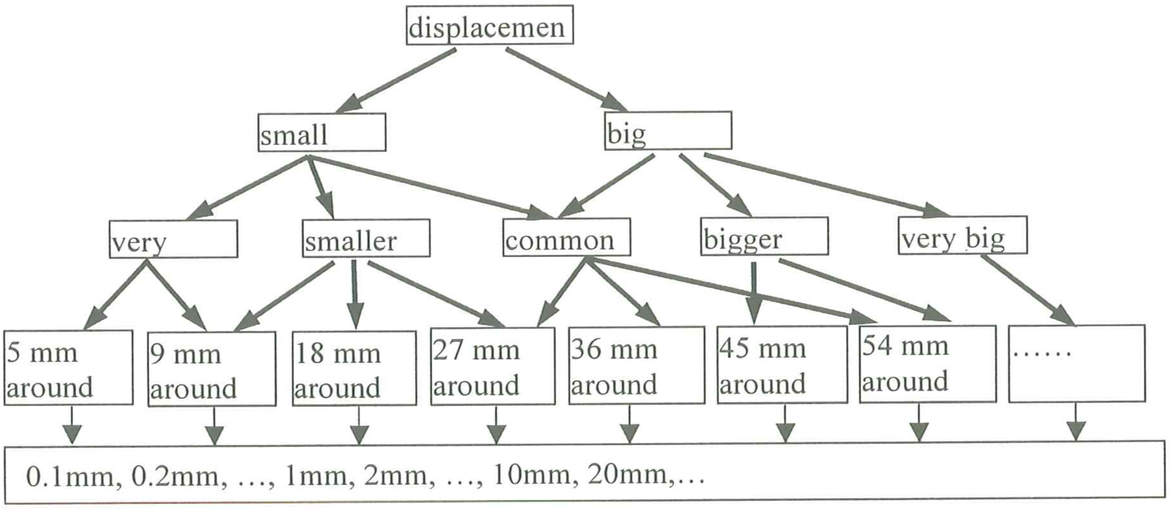


Figure 3. A pan-concept hierarchical tree from displacement data

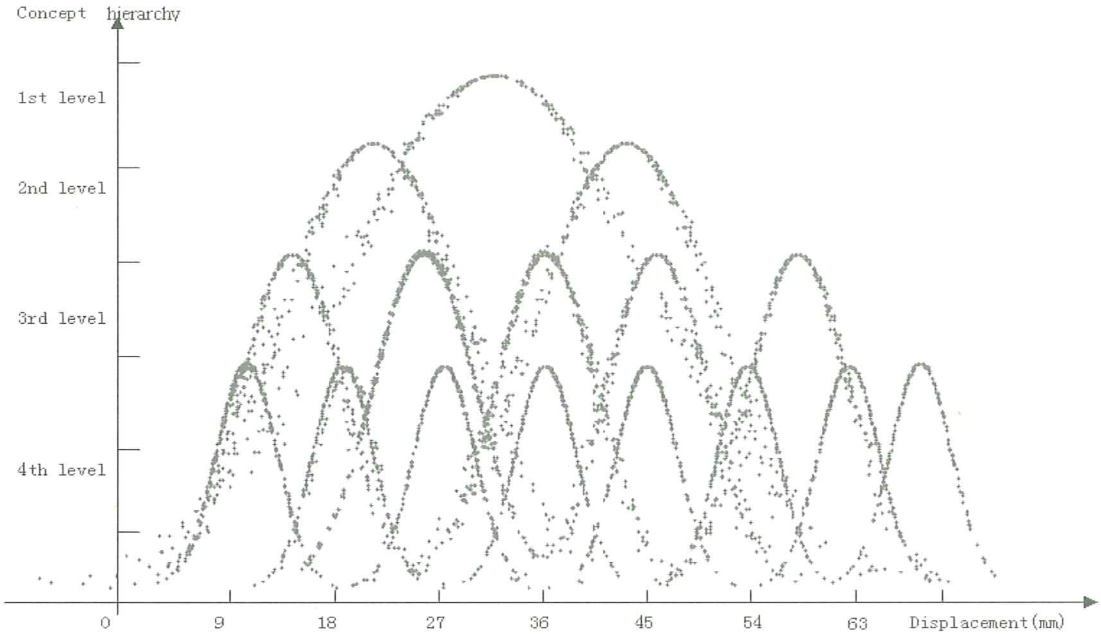
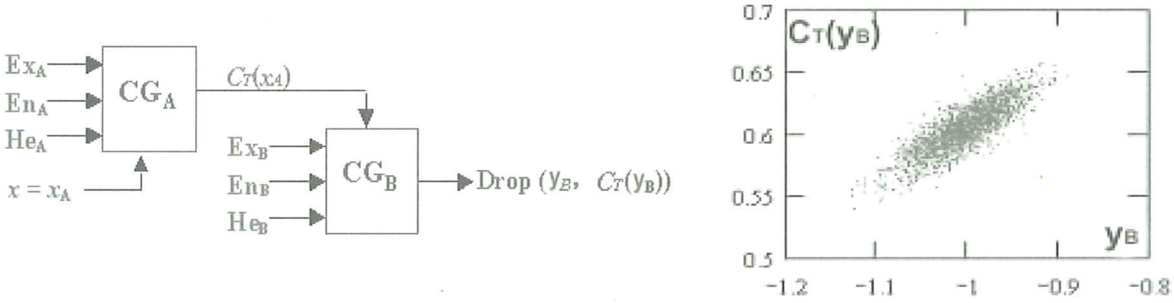


