International Journal of Geographical Information Science
Vol. 24, No. 4, April 2010, 559–576

Using rough set theory to identify villages affected by birth defects: the example of Heshun, Shanxi, China

Hexiang Bai, Yong Ge*, Jin-Feng Wang and Yi Lan Liao

State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

(*Received 6 September 2008; final version received 8 April 2009*)

This article uses rough set theory to explore spatial decision rules in neural-tube birth defects and searches for novel spatial factors related to the disease. The whole rule induction process includes data transformation, searching for attribute reducts, rule generation, prediction or classification, and accuracy assessment. We use Heshun as an example, where neural-tube birth defects are prevalent, to validate the approach. About 50% of the villages in Heshun are used as the sample data, from which all of the rules are extracted. Meanwhile, the other villages are used as reference data. The rules extracted from the training data are then applied to the reference data. The result shows that the rules' generalization is reasonably good. Moreover, a novel relationship between the spatial attributes and the neural-tube birth defects was discovered. That is, the villages that lie in Watershed 9 of this district and that are also associated with a gradient of between 16° and 25° are vulnerable to neural-tube birth defects. This result paves the road for predicting where high rates of neural-tube birth defects will occur and can be used as a preliminary step in finding a direct cause for the disease.

Keywords: rough set; neural-tube birth defects; spatial analysis

1. Introduction

Epidemiologists have traditionally used maps when analyzing associations between location, environment, and disease (Gesler 1986). GIS is particularly well suited for studying these associations because of its spatial analysis and display capabilities (Clarke *et al.* 1996). Recently, many researchers have concentrated on spatial analysis in epidemiology. In epidemiology, Graham *et al.* (2004) introduced a range of techniques relevant to epidemiology that are used in remote sensing, GIS and spatial analysis, and gave some suggestions for possible future directions for the application of remote sensing, GIS, and spatial analysis; Ostfeld *et al.* (2005) briefly described approaches to spatial epidemiology that are spatially implicit; Meng *et al.* (2005) measured the spatial contagion of severe acute respiratory syndrome (SARS) in Beijing and tested the different epidemic factors of the spread of SARS over different periods using spatial statistics; Wang *et al.* (2006) used geographical techniques to identify and map spatial patterns of risk exposures, and mathematical modeling techniques to quantify the temporal spread of SARS in Beijing in the spring of 2003; and Ulugtekin *et al.* (2007) used GIS for tracking the distribution of measles in a district of Istanbul and presented the available questionnaire data by means of maps in which the

*Corresponding author. Email: gey@lreis.ac.cn

ISSN 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/13658810902960079
http://www.informaworld.com

relationship and the distribution of individual cases are shown both spatially and over time. Meanwhile in neural-tube birth defects (NTD), Wu *et al.* (2004) performed exploratory spatial data analysis to identify the risk factors for birth defects and found that there were two typical hot spot clustering patterns that suggest that the risk for neural-tube birth defects exists at two spatial scales; Li *et al.* (2006b) used a Bayesian method to investigate the relationship between the geological background and the occurrence ratio of neural-tube birth defects; Gu *et al.* (2007) proposed an approach combining population and hospital-based methodologies to establish baseline rates for NTD and discuss the risk factors associated with sociodemographic characteristics, maternal characteristics, as well as geographical factors; Liao *et al.*(2008) propose an approach for estimating population based on integration of Genetic Programming (GP) and Genetic Algorithms (GA) techniques with GIS; and Wang *et al.* (2008) use geographical detector-based health risk-assessment methodology to study the NTD of the Heshun Region, China.

Rough set theory, proposed by Pawlak (1982), is an extension of standard crisp set theory. In this extension, an uncertain set can be expressed by the lower and upper approximations, which are defined based on two extreme cases (full inclusion or nonempty overlap) regarding the relationships between an equivalence class and a target set (Yao 2008). The main goal of the rough set-based analysis is to synthesize an approximation of concepts from the acquired data (Komorowski *et al.* 1999). It has been successfully applied to pattern recognition, machine learning, and automated knowledge acquisition (see, e.g., Yasdi 1996, Polkowski and Skowron 1998, Polkowski *et al.* 2000, Leung and Li 2003, Leung *et al.* 2006). In the field of spatial data analysis, some researchers have argued for the advantages of using rough sets and have also provided in recent works some methodologies for handling spatial data. Worboys (1998a, 1998b) discusses an approach for handling and the reason from such representations using techniques that bear resemblance in places to rough set theory. This approach was further developed by Stell and Worboys (1998) and Bittner and Stell (2002). Wang *et al.* (2002) simplify and standardize rough set symbols in terms of rough interpretation and specialized indication, and they proposed rough spatial entities and their topological relationships in rough space. Thus a universal intersected equation is developed, and a rough membership function is further extended with the gray scale in their case study (Wang *et al.* 2002). For data mining in spatial databases, Aldridge (1998) has developed a rough set methodology for obtaining knowledge from multi-theme geographic data and applied the classical rough set method to estimate landslide hazards in New Zealand. Wang *et al.* (2001) have employed the rough set method to discover land control knowledge, with a case study indicating its feasibility. Ahlqvist *et al.* (2000, 2003) and Ahlqvist (2005) have also applied the rough set method for spatial classification and uncertainty analysis. The rough set theory is also used in preprocessing and the classification of remotely sensed imagery and attributes analysis in GIS (Li *et al.* 2006a). These studies, however, have not explicitly studied the mining of rules for the classification of spatial data (Leung *et al.* 2007). So Leung *et al.* (2007) proposes a novel rough set approach for discovering classification rules in general and in remotely sensed data in particular.

Rough set theory has also been used successfully in medicine. Broadly speaking, previous applications of rough sets in medicine can be classified into two categories; diagnosis and outcome prediction, and attribute selection, with an overwhelming majority of articles falling into the former category (Øhrn 1999). Recently, Wilk *et al.* (2005) used rough set theory to select the most relevant clinical features and generate decision rules based on attributes from a medical data set with missing values. Thangavel and Pethalakshmi (2006) proposed the Improved Quickreduct Algorithm to select features from the information system. Su *et al.* (2006) used rough set theory to select the relevant features from the data to predict diabetes. Tsumoto (2007) focuses on several models of medical reasoning that show that the core ideas of rough set theory can be observed in diagnostic models. However, none of these studies explicitly studied the mining rules between spatial factors and epidemiology.

In this article, we focus on extracting decision rules in epidemiology using rough set theory. Birth defects are the major cause of infant mortality and a leading cause of disability. Preventing birth defects requires a comprehensive a priori and accurate understanding of the risk factors that correlate to birth defects. As an example, we analyze the NTD that are prevalent in Heshun, Shanxi, China using rough set theory. Half of the villages in the district are used as sample data, and the data from the other villages are used as reference data. The decision rules found in the sample data are used to classify the reference data. The accurate assessment shows a high agreement between the actual and the identified state if NTD occured in the villages. We also found novel relationships between the spatial attributes and the incidence of NTD.

2. Baby-birth defects

Birth defects, formally defined by the March of Dimes Birth Defects Foundation, refer to any anomaly, functional or structural, that presents in infancy or later in life and is caused by events preceding birth, whether inherited, or acquired. These range from minor cosmetic irregularities to life-threatening disorders. Birth defects are the major cause of infant mortality and a leading cause of disability (Carmona 2005). Birth defects can cause lifelong problems with health, growth, and learning. Most birth defects are thought to be caused by a complex mix of factors. These factors include our genes, our behavior, and environmental influences. Some birth defects can be attributed to a specific cause, but most cannot. We also have a poor understanding of the extent to which factors work together to cause birth defects (CDC 2008a).

Neural-tube defects are major birth defects of a baby's brain or spine. They happen when the neural tube (which later turns into the brain and spine) fails to form correctly, and the baby's brain or spine is damaged as a result. This happens within the first few weeks of pregnancy, often before a woman even knows she is pregnant. The two most common NTD are spina bifida and anencephaly. These birth defects can result in lifelong disabilities or death. Many of these defects could be prevented if all women were to receive adequate levels of B vitamin folic acid on a daily basis, starting from before becoming pregnant.

Spina bifida occurs when the spine and backbones do not close completely. When this happens, the spinal cord and backbones do not form as they should. A sac of fluid penetrates through an opening in the baby's back. Much of the time, part of the spinal cord is in this sac, and is damaged. Most children born with spina bifida live full lives, but they often have lifelong disabilities and need many surgeries. With the right care, most of these children will grow up to lead full and productive lives.

Anencephaly occurs when the brain and skull bones do not form correctly. When this happens, part or all of the brain and skull bones are missing. Babies with this defect die before birth (miscarriage or stillbirth) or shortly after birth. The average cost of caring for a child born with spina bifida for life is about $636,000.00 per child. This is only an average cost, and for many families the total cost might be well above $1,000,000.00. More importantly, the physical and emotional tolls upon the families affected are high as well (CDC 2008b).

To find ways to prevent birth defects, we need to know their cause. Research gives us important clues about the factors that might raise or lower the risk of having a baby with a birth defect. Those clues help us develop sound public health policies for prevention. According to the results of research into birth defects, the probability of a birth defect caused by genetic factors is likely to be similar between regions. However, environmental risk factors, such as chemicals, toxins, and environmental pollution account for different rates of birth defects between regions. These environmental risk factors, including socioeconomic status and geographical elements, often have spatial associations as well as displaying various patterns. This article attempts to uncover the relationships between spatial attributes and NTD.

3. Using rough set theory to extract spatial decision rules of birth defects

Rough set theory was first introduced by Pawlak in 1982. It can extract decision rules from a decision system. A decision system is a special case of an information system. An information system, formally, is a pair $S = (U, A)$, where U is a nonempty finite set of objects called the universe and A is a nonempty finite set of attributes such that $a : U \rightarrow V_a$ for every $a \in A$. The set Va is called the value set of a. A decision system is any information system of the form $D = (U, A \cup \{d\})$, where $d \notin A$ is the decision attribute. The elements of A are called conditional attributes or simply conditions. The decision attributes may take several values though binary outcomes are rather frequent (Komorowski *et al.* 1999). In epidemiology, the universe U may be all the villages of a district, and the decision attribute d can be the state of the village. For example, one can associate the number 1 with the existence of patients in a village, with 0 representing no patients in the village. The conditional attribute A can be the factors associated with the rate of prevalence of a disease, such as the watershed, land use, and soil type.

3.1. Rough set approximation

In a decision system, two objects in a universe may be indiscernible. For example, we may have only two attributes associated with each of two patients, their age, and whether either has a headache. If they are of the same age and neither has a headache, then the two patients are indiscernible according to these two attributes and they will be treated as the same in the rough set analysis. Without loss of generality, the indiscernibility relation is based on an equivalence relationship in classical rough set theory. An equivalence relation is a binary relation R that is reflexive (xRx), symmetric ($xRy \Rightarrow yRx$), and transitive ($xRy \wedge yRz \Rightarrow xRz$). The equivalence class of an element $x \in S$ consists of all objects $y \in S$ such that xRy. For the information system $S = (U, A)$, with any $B \subseteq A$, there is associated an equivalence relation $\mathrm{IND}_S(B) = \{(x, x') \subseteq U^2 | \forall a \in B \; a(x) = a(x')\}$, called the B-indiscernibility relation, which can also be denoted as $\mathrm{IND}(B)$ for simplicity. If $(x, x') \in \mathrm{IND}_S(B)$, then objects x and x' are indiscernible from each other by attributes from B. The equivalence class of the B-indiscernibility relations is denoted by $[x]_B$ (Komorowski *et al.* 1999).

The main goal of the rough set analysis is to synthesize an approximation of concepts from the acquired data. Some concepts cannot be fully defined in a crisp manner by attributes at hand, e.g., some diseases cannot be fully diagnosed by the symptoms the patient has and the land cover type cannot be completely classified only by spectral attributes. Rough sets use upper and lower approximations to describe these uncertain concepts roughly. Formally, for an information system $S = (U, A)$, let $B \subseteq A$ and $X \subseteq U$. We can approximate X using only the information contained in B by constructing the B-lower and B-upper approximations of X, denote $\underline{B}X$ and $\overline{B}X$, respectively, where $\underline{B}X = \{x|[x]_B \subseteq X\}$ and $\overline{B}X = \{x|[x]_B \cap X \neq \phi\}$. The set $BN_B(X) = \overline{B}X - \underline{B}X$ is called the B-boundary region of X. A set is said to be rough if the boundary region is nonempty, whereas it is crisp if the boundary region is empty (Komorowski *et al.* 1999).

3.2. Reduction

The number of attributes at disposal is usually very large in real-life applications. This number can easily become of the order of a few dozens or even hundreds. So it is necessary to reduce the number of attributes to a sufficient minimum (Theodoridis and Koutroumbas 2003). In rough set theory, this step is completed by finding reducts. The basic idea of reducts is to retain only those attributes that preserve the indiscernibility relation and, consequently, the set approximation. The rejected attributes are redundant because their removal cannot worsen the classification. There are usually several such subsets of attributes, and those that are minimal are called reducts. Given an information system $S = (U, A)$, a reduct of A is a minimal set of attributes $B \subseteq A$ such that $\mathrm{IND}(A) = \mathrm{IND}(B)$. In other words, a reduct is a minimal set of attributes from A that preserves the partitioning of the universe and hence the ability to perform classifications as does the whole attribute set A (Komorowski *et al.* 1999).

Let S be an information system with n objects. The discernibility matrix of A is a symmetric $n \times n$ matrix with entries c_{ij} as given below.

$$c_{ij} = \{a \in A|a(x_i) \neq a(x_j)\} \quad \text{for} \quad i,j = 1, \ldots, n$$

Each entry thus consists of the set of attributes upon which objects x_i and x_j differ. A discernibility function f_S for an information system S is a Boolean function of m Boolean variables $a_1^*, \ldots, a_m^*$ (corresponding to the attributes $a_1, \ldots, a_m$) defined as follows:

$$f_S(a_1^*, \ldots, a_m^*) = \wedge\{\vee c_{ij}^* | 1 \leq j \leq i \leq n, c_{ij} \neq \phi\}$$

where $c_{ij}^* = \{a^*|a \in c_{ij}\}$. The set of all prime implicants of f_S determines the set of all reducts of S. Finding a minimal reduct (i.e., a reduct with a minimal cardinality of attributes among all reducts) is NP-hard (Komorowski *et al.* 1999). The high complexity of this problem has motivated investigators to apply various approximation techniques to find near-optimal solutions (see Skowron and Rauszer 1992, Wroblewski 1995, Komorowski *et al.* 1999, Vinterbo and Øhrn 2000, Wang and Wang. 2001, Jensen and Shen 2005). In our experiment, we select a genetic algorithm for finding ε-approximate (Vinterbo and Øhrn 2000) reducts.