Figure 4. Cloud model-based representation of the pan-concept hierarchical tree

series of random values $C_T(x_A)$. These values are considered as the activation degree $C_T(x_B)$of the rule and input to CG$_B$, i.e., $C_T(x_A) = C_T(x_B)$. The final outputs are cloud drops, which forms a new cloud. So the reasoning algorithm is the combination of the algorithm of X and Y conditional cloud generators (Di, 1999; Wang, 2002). Figure 5(b) is the output cloud of a one-factor and one-rule generator with one input. It shows that cloud model-based reasoning generates an uncertain result, a piece of cloud composed of many cloud drops. The uncertainty of the linguistic terms in the rule is propagated during the reasoning process. The closer to the expected values, the more focusing the cloud band, while the farther to the expected value, the more dispersed the cloud band, which matches the intuition of human being. Because the rule output is a piece of cloud instead of a datum, the final result may be given in several forms. That is, (a) one random value; (b) several random values as sample results; (c) expected value that is the mean of many sample results; and (d) linguistic term represented by a cloud model whose three parameters are obtained by backward cloud generator (Li, 1997). The above mentioned further indicates that the cloud model-based uncertain reasoning is more flexible and powerful than the conventional reasoning methods, i.e., fuzzy reasoning.

In the contexts of the one-factor and one-rule reasoning and cloud generator algorithms, it is able to contextually give the algorithms of other uncertain reasoning forms. Usually, there are many rules in a real knowledge base. Multi-rule reasoning is frequently used in an intelligent GIS or a spatial decision support system. The main idea of the multi-rule reasoning algorithm is that when several rules are activated simultaneously, a virtual cloud is constructed by the geometric cloud method. Because the property of least square fitting, the final output is more likely to close to the rule of high activated degree. The one-factor and one-rule reasoning method can be easily extended to multi-factor multi-rule reasoning on the basis of multi-dimensional cloud models. For example, the two-factor and two-rule generator may combine a two-dimen-



(a) Rule generator (b) Output cloud with one input

Figure 5. Rule generator and output cloud of one-factor and one-rule reasoning

sional X-conditional cloud generator and a two-dimensional Y-conditional cloud generator.