3.3. Decision rule and classifier

Once the reducts have been found, the rules are easily constructed by overlaying the reducts over the originating decision table and reading off the values. To facilitate an understanding of the decision rule synthesis process, some related concepts are given here. Let $D = (U, A \cup \{d\})$ be a decision system and let $V = \cup\{V_a|a \in A\} \cup V_d$. Atomic formulae over $B \subseteq A \cup \{d\}$ and V are expressions of the form $a = v$; they are called descriptors over B and V, where $a \in B$ and $v \in V_a$. The set *F(B,V)* of formulae over B and V is the least set containing all atomic formulae over B and V and is closed with respect to the propositional connectives $\wedge$ (conjunction), $\vee$ (disjunction), and $\neg$ (negation). Given $\varphi \in F(B, V)$ and φ

is of the form $a = v'$, a decision rule for D is any expression of the form $\varphi \Rightarrow d = v$, where $v \in V_d$ and $\|\varphi_D\| = \{x \in U|a(x) = v^{'}\}$ (which denotes the meaning of φ in the decision table D) is a nonempty set. Formulae φ and $d = v$ are referred to as the predecessor and the successor of the decision rule $\varphi \Rightarrow d = v$ (Komorowski *et al.* 1999). The decision rules can also be denoted as *if* φ *then d* = *v*. All the rules constitute a rule set.

When a set of rules has been induced from a decision table containing a set of training examples, it can be inspected to see whether they reveal any novel relationships between attributes that are worth pursuing for further research. Furthermore, the rules can be applied to a set of unseen cases in order to estimate their classificatory power (Komorowski *et al.* 1999). In the experiment, the standard voting (Øhrn 1999) is chosen as the classifier.

3.4. Discretization

In the previous sections, it was assumed that the attributes' values are discretized data. If the attribute values are continuous, they should be discretized beforehand. Meanwhile, the discretization step determines how coarsely we want to view the world (Komorowski *et al.* 1999). For instance, body temperature, which is usually measured in real numbers, can be divided into three ranges in medicine. If your temperature is higher than 37.5°C, then you may have a fever. If your temperature falls between 36°C and 37.5°C, then your temperature is in a normal state. And if your temperature is less than 36°C, your temperature is too low. One can easily see that the selection of appropriate intervals and the partitioning of attribute value sets is a complex problem and its complexity can grow exponentially with the number of attributes to be discretized. Discretization is a step that is not specific to the rough set approach but that most rule or tree induction algorithms currently require for them to perform well (Komorowski *et al.* 1999).

In earlier days, simple techniques were used such as equal-width (equal-interval) and equal-frequency (equal-quantile, or, a form of binning) to discretize. As the need for an accurate and efficient classification grew, the technology for discretization developed rapidly. Over the years, many discretization algorithms have been proposed and tested to show that discretization has the potential to reduce the amount of data while retaining or even improving predictive accuracy. Discretization methods have been developed along different lines because of different needs: supervised versus unsupervised, dynamic versus static, global versus local, splitting (top-down) versus merging (bottom-up), and direct versus incremental (Liu *et al.* 2002). According to the experimental results obtained by Liu *et al.* (2002), it is suggested that if we simply want to discretize data, other things being equal, entropy (MDLP – the minimum description length principle) should be the first consideration.

3.5. Modeling process

The whole rough set analysis progress can be divided into four steps (Figure 1): (1) construct a decision information system based on the data at hand; (2) find minimal reducts of conditional attributes; (3) generate rules according to the reducts; and (4) apply the rules to unseen cases and perform error analysis. In the first step, the original data should be transformed to a decision system. This step is extremely important when rough set theory is used in spatial data analysis. Most spatial data are presented in a map, so the objects and their attributes in spatial data should be extracted from maps and presented in table format. Furthermore, some spatial attributes, which cannot be obtained directly, must be calculated according to maps. Many spatial attributes are continuous variables. Under such

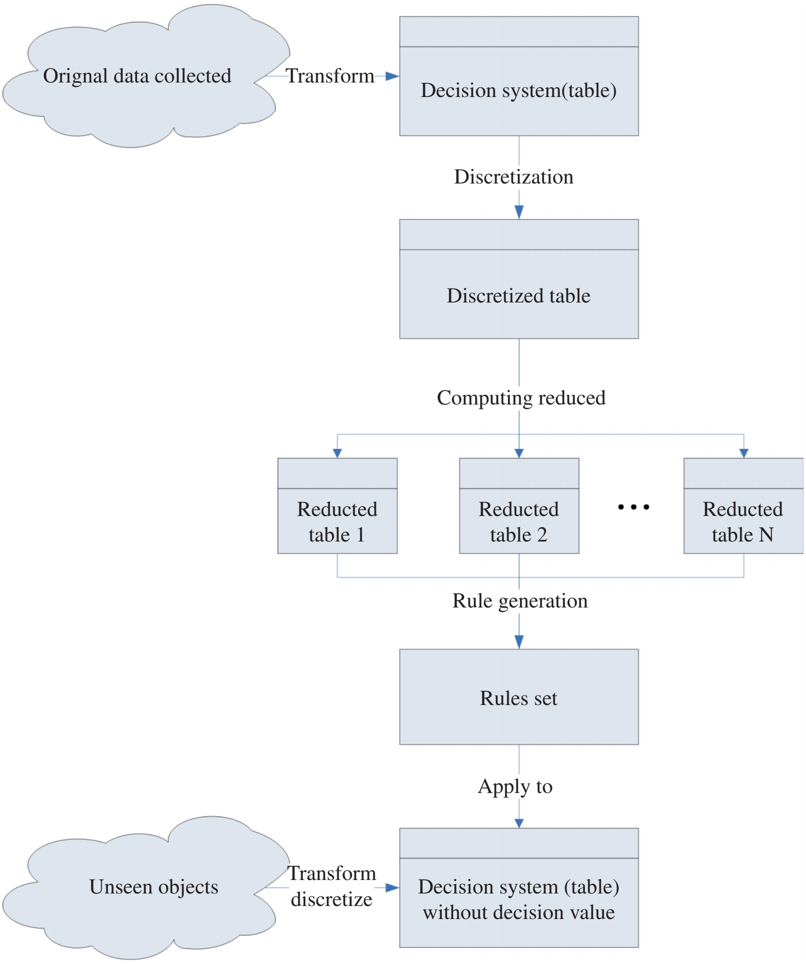

Figure 1. Rough set-modeling process for extracting spatial decision rules of birth defects.

circumstances, the continuous attributes should be discretized before performing rough set analysis. In our experiment, we chose the MLDP algorithm to discretize the continuous attributes. Next, as is described in Section 3.2, finding reducts is used as the attribute selection method. Not all attributes transformed or collected previously are the key factors that are related to the decision attribute. So it is important to select attributes that are in close relation to the decision. In rough set analysis, the number of conditional attributes is reduced without impacting the classificatory ability of the decision system, i.e., the minimal set from all attributes at hand, which preserves the partitioning of the universe as the whole attribute set does of attributes, is selected. Then, the rules are generated from the two reducts by overlaying the reducts over the originating decision table and reading off the values. These rules are the knowledge induced from the decision system and can be applied to unseen objects to validate its classificatory ability. Therefore, the same attributes of the unseen spatial objects should be collected according to the original decision system that generates the spatial decision rules, and also be transformed and discretized in the same way with original data. Finally, an accuracy assessment of the classification result should be performed to validate the decision rules' accuracy. In this article, the training and validation areas do not overlap spatially.