Representation of spatial knowledge

Natural language of human is one of the best ways to represent spatial knowledge (Li, 1997). The discovered knowledge is associated with spatial objects at the cognitive hierarchy, and it is essentially important to properly represent them in spatial data mining. It is acknowledged that spatial qualitative concept is more understandable, direct and precise than spatial data. Among various opinions on how concept is formed, both feature opinion and prototype opinion agree that all the samples are figured by a set of datum (Li et al., 2000). Describing the quantitative concept with linguistic terms obviously weights more than that with precise math equations. The more abstract the knowledge to be discovered, the greater the advantage. Being the carrier of thinking, natural language takes linguistic term as the basic unit, and the linguistic atom of minimum linguistic term is the basic cell of human thinking. The linguistic atom is corresponding to the most elementary concept. With the most elementary concepts and their various combinations, natural language is able to describe a complicated concept. Moreover, natural language can well deal with the spatial uncertainties, especially the randomness and fuzziness, between qualitative concept and quantitative data. The advantage of cloud model is much similar with natural language.

Cloud model depicts the granularity and potential of spatial concepts naturally. The attribute helps to discover the relationship between attributes, the feature perceives the consistency and difference between spatial objects, and the knowledge indicates the increase of concept granularity of attributes value. The intension and extension of various conceptions should be clearly defined so that we can know whether the concept granularity is big or small, and know the conceptions' inner relationship, resembling one or subordinative one. The granularity of a basic conception can be measured by its representative cloud's three mathematical characteristics. Ex marks the center of the cloud, En signs the size of the cloud, and He denotes the agglutinability of the cloud drops and shows the constrictive level of the whole cloud. For example, the category of "displacement" can be looked upon as a language variable of many language terms with different granularity, i.e., indicate 9 millimeters around, very small, small, etc. And the conception of "a small displacement" has larger En than the conception of "9 millimeters around", which means the former has larger granularity than the latter (Li et al., 2001). Each datum makes its own contribution to forming the conception, which may be measured by the potential of its corresponding cloud drop in the conceptual space. The potential is determined by both the position and certainty of a cloud drop. With a higher certainty, the cloud drop may have greater contribution to the potential in the conceptual space. That people observe and analyze the spatial databases from different perspectives is similar to select different functions to calculate the potential of a point in the conceptual space. The isopotential of a large number of cloud drops (Wang, 2002) will naturally take shape a pan-concept hierarchy (Figure 4). Moreover, different attributes have different pan-concept trees, and there are lots of combination states of spatial objects, which are usually decided by spatial data mining task.

Cloud model represent the discovered knowledge with linguistic rules made by linguistic terms of qualitative concepts. It is known that the attributes of objects in spatial databases, e.g., location, elevation, distance, direction, are the linguistic variables. And each linguistic variable is mapping to several linguistic terms with various granularities and potentials. These linguistic terms can be characterized via cloud model. A one-dimensional cloud represents the linguistic terms for elevation and distance, such as high, low, near, far, etc. And a two-dimensional cloud represents the linguistic terms of variable location, e.g., southeast, southwest, northeast, and northwest. So cloud models represent the fuzziness and randomness of the knowledge in a unified way. If we have enough sample data for a concept, three cloud parameters can be automatically derived by cloud transform method, which decomposes a data distribution to the sum of several clouds. Otherwise, the user has to manually give the parameters of some key clouds and the other clouds will be automatically constructed by virtual cloud methods when needed (Di, 1999).

Conceptual space represents different concepts in the same characteristic category, while feature space depicts complicated spatial objects with multi-properties. The same idea can also be extended to spatial objects with attributes in their feature space. The granularity world becomes bigger when spatial data mining jumps up from the conceptual space to the feature space. N properties or characteristics compose an N-dimension feature space. Each spatial object becomes a point in the feature space, and thousands of objects become thousands of points in the space. Each point makes its own contribution to forming the spatial category in the feature space. All the points form the isopotential spontaneously. Intuitively, these points can be grouped naturally into clusters, and the isopotential of all objects will automatically take shape clusters and clustering hierarchy at this time. These clusters represent different kinds of spatial objects recorded in the database, and naturally form the cluster spectrum graph. In the feature space, potential presents the contributions made by all samples in the original data set, and the clustering result is conducted with the whole data set. Ordinary clustering methods always split the original data set into two parts, the training data set to generate the clusters and the testing data set to prove the validity of the methods. However, it is unable to naturally and convincingly determine the proportion of the two parts and choose the samples to be the training part. Furthermore, each sample has its own contribution to the clustering result, but ordinary methods neglect the samples in the testing data set. Hence, the clustering result from the feature space with the potential of cloud drops is more reasonable.

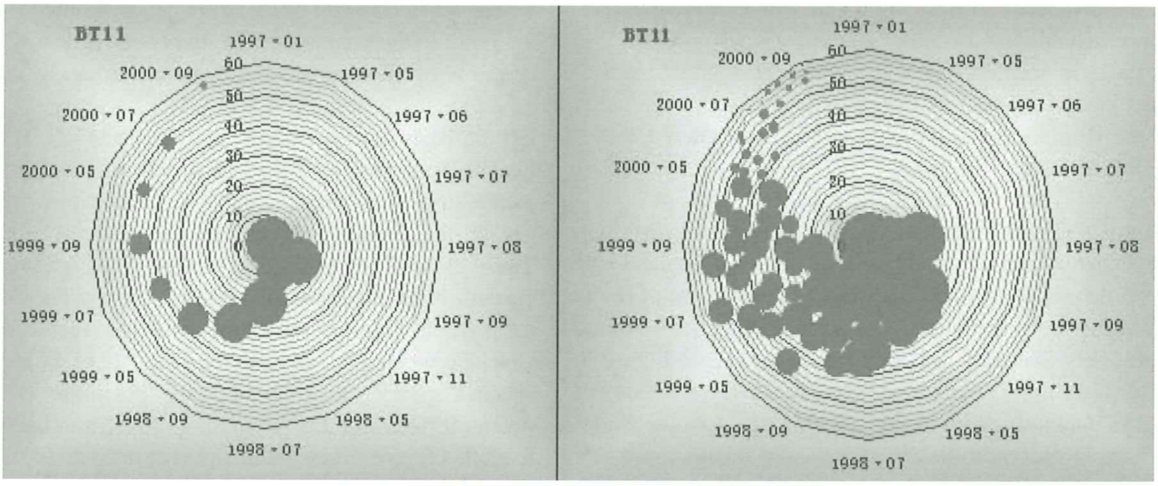
IV. CASE STUDY

In order to verify the feasibility and effectiveness of the cloud model for spatial data mining, Baota landslide-monitoring data mining is studied as a case. Baota landslide locates in Yunyang, Chongqing, China, in the region of Three Gorge on Yangtze River. The landslide monitoring started from June 1997. Up to now, this database has accumulated to 1G bytes, and all the attributes are numerical displacements, i.e. *dx*, *dy,* and *dh*. Respectively, 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, and |*dx*|, |*dy*| an |*dh*| are their absolute values. In the following, it is noted that all spatial knowledge is discovered from the databases with the properties of *dx*, *dy,* and *dh*, while |*dx*|, |*dy*| an |*dh*| are only used to visualize the results of spatial data mining. And the properties of *dx* are the major examples.

The linguistic terms of different displacements on *dx*, *dy* and *dh* are depicted by the pan-concept hierarchy tree in the conceptual space, which are formed by cloud models on the basis of monitoring data (Figure 3, Figure 4). 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. Figure 6 further gives such an example on landslide-monitoring point BT11 in X direction with the linguistic concept "the displacements are big south, high scattered and instable".