4. An empirical study

Based on our research, the ratio of birth defect occurrences is estimated to be about 40–50% in P. R. China. Shanxi province, in the northern region of China, has the highest ratio of NTD in the world. To reach prepotency, we selected Heshun, one county of Shanxi, as an experimental region for study (Figure 2). This county lies in the Taihang Mountain region and forms a relatively closed area. Most of the people in this county are farmers and seldom change their living environment. Furthermore, there have been no large-scale movements of people in the history of this region. The inherited and congenital causes of birth defects are

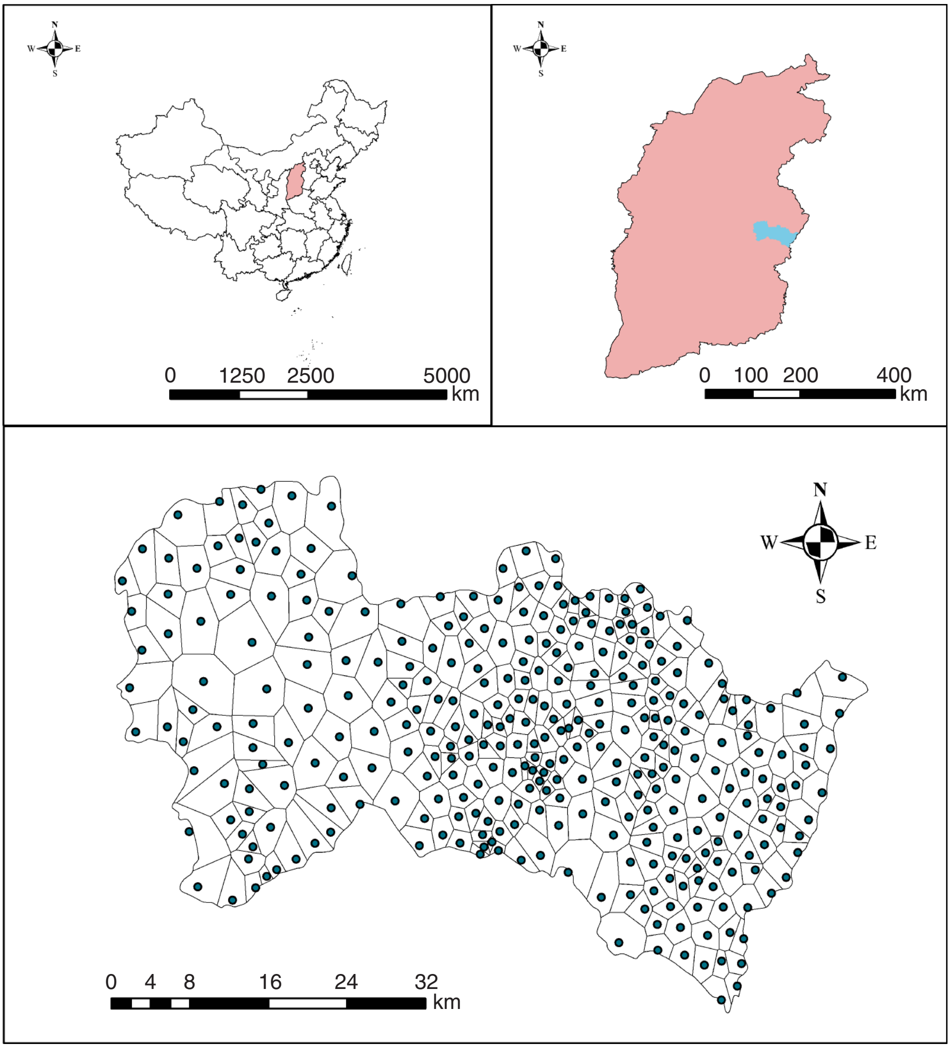

Figure 2. The position of the study area and villages in the study region with their Voronoi polygons.

similar among the people in this region, and those causes explain only a small fraction of all the NTD cases. In our experiment, rough set theory is used to study the relationships between spatial factors, which mostly concern the spatial relationships and spatial attributes, and NTD. Furthermore, all the villages in the county are divided into two parts. One is used as the sample area and the other is used as the test area. The decision rules mined from the sample area are used to identify the decision attribute of the test area, and an error matrix is build to show its power of identification.

4.1. Data description and transformation

There were 322 villages and one town in the study area. The locations of the 322 villages were determined by the Geographical Information System for spatial analysis. As there were no boundaries defined for the villages, we drew them for each village using a Voronoi polygon (Figure 2). We collected information on both spatial and nonspatial attributes of all the villages in the test area. For example, the nonspatial data include GDP, number of children born, number of children who have had NTD, fertilizer use in the area (fertilizer), access to a doctor (Doctor), production of fruit (Fruit), and production of vegetables (Vegetables). The spatial data include elevation (Figure 3), soil type, rivers, roads, lithology type, land cover type (Figure 4), faulting attributes, etc. Some of the spatial attributes, such as soil type, lithology type, and land cover type, can be used directly in Decision Systems. However, other spatial attributes, such as rivers, roads, and faulting attributes need to be transformed in order to analyze the information they carry.

The first step in the rough set analysis model is to transform the data at hand to a decision system $D = (U, A \cup \{d\})$. We want to inspect which factors closely correlate to the NTD, so the decision attribute d means whether the village has NTD instances. If the village has not had at least one instance from 1998 to 2003, then its decision attribute is assigned 1, which means the village had NTD instances during this period. Otherwise, the decision attribute is

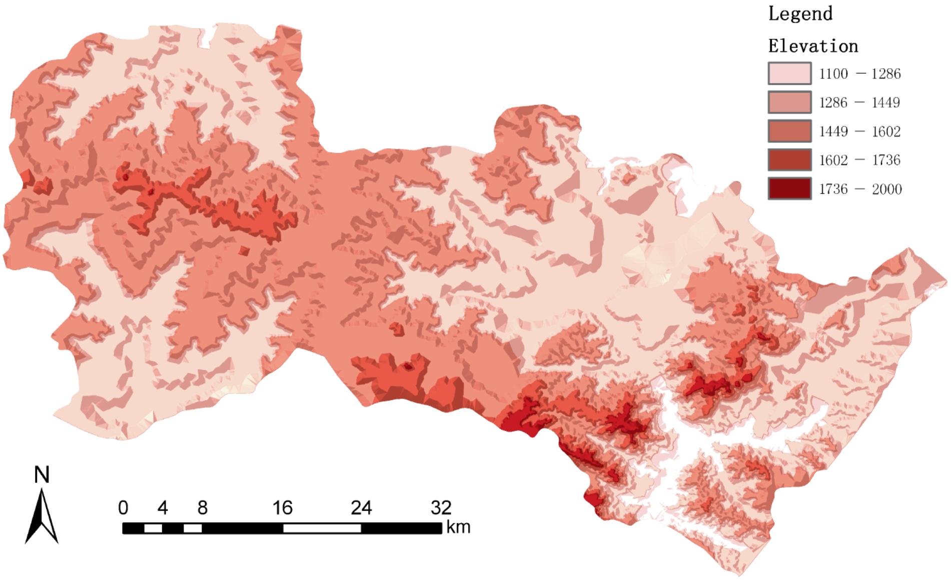

Figure 3. Elevation of the study area.