Seen from Figure 6, the consistence of the collective distribution between them is still obvious although there are differences between the synthetic landslide-monitoring values and the observed ones. Therefore, 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. According to the landslide-monitoring characteristics and demands, let the linguistic concepts of "smaller(0~9mm), small(9~18mm), big(18~27mm), bigger(27~36mm), very big(36~50mm), extremely big(>50mm)" with Ex, "lower (0~9), low(9~18), high(18~27), higher(27~36), very high(36~50), extremely big(>50)"with En, "more stable (0~9), stable (9~18), instable(18~27), more instable (27~36), very instable (36~50), extremely instable (>50)" with He respectively depict the movements, scattering levels and stabilities of the displacements, then the rules on Baota landslide-monitoring in X direction can be discovered from the databases in the conceptual space. Figure 7 is the cloud based knowledge on Baota landslide monitoring in X direction, which is the focus vertical direction of Yangtze River. In Figure 7, 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 7 indicates that all landslide monitoring points move to the direction of Yangtze River, i.e., south, or the negative direction of X axle. However, the displacements are different from each other. The displacements of monitoring point BT21 are extremely big south, extremely high scattered and extremely instable, and BT31is behind BT21. The displacements are smaller south, lower scattered and more stable, which is the movements of monitoring point BT14. Generally speaking, 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. These cloud model-based rules in Figure 7 may also be described as the following qualitative concepts.



(a) 17 observed values of *dx* (b) 100 synthetic values of *dx*

Figure 6. Gain the concept and produce synthetic values

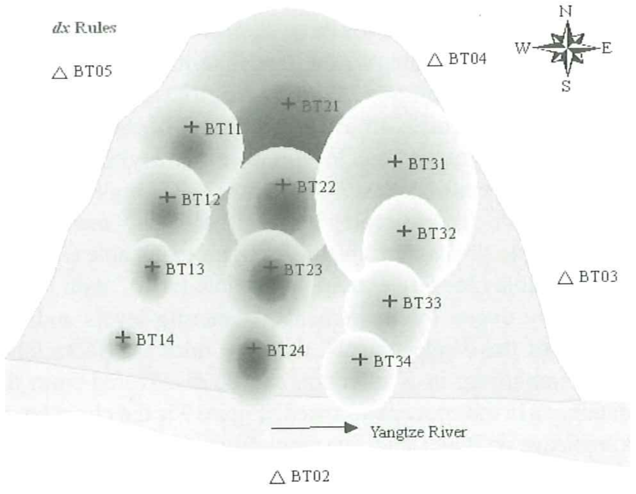


Figure 7. Spatial rules on monitoring points of Baota landslide

BT11: the displacements are big south, high scattered and instable;
BT12: the displacements are big south, high scattered and very instable;
BT13: the displacements are small south, lower scattered and more stable;
BT14: the displacements are smaller south, lower scattered and more stable;
BT21: the displacements are extremely big south, extremely high scattered and extremely instable;
BT22: the displacements are bigger south, high scattered and instable;
BT23: the displacements are big south, high scattered and extremely instable;
BT24: the displacements are big south, high scattered and more instable;
BT31: the displacements are very big south, higher scattered and very instable;
BT32: the displacements are big south, low scattered and more instable;
BT33: the displacements are big south, high scattered and very instable; and
BT34: the displacements are big south, high scattered and more instable;

Figure 8 is the generalized result at a higher hierarchy than that of Figure 7 in the feature space, i.e. the displacement rule of the whole landslide. It is "the whole displacement of Baota landslide are bigger south (to Yangtze River), higher scattered and extremely instable". Based on Figure 7 and Figure 8, spatial data mining is particular views for a viewer to look at the spatial database on the displacements of Baota landslide-monitoring by different distances only, and a longer distance leads a piece of more meta-knowledge to be discovered. Because large amounts of consecutive data are replaced by discrete linguistic terms, the efficiency of spatial data mining is improved. Meanwhile, the final result mined is also stable due to the randomness and fuzziness of concept indicated by the cloud model.