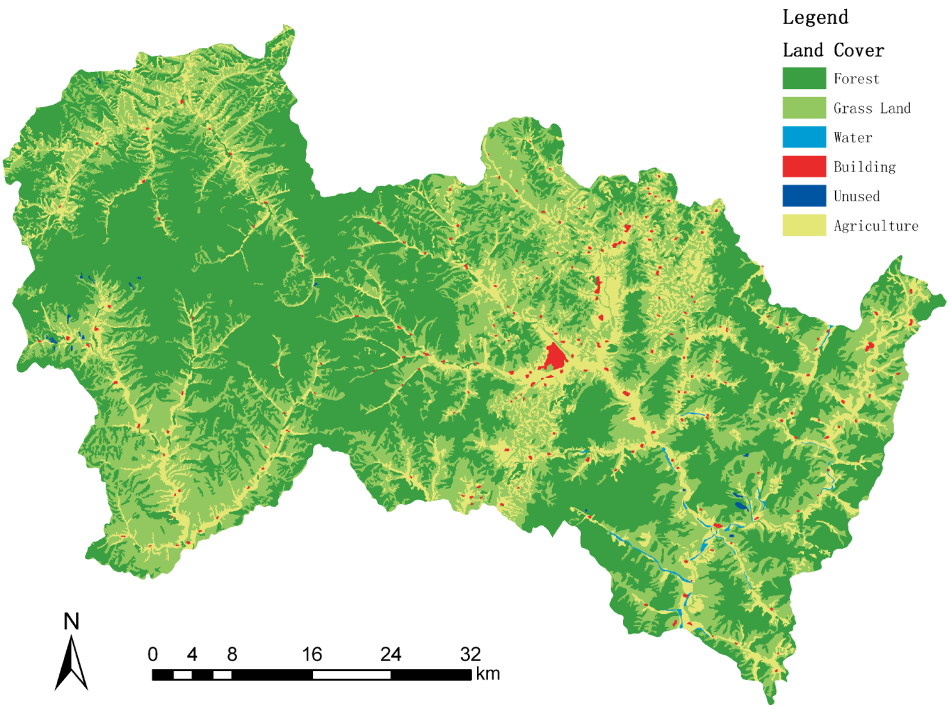

Figure 4. Landcover of the study area.

assigned 0. Because the main object of this study is to discover the relationships between spatial factors and NTD, the villages that did not have new babies are not included. So there are 98 villages that have NTD instances and 217 villages that have had no NTD instances.

The nonspatial conditional attributes used in our experiment include GDP, Doctor, Fruit, Fertilizer, and Vegetable. According to the GDP, which has been converted into 1970 US dollars, of the study area, the villages are divided into three groups: villages that have not entered the first stage of industrialization (GDP $< \$280$), villages that are in the first stage of industrialization ($\$280 \leq \text{GDP} < \$560$), and villages that are in the second stage of industrialization ($\$560 \leq \text{GDP} < \$1120$) (Hollis *et al.* 1986) denoted as A, B, and C, respectively. For each village, we also know the number of doctors, the quantity of fruit and vegetable produced per year, and the amount of fertilizer used on the farms each year. These attributes are discretized using the MLDP discretization method.

Some spatial conditional attributes are not in need of discretizations. These attributes include soil type, lithology type, land cover type, gradient, and watershed. The soil type includes leaching cinnamon soil, cinnamon soil, calcareous cinnamon soil, infant cinnamon soil, neural lithosols, neural skeletisols and calcareous skeletisols, and fluvo-aquic soil. The lithology type includes Carbon (C), Changchen (CH), Ordo (O), Permian (P), Quaternary (Q), Triassic (T), and Wutai (Ar). The land cover type includes grassland, forests, water, building, agriculture, and others. The gradients in this area are divided into five categories: $0°–8°$, $9°–16°$, $17°–24°$, $25°–32°$, and $33°–90°$. This district has nine watersheds, denoted as digits 1–9.

Other spatial conditional attributes are continuous attributes that need to be discretized before they are used in rough set synthesis. These attributes are the road buffer, river buffer, fault buffer, elevation, and neighbor. Buffer regions of road, river, and fault zone were built into a GIS environment, with buffer distances of 2, 4, 6, 8, 10, 12, 14, or 16 km. As a result, the study region was divided into several zones with buffer distance ranges of 0–2, 2–4, 4–6, 6–8, 8–10, 10–12, 12–14, and 14–16 km. The attribute neighbor is the number of adjacent villages that have NTD instances. These attributes are also discretized using the MLDP discretization method. Furthermore, for convenience, all the spatial and nonspatial attributes concerned are summarized in Table 1. In the table, the first column is the name of the attribute, the second column is the meaning of the attribute, and the last column shows how the attribute is discretized.

4.2. Rule extraction and identification of unseen objects

First, the genetic algorithm discussed in Section 3.2 is applied to find the minimal reducts. We obtained two reducts from the decision system built in Section 4.1, each of which contains three spatial attributes. This means that at least three spatial attributes {watershed, gradient, neighbor} or {gradient, landcover, neighbor}, should be used to synthesize the decision rules. So the reducts reduce the number of attributes by 20%. The two reducts share two common spatial attributes: neighbor and gradient. This attribute is important because the NTD disease is distributed in clusters spatially. This can be seen from the Moran's Index computed in ArcGIS. Its value is 0.06 and *Z* score is 6.68. There is a less than 1% likelihood that this clustered pattern could be the result of random chance. It has significant effects on the classification result. It produces a sound result for attribute selection highlighting the discriminatory power in different combinations of spatial and nonspatial attributes.

Then, according to the two reducts, we can generate 50 rules using methods introduced in Section 3.3. These rules, which discovered the relationship between attributes and decision, are the decision rules. Furthermore, we can use these rules to identify unseen objects and provide a knowledge base to find the real cause of the disease. Meanwhile, it can

Table 1. Summary of spatial and nonspatial attributes.

Attribute name	Meaning	Discretization method
GDP	Average GDP during the 6 years	Manually discretized
Doctor	The number of doctors of the village	MDLP
Fruit	The quantity of fruit produced per year of each village	MDLP
Fertilizer	The quantity of fertilizer used in the farm per year of each village	MDLP
Vegetable	The quantity of vegetable produced per year of each village	MDLP
Soil type	The main soil type of the village	Not discretized
Lithology type	The main lithology type of the village	Not discretized
Land cover type	The main land cover type of the village	Not discretized
Gradient	The gradient calculated from DEM for each village	Not discretized
Watershed	In which watershed the village lies	Not discretized
Road buffer	Distance to the main road from the village	MDLP
River buffer	Distance to the main river from the village	MDLP
Faultage buffer	Distance to the faultage from the village	MDLP
Elevation	The altitude of the village	MDLP
Neighbor	How many neighbor villages have NTD instance from 1998 to 2003	MDLP
Decision attribute	Whether the village has NTD instances from 1998 to 2003	Not discretized

Table 2. Error matrix.