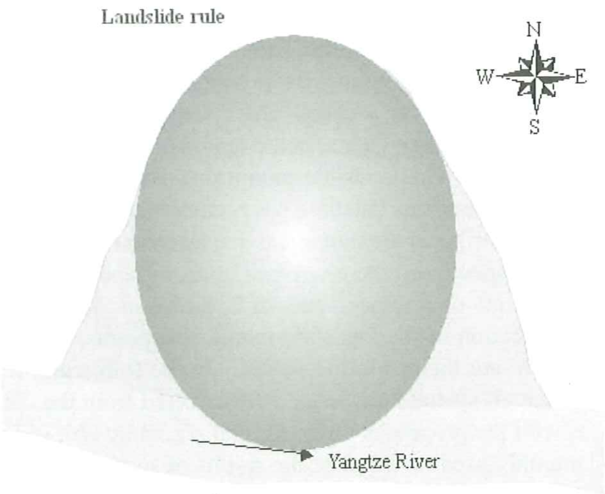


Figure 8. Spatial rules on Baota landslide

Further, let the |*dx*|-axis, |*dy*|-axis respectively depict the absolute displacement values of the landslide- -monitoring points. The certainty of the cloud drop (dx_i, $C_T(dx_i)$), $C_T(dx_i)$ is also defined as,

$$C_T(dx_i) = \frac{dx_i - \min(dx)}{\max(dx) - \min(dx)} \qquad (1)$$

where, max(*dx*) and min(*dx*) are the maximum and minimum of $dx = \{dx_1, dx_2, \ldots, dx_i, \ldots, dx_n\}$. The potential *p* of a cloud drop in the number universe is the sum of all data potentials.

$$p = k \cdot e^{-\sum_{i=1}^{N} \frac{r_i^2}{C_T(dx_i)}} \qquad (2)$$

where, k is a constant of radiation gene, r_i is the distance from the point to the position of the *i*th observed data, $C_T(dx_i)$is the certainty of the *i*th data, and *N* is the amount of the data.

On the basis of equation (1) and equation (2), all the above landslide-monitoring points form the potential field and the isopotential lines spontaneously in the feature space. Intuitively, these points can be grouped naturally into clusters. These clusters represent different kinds of spatial objects recorded in the database, and naturally form the cluster spectrum graph. Figure 9 shows the visualized results of the landslide-monitoring points with the property of dx in the feature space. Figure 9 (a) depicts all points' potential. They form the potential field and the isopotential lines spontaneously. Seen from this figure, when the hierarchy jumps up from Level 1 to Level 5, i.e. from the fine granularity world to the coarse granularity world, these landslide-monitoring points can be intuitively grouped naturally into different clusters at different hierarchies of variant levels. That is,

[1] No clusters at the hierarchy of Level 1;