	Actual village status			
	Have no NTD	Have NTD	Undefined	Row total
Identified village status				
Have no NTD	106	0	0	106
Have NTD	0	47	0	47
Undefined	1	3	0	4
Column Total	107	50	0	157

Producer's accuracy	User's accuracy
Have NTD = 100%	Have NTD = 99.1%
Have no NTD = 100%	Have no NTD = 94.0%
Overall Accuracy = 97.5%	

supervise the decision maker to make strategies to defend or even control the disease through changing local environments or improving living conditions. Some disease, which may not have been recognized or conquered by the human race, can be controlled through changing some of the factors related to the disease, for example, cholera that was ever prevalent in London in the mid-nineteenth century.

The classification rules are derived from the villages from which the sample data set is generated. To test the effectiveness of the rough set method, the remaining villages are used as reference data for accuracy verification. The composition of the error matrix helps generate standard accuracy indices, including the producer's accuracy and user's accuracy for individual classes, as well as an overall accuracy and Kappa coefficient of agreement for the entire data set. The error matrix is summarized in Table 2.

The actual and identified values associated with whether a village has NTD instances are shown in Figure 5, respectively. There are only four villages in which the identified decision values and the actual decision values are different, and their identified values are undefined.

5. Discussion

From the error matrix summarized in Table 2, the overall accuracy of the identification is 97.5%. Meanwhile, the user's accuracies for determining villages that have NTDs and villages that have no NTDs are 99.1 and 94.0%, respectively, and both producers' accuracies are 100%. This shows the high agreement between the actual and the identified decision data. It has demonstrated that the generalization of the rules is reasonably good or, alternatively, that the situation of over-generalization will not occur. Today, the causes of birth defects in 65–80% of cases are unknown. Without knowing the causes of birth defects, we are helpless in terms of being able to prevent them. Our lack of knowledge about how to prevent birth defects is surprising in light of the advances we have made in combating diseases and death in infants over the past 50 years. Rough set-based rule synthesis can also be used to identify the outcome in villages yet unseen, based on knowledge found previously. In our experiment, there are only four villages for which the identifications were wrong. The four villages, which are called Lingnan, Yinmachi, Yangjuan, and Chanyaogou, have attribute values that do not appear in the sample data, so they cannot be identified correctly.

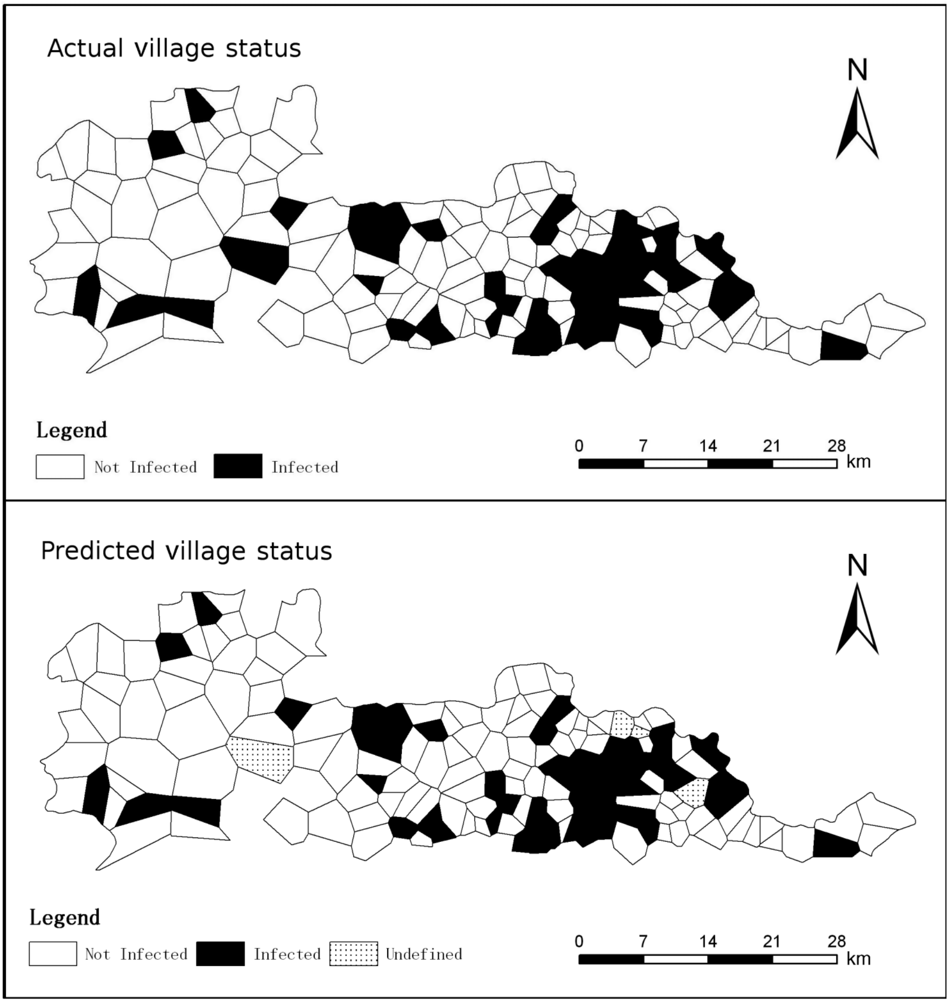

Figure 5. The actual decision value and the identified decision value.

Although the four spatial attributes in the two reducts {watershed, gradient, landcover, neighbor} are probably not the direct cause of the NTD, they can be used to identify the NTD instances. There probably exist underlying relationships between the direct cause of the disease and the four indices. Many articles have reported extensively on the possible causes of NTD. These include social, behavioral, and environmental factors, such as smoking, prescription medication, and alcohol use by pregnant women, as well as conditions that cause injury that are outside personal control, such as motor vehicle accidents and the presence of toxic agents in the environment such as radiation, chemicals and metals (Weicker, L. 1999). In Meng's doctoral dissertation (2005), it was shown that NTD have statistically significant associations with geographical factors such as elevation, soil type, lithology, and watershed. And Li X.H. (2007) proved that areas, which have gradients between 8° and 24°, are vulnerable to NTD. In our experiment, it can be seen that watershed, gradient, land cover type, and neighbor can be used to identify the occurrence of the NTD. We select six rules synthesized by the rough set here, in which the predecessor covers

Table 3. Decision rules.

ID	Rule content
1	Watershed = 8 AND Gradient between 0° and 8° AND Neighbor < 2 ⇒ decision = 0
2	Watershed = 5 AND Gradient between 0° and 8° AND Neighbor < 2 ⇒ decision = 0
3	Gradient between 0° and 8° AND landcover (33) AND Neighbor < 2 ⇒ decision = 0
4	Gradient between 0° and 8° AND landcover (32) AND Neighbor < 2 ⇒ decision = 0
5	Watershed = 8 AND Gradient between 0° and 8° AND Neighbor>2 ⇒ decision = 1
6	Watershed = 5 AND Gradient between 0° and 8° AND Neighbor>2 ⇒ decision = 1

at least 10% of the sample, and analyze its effectiveness. The decision rules are summarized in Table 3, and we also highlight on a map the villages that these six rules identify in the sample data (Figure 6). In Figure 6, 0 represents villages that cannot be identified by any of the six rules, 1 represents villages that can be identified by the first rule in Table 3, 2 represents villages that can be identified by the second rule in Table 3, . . ., and, 6 represents villages that can be identified by the first rule in Table 3.

From these rules, we can see that the neighbor is the key factor that associates with the occurrence of NTD instances. This is sound, because NTD are always spatially autocorrelated, which can be seen from Moran's I. But according to the spatial attribute neighbor alone cannot identify if a village has NTD instances. This can be seen from the rule below:

$$\text{Watershed} = 9 \text{ AND Gradient between } 17^\circ \text{ and } 24^\circ \text{ AND Neighbor} > 2 \Rightarrow \text{ decision } = 1$$

This rule discovers that if the watershed is 9 and gradient is between 17° and 24° then the village will have NTD instances. Furthermore, there are also two more rules that are associated with watershed 9, shown below:

$$\text{Watershed} = 9 \text{ AND Gradient between } 0^\circ \text{ and } 8^\circ \text{ AND Neighbor} > 2 \Rightarrow \text{ decision } = 1$$

$$\text{Watershed} = 9 \text{ AND Gradient between } 9^\circ \text{ and } 16^\circ \text{ AND Neighbor} > 2 \Rightarrow \text{decision} = 0$$

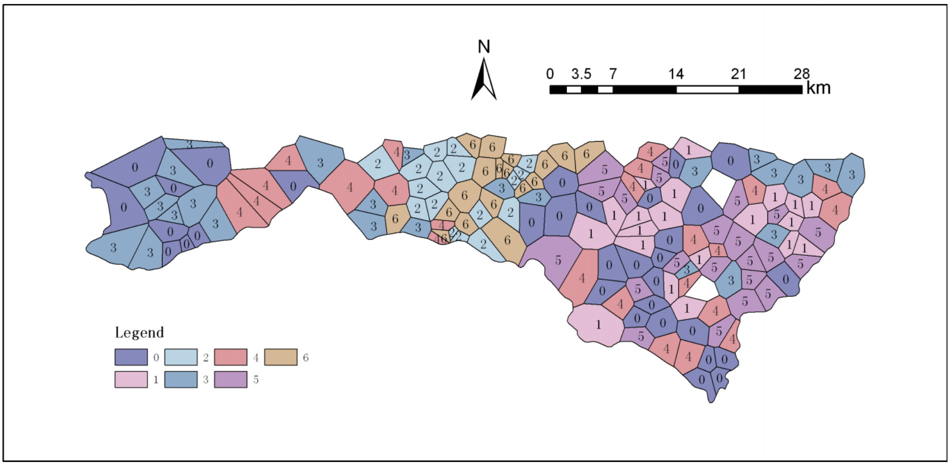

Figure 6. The villages identified by six rules in Table 3 in the sample data.

These also support our judgment that the villages that lie in watershed 9 and have gradients between 17° and 24° are vulnerable to NTD. And to make it clear, we observe villages that have gradients between 17° and 24° but with the other watersheds. There is only one such rule: Watershed = 8 AND Gradient between 17° and 24° AND Neighbor $> 2 \Rightarrow$ decision = 0. Therefore, for an unseen village, if it lies in the same watershed or watershed with a similar geographical condition as watershed 9, and it also has a gradient between 17° and 24°, then there must be some measures taken on the village. The villages in this watershed and those that have this gradient may be subjected to some underlying unknown environmental or social factors that are critical to the occurrence of NTD. For example, if the gradient is steep, then pregnant women may suffer more unexpected injuries than those living on flat areas. And there probably exist some types of microelements in the rock in the area that may directly cause the NTD defect in which the watershed is 9. Meanwhile, currents are rapid on steep gradients. Therefore, microelements are easily assimilated into the water and further absorbed by people. For the discussion above, we find that the decision rules synthesized by rough set analysis are a preliminary step toward identifying the direct cause of NTD.

Rough set theory is completely data-driven, and needs no prior knowledge. However, because all the rules are derived from the sample data directly, the rules cannot cover all possible cases. Hence there are probably undefined cases in the identified result. In this article, there are four ‘undefined’ villages. These four villages cannot be matched to any rules derived from the training set. The first undefined village is Lingnan. The gradient of Lingnan is between 33° and 90° however, because this gradient value does not appear in the training set. So there is no rule for identifying this village. The second undefined village is Yinmachi. Because there is no village in the training set that has gradient between 0° and 8° and, at the same time, has grass landcover. So there is no rule for identifying this village. The third undefined village is Yangjuan. Although in the training sets there are villages that have gradient between 0° and 8°, and also possess its building land cover type, there are no villages that have more than two adjacent villages with NTD instances, therefore it cannot be identified. The final undefined village is Chanyaogou. Although there are villages that have a gradient between 9° and 16° and possess watershed 4, there are no villages that have more than two adjacent villages with NTD instances, thus, it cannot be identified. In our future work, special attention should be paid to the effective use of at-hand sample data, as well as attention to methods for designing sample schemes for rule extraction of rough sets.

6. Conclusion

In this article, we propose a general framework based on rough set theory to extract spatial decision rules for NTD. The spatial decision rules found can be used to identify NTD for unseen villages. From the accuracy assessment of the identification through an error matrix, it can be seen that the spatial rules synthesized by the rough set have reasonably good generalization. On the other hand, as a preliminary step to finding the direct causes of the disease, many useful conclusions can be drawn from the rules generated. For example, the rules found in our experiment reveal that the watershed and gradient are important. Then further in situ inspections and experiments of the area need to be considered in order to identify the real causes that hide behind the simple relationship. And in our future work, it is also important to compare the result from rough set theory with results from some traditional methods, for example, Poisson regression.

Rough set theory is data-driven, and the advantage of this approach is that it does not require the user to make any a priori assumptions about the data. It can be used to find attributes efficiently, a critical step in the decision-making process, through the computing of reducts. Furthermore, the decision rule generated can be used to predict or classify unseen objects. In epidemiology, the original data acquired are always accompanied by missing data or errors. These factors will result in an incomplete decision system or bring inconsistency to the decision-making system. We can use the extended model of rough set theory to handle an incomplete or inconsistent decision-making system. For example, the missing data can be completed or directly analyzed through maximum consistent block technology. The inconsistency in the decision-making system can be handled by the concept of a probability rough set. Sometimes, the attributes are generated from fuzzy concepts. Fuzzy rough set theory can be used in such situations.

Acknowledgments

This work was supported is part by the National 973 Program 2006CB701305, by the National Natural Science Foundation of China under Grant 40671136 and 40471111, and by the Ministry of Science and Technology of the People's Republic of China 2007DFC20180.

References

Ahlqvist, O., 2005. Using uncertain conceptual spaces to translate between land cover categories. *International Journal of Geographical Information Science*, 19 (7), 831–857.

Ahlqvist, O., Keukelaar, J., and Oukbir, K., 2000. Rough classification and accuracy assessment. *International Journal of Geographical Information Science*, 14 (5), 475–496.

Ahlqvist, O., Keukelaar, J., and Oukbir, K., 2003. Rough and fuzzy geographical data integration. *International Journal of Geographical Information Science*, 17 (3), 223–234.

Aldridge, C.H., 1998. *A theory of empirical spatial knowledge supporting rough set based knowledge discovery in geographic databases*. Thesis (PhD). University of Otago, Dunedin, New Zealand.

Bittner, T. and Stell, J.G., 2002. Vagueness and rough location. *Geoinformatica*, 6 (2), 99–121.

Carmona, R.H., 2005. The global challenges of birth defects and disabilities. *Lancet*, 366, 1142–1144.

CDC, 2008a, Birth Defects Research. Available from: http://www.cdc.gov/ncbddd/bd/research.htm [Accessed 05 August 2008].

CDC, 2008b, Questions and Answers. Available from: http://cdc.gov/ncbddd/folicacid/faqs.htm [Accessed 05 August 2008].