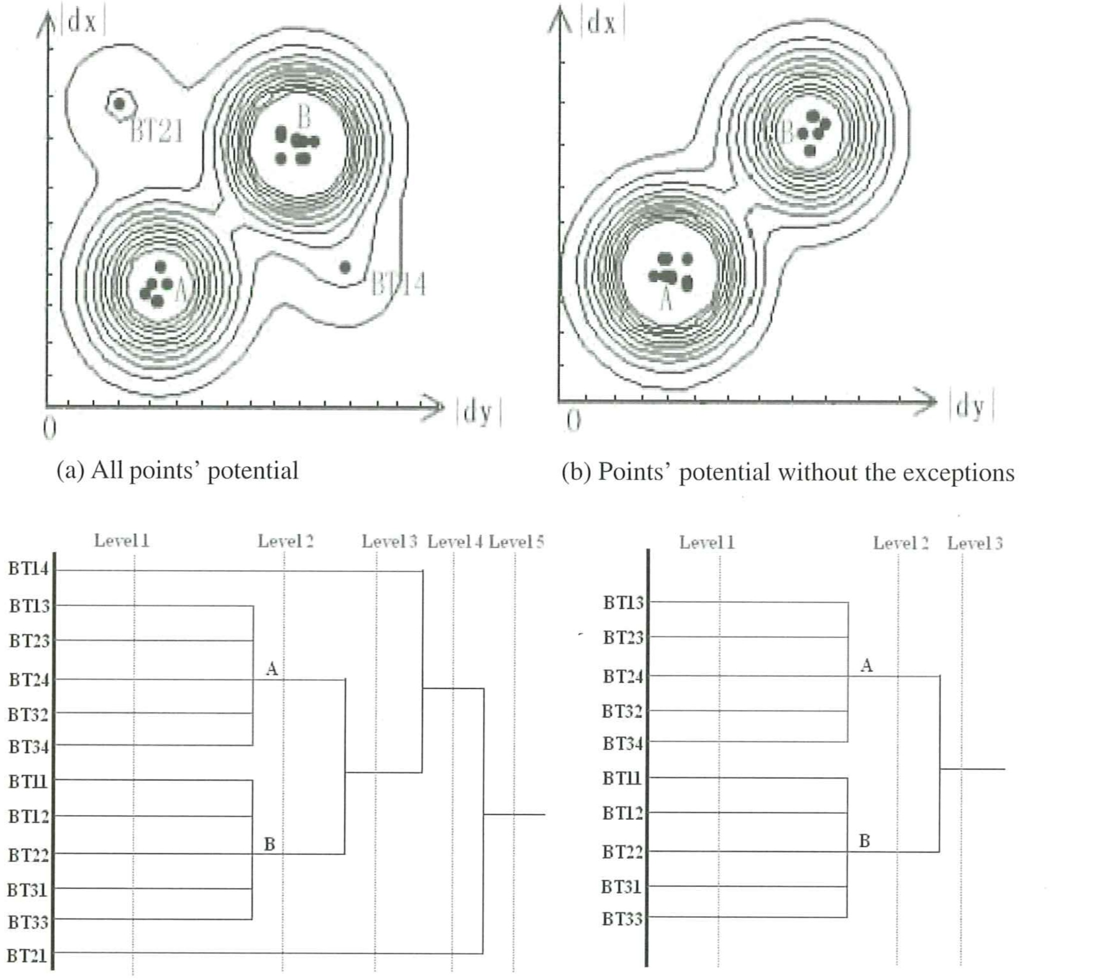
[2] Four clusters at the hierarchy of Level 2 that are cluster BT14, cluster A (BT13, BT23, BT24, BT32, BT34), cluster B (BT11, BT12, BT22, BT31, BT33) and cluster BT21;

[3] Three clusters at the hierarchy of Level 3 that are cluster BT14, cluster (A, B) and cluster BT21;

[4] Two clusters at the hierarchy of Level 4 that are cluster (BT14, (A, B)) and cluster BT21; and

[5] One cluster at the hierarchy of Level 5 that is cluster ((BT14, (A, B)), BT21).

Respectively, they denote, [1] the displacements of landslide-monitoring points are separate at the lowest hierarchy; [2] at the lower hierarchy, the displacements of landslide-monitoring points (BT13, BT23, BT24, BT32, BT34) have the same trend of "the displacements are small", and the same with (BT11, BT12, BT22, BT31, BT33) of "the displacements are big", while BT14, BT21 show the different trend with both of them, and each other i.e. the exceptions, "the displacement of BT14 is smaller", "the displacement of BT21 is extremely big"; [3] when the hierarchy becomes high, the displacements of landslide-monitoring points (BT13, BT23, BT24, BT32, BT34) and (BT11, BT12, BT22, BT31; BT33) have the same trend of "the displacements are small", however, BT14, BT21 are still unable to be grouped into this trend; [4] when the hierarchy gets higher, the displacements of landslide-monitoring point BT14 can be grouped into the same trend of (BT13, BT23, BT24, BT32, BT34) and (BT11, BT12, BT22, BT31, BT33) that is "the displacements are small", however, BT21 is still an outliner ; [5] the displacements of landslide-monitoring points are unified at the highest hierarchy, that is, the landslide is moving. Simultaneously, these clusters represent different kinds of landslide-monitoring points recorded in the database. And they can naturally form the cluster spectrum figures as Figure 9(c) and Figure 9(d). Seen from these figures, the displacements of landslide-monitoring points (BT13, BT23, BT24, BT32, BT34) and (BT13, BT23, BT24, BT32, BT34) firstly compose two new classes, cluster A and cluster B, then the two new classes compose a larger class with cluster BT14, and they finally compose the largest class with cluster BT21.