Clarke, K.C., Mclafferty, S.L., and Tempalski, B.J., 1996. On epidemiology and geographic information systems: a review and discussion of future directions. *Emerging Infectious Diseases*, 2 (2), 85–92.

Gesler, W., 1986. The uses of spatial analysis in medical geography: a review. *Social science and Medicine*, 23 (10), 963–973.

Gu, X., *et al.*, 2007. High prevalence of NTDs in Shanxi province: a combined epidemiological approach. *Birth Defects Research (Part A): Clinical and Molecular Teratology*, 79 (10), 702–707.

Graham, A.J., Atkinson, P.M., and Danson, F.M., 2004. Spatial analysis for epidemiology. *Acta Tropica*, 91 (3), 219–225.

Hollis, B.C., Sherman, R., and Moises, S., 1986. *Industrialization and growth: a comparative study.* New York: Published for the World Bank [by] Oxford University Press.

Jensen, R. and Shen, Q., 2005. Fuzzy-rough data reduction with ant colony optimization. *Fuzzy Sets and Systems*, 149 (1), 5–20.

Komorowski, J., *et al.*, 1999. Rough sets: A tutorial. *In*: S. Pal and A. Skowron, eds. *Rough fuzzy hybridization: a new trend in decision-making*. Singapore: Springer-Verlag, 3–93.

Leung, Y. and Li, D.Y., 2003. Maximal consistent block technique for rule acquisition in incomplete information systems. *Information Sciences*, 153, 85–106.

Leung, Y., Wu, W.Z., and Zhang, W.X., 2006. Knowledge acquisition in incomplete information systems: a rough set approach. *European Journal of Operational Research*, 168 (1), 164–180.

Leung, Y., *et al.*, 2007. A rough set approach to the discovery of classification rules in spatial data. *International Journal of Geographical Information Science*, 21 (9), 1033–1058.

Li, X.H., 2007. *Environment factors analysis for the human birth defects risk and the environmental mechanism exploration*. Thesis (Unpublished PhD). Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Science, Beijing, China.

Li, D.R., Wang, S.L., and Li, D.Y., 2006a. *Spatial data mining theories and applications*. Beijing: Science Press, 487–500.

Li, X.H., *et al.*, 2006b. A geological analysis for the environmental cause of human birth defects based on GIS. *Toxicological and Environmental Chemistry*, 88 (3), 551–559.

Liao, Y.L., *et al.*, 2010. Integration of GP and GA for mapping population distribution. *International Journal of Geographical Information Science*, 24 (1), 47–67.

Liu, H., *et al.*, 2002. Discretization: an enabling technique. *Data mining and knowledge discovery*, 6 (4), 393–423.

Meng, B., 2005. *Researches on identifying the factors with spatial processes and scaling with geo-data*. Thesis (Unpublished PhD). Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Science, Beijing, China.

Meng, B., *et al.*, 2005. Understanding the spatial diffusion process of severe acute respiratory syndrome in Beijing. *Public Health*, 119 (12), 1080–1087.

Øhrn, A., 1999. *Discernibility and rough sets in medicine: tools and applications*. Thesis (PhD). Norwegian University of Science and Technology, Norway.

Ostfeld, R.S., Glass, G.E., and Keesing, F., 2005. Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends in Ecology and Evolution*, 20 (6), 328–336.

Pawlak, Z., 1982. Rough sets. *International Journal of Information and Computer Sciences*, 11 (5), 341–356.

Polkowski, L. and Skowron, A., eds., 1998. *Rough sets in knowledge discovery 1: Methodology and applications, 2: Applications*. Heidelberg: Physica-Verlag.

Polkowski, L., Tsumoto, S., and Lin, T.Y., 2000. *Rough set methods and applications*. Heidelberg: Physica-Verlag.

Skowron, A. and Rauszer, C., 1992. The discernibility matrices and functions in information systems. *In*: R. Sł owiński, ed. *Intelligent decision support: handbook of applications and advances of the rough sets theory*. Norwell, MA, USA: Kluwer Academic Publishers, 331–362.

Stell, J.G. and Worboys, M.F., 1998. Stratified map spaces: a formal basis for multiresolution spatial databases. *In*: T.K. Poiker and N. Chisman, eds. *SDH'98 Proceedings 8th international symposium on spatial data handling*, 11–15 July 1998 Vancouver, BC, Canada. Vancouver: International Geographical Union, 180–189.

Su, C.T., *et al.*, 2006. Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data. *Computers and Mathematics with Applications*, 51 (6–7), 1075–1092.

Thangavel, K. and Pethalakshmi, A., 2006. Feature selection for medical database using rough system. *International Journal on Artificial Intelligence and Machine Learning*, 6 (1), 11–17.

Theodoridis, S. and Koutroumbas, K., 2003. *Pattern recognition*. 2nd ed. America: Academic Press, 163–207.

Tsumoto, S. 2007. Medical Reasoning and Rough Sets. *In*: J.G. Carbonell and J. Siekmann, eds. *Rough sets and intelligent systems paradigms*. Heidelberg: Springer Berlin, 90–100.

Ulugtekin, N., *et al.*, 2007. Use of GIS in epidemiology: a case study in Istanbul. *Journal of Environmental Science and Health, Part A*, 41 (9), 2013–2026.

Vinterbo, S. and Øhrn, A., 2000. Minimal approximate hitting sets and rule templates. *International Journal of Approximate Reasoning*, 25 (2), 123–143.

Wang, J. and Wang, J., 2001. Reduction algorithms based on discernibility matrix: the ordered attributes method. *Journal of Computer Science and Technology*, 16 (6), 489–504.

Wang, S.L., Wang, X.Z., and Shi, W.Z., 2001. Development of a data mining method for land control. *Geo-Spatial Information Science*, 4 (1), 68–76.

Wang, S.L., *et al.*, 2002. Geo-rough space. *Geo-Spatial Information Science*, 6 (1), 54–61.

Wang, J.F., *et al.*, 2006. Spatial dynamics of an epidemic of severe acute respiratory syndrome in an urban area. *Bulletin of the World Health Organization*, 84 (12), 965–968.

Wang, J.F., *et al.*, 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science*, 24 (1), 107–127.

Weicker, L., 1999. *Healthy from the start*. Available from: http://healthyamericans.org/reports/files/healthystart.pdf [Accessed 05 August 2008].

Wilk, S., *et al.*, 2005, Supporting triage of children with abdominal pain in the emergency room. *European Journal of Operational Research*, 160, 696–709.

Worboys, M.F., 1998a. Computation with imprecise geographical data. *Computers Environment and Urban Systems*, 22 (2), 85–106.

Worboys, M.F., 1998b. Imprecision in finite resolution spatial data. *GeoInformatica*, 2 (3), 257–279.

Wroblewski, J., 1995. Finding minimal reducts using genetic algorithm. *Warsaw University of Technology: ICS Research Report*, 16–95.

Wu, J.L., *et al.*, 2004. Exploratory spatial data analysis for the identification of risk factors to birth defects. *BMC Public Health*, 23 (4), 23–32.

Yao, Y., 2008. Probabilistic rough set approximations. *International Journal of Approximate Reasoning*, 49 (2), 255–271.

Yasdi, R., 1996. Combining rough sets learning and neural learning: method to deal with uncertain and imprecise information. *Neuralcomputing*, 7 (1), 61–84.