(a) All points' potential

(b) Points' potential without the exceptions

(c) All points' cluster spectrum

(d) Points' cluster spectrum without the exceptions

Figure 9. Clusters and cluster spectrum of Baota landslide monitoring points

When the Committee of Yangtze River (Zeng, 2001) investigated in the region of Yunyang Baota landslide, they found out that the landslide had moved to Yangtze River. By the landslide-monitoring point BT21, a small size landslide had taken place. Now there are still two pieces of big rift. Especially, the wall rift of the farmer G. Q. Zhang's house is nearly 15 millimeters. These results from the facts match the discovered spatial knowledge very much, which indicates that the techniques of cloud model-based spatial data mining are practical and creditable.

V. CONCLUSIONS

This paper proposed cloud model-based spatial data mining in the aspects of data preprocessing, uncertain reasoning and knowledge discovery. It was argued that the conceptual space represents different concepts in the same characteristic category, while the feature space depicts complicated spatial objects with multi-properties.

The method integrated the fuzziness and randomness in a unified way via the algorithms of forward and backward cloud generators in the contexts of three numerical characteristics, {Ex, En, He}. It took advantage of human natural language, and might search for the qualitative concept described by natural language to generalize a given set of quantitative datum with the same feature category. Moreover, the cloud model could act as an uncertainty transition between a qualitative concept and its quantitative data. With this method, it was easy to build the mapping relationship inseparably and interdependently between qualitative concept and quantitative data, and the final discovered knowledge with hierarchy could match different demands from different level users. The method would further improve the implementation efficiency, and enhance the comprehension of the discovered spatial knowledge.

The experimental results on Baota landslide monitoring show the cloud model-based spatial data mining can reduce the task complexity, improve the implementation efficiency, and enhance the comprehension of the discovered spatial knowledge.

ACKNOWLEDGEMENTS

This study is supported by the funds from Hong Kong RGC (Project No.3_ZB40), the Hong Kong Polytechnic University (Project No.1.34.9709), and National High Technology R&D Program (863) (Project No.2001AA135081).

REFERENCES

[1] Di, K. C., 1999, The Theories and Methods of Spatial Data Mining and Knowledge Discovery. *Ph.D. Thesis* (Wuhan: Wuhan Technical University of Surveying and Mapping).

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

[3] Han, J., Kamber, M., 2001, *Data Mining: Concepts and Techniques* (San Francisco: Academic Press).

[4] Li, D.R., Wang, S.L., Shi, W.Z., and Wang, X. Z., 2001, On spatial data mining and knowledge discovery (SDMKD), *Geomatics and Information Science of Wuhan University*, 26(6):491-499.

[5] Li, D.R., Wang, S.L., Li, D.Y., and Wang, X. Z., 2002, Theories and technologies of spatial data mining and knowledge discovery. *Geomatics and Information Science of Wuhan University*, 27(3): 221-233.

[6] Li, D.Y., Meng, H. J. and Shi, X. M., 1995, Membership clouds and their generators. *Journal of Computer Research and Development*, 42(8):32-41.

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

[8] Miller, H. J., Han, J., 2001, *Geographic Data Mining and Knowledge Discovery* (London and New York: Taylor and Francis).

[9] PiatetskY-Shapiro G., 1994, An overview of knowledge discovery in databases: recent progress and challenges. In *Rough Sets, Fuzzy Sets and Knowledge Discovery*, edited by Wojciech P. Ziarko (Berlin: Springer-Verlag),1-10

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

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