J. Geogr. Sci. 2011, 21(1): 101-117
DOI: 10.1007/s11442-011-0832-y

Application of rough set-based analysis to extract spatial relationship indicator rules:
An example of land use in Pearl River Delta

GE Yong[1], CAO Feng[1], DU Yunyan[1], LAKHAN V. Chris[2], *WANG Yingjie[1], LI Deyu[3]

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;
2. Department of Earth and Environmental Sciences, University of Windsor, ON N9B 3P4, Canada;
3. School of Computer & Information Technology of Shanxi University, Taiyuan 030006, China

Abstract: Spatial relations, reflecting the complex association between geographical phenomena and environments, are very important in the solution of geographical issues. Different spatial relations can be expressed by indicators which are useful for the analysis of geographical issues. Urbanization, an important geographical issue, is considered in this paper. The spatial relationship indicators concerning urbanization are expressed with a decision table. Thereafter, the spatial relationship indicator rules are extracted based on the application of rough set theory. The extraction process of spatial relationship indicator rules is illustrated with data from the urban and rural areas of Shenzhen and Hong Kong, located in the Pearl River Delta. Land use vector data of 1995 and 2000 are used. The extracted spatial relationship indicator rules of 1995 are used to identify the urban and rural areas in Zhongshan, Zhuhai and Macao. The identification accuracy is approximately 96.3%. Similar procedures are used to extract the spatial relationship indicator rules of 2000 for the urban and rural areas in Zhongshan, Zhuhai and Macao. An identification accuracy of about 83.6% is obtained.

Keywords: rough set theory; spatial relations; spatial indicator rules; land use change

1 Introduction

Spatial relations contain topological relation, direction relation, sequence relation and distance relation, and are of importance in spatial data query/retrieval (Egenhofer, 1997; Papadias and Theodoridis, 1997), spatial data mining (Clementini *et al.*, 2000), spatial analysis

Received: 2010-03-23 **Accepted:** 2010-08-31
Foundation: National Natural Science Foundation of China, No.40971222; State Key Laboratory of Independent Innovation Team Project, No.O88RA203SA; National Natural Science Foundation of China, No.60970014, 60875040; Foundation of Doctoral Program Research of the Ministry of Education of China, No.200801080006; Natural Science Foundation of Shanxi Province, No.2010011021-1
Author: Ge Yong (1972–), Ph.D, specialized in spatial data analysis and accuracy assessment. E-mail: gey@lreis.ac.cn
***Corresponding author:** Wang Yingjie (1961–), Professor, E-mail: wangyj@igsnrr.ac.cn

and spatial reasoning (Egenhofer, 1991; Frank, 1992; Guo and Du, 2009). Spatial relations can reflect the spatial features of geographical issues, and different spatial relations can be expressed by various spatial indicators. The spatial relationship indicator rules can reflect the spatial relationship of spatial objects for significant geographical issues. Since urbanization is an important geographical issue this paper focuses on the extraction of spatial indicator rules with the use of data from urban and rural areas. Rough set theory (RST) is used to extract the spatial indicator rules thereby advancing the field of research on the urbanization process.

In previous studies researchers have demonstrated that urbanization could be studied with the use of factor analysis of urban and rural areas (Belsky, 1990; Potter and Unwin, 1995; Tacoli, 1998; Lin, 2001; Verburg *et al*., 2004a; Verburg *et al*., 2004b; Oluwasola *et al*., 2008), and urban network theories (Batty and Xie, 1994). In addition, urbanization has been studied with cellular automata (Li and Yeh, 2000; Yeh and Li, 2001; Li and Yeh, 2002), and statistics and regression analytical techniques (Verburg *et al*., 2004a; Oluwasola *et al*., 2008). Since research on the urbanization process must consider the spatial relations between urban and other spatial entities such as roads, rivers (Yeh and Li, 2003; Yeh and Li, 2006), this paper studies the spatial relations of urban and rural areas with several spatial relationship indicators and spatial indicator rules. The extracted spatial relationship indicator rules will facilitate the analysis of urbanization. Rough set is used to extract the spatial relationship indicator rules of urban and rural areas from aspects of spatial data mining.

2 Brief remarks on rough set theory and its applications

Rough Set Theory (RST), proposed by Pawlak (1982), is an extension of classical set theory for use when representing incomplete knowledge. Though RST does not need prior information in analyzing data, it can derive a classification or decision rules according to knowledge reduction procedures. Hence, RST can mine the objective and inherent rules implicit in the data. Since RST is useful in analyzing data with insufficient and incomplete knowledge it has been applied in many traditional domains including finance, medicine, telecommunications, vibration analysis, control theory, signal analysis, pattern recognition, and image analysis (Yasdi, 1996; Polkowski and Skowron, 1998; Polkowski *et al*., 2000; Skowron, 2001; Leung and Li, 2003). The literature has also revealed that RST has been applied to problems of spatial analysis (Bittner, 2001; Bittner and Stell, 2001); spatial classification and uncertainty analysis (Ahlqvist *et al*., 2000, 2003), geo-knowledge discovery (Wang *et al*., 2001; Beaubouef *et al*., 2007), remote sensing image classification (Dong *et al*., 2007), and in the extraction of decision rules in GIS and remote sensing (Berger, 2004; Leung *et al*., 2007; Bai *et al*., 2009). Moreover, Cao *et al*. (2009) initially attempted to extract the spatial relationship indicator rules based on RST. Though the results demonstrated that the spatial relationship indicator rules can reflect to some extent the spatial features of urban and rural areas, the study did not explain the causes of extracted rules and assess the applicability of these rules. To further analyze the applicability of extracted rules and extend the application of RST, the spatial issue of land use in two regions of urban and rural areas in different years is considered in this paper.

3 Use of RST to address the issue of land use

The land use vector data of urban and rural areas in the Pearl River Delta are used to illustrate how the RST is used to extract the spatial relationship indicator rules. The Pearl River Delta is one of the fastest growing metropolitan regions in China. Economic reforms are stimulating the acceleration of land use. The spatial relationship indicator rules of urban and rural areas of Shenzhen and Hong Kong, located in the Pearl River Delta, are extracted for the year 1995. The extracted spatial relationship indicator rules are then used to identify, for 1995, the urban and rural areas in Zhongshan, Zhuhai and Macao which are other regions in the Pearl River Delta. In order to explain the effectiveness of the employed method, the spatial relationship indicator rules of urban and rural areas in Shenzhen and Hong Kong for the year 2000 are also extracted and then used to identify the urban and rural areas in Zhongshan, Zhuhai and Macao for the year 2000.

4 Applying RST to extract spatial relationship indicator rules

The extraction process of the spatial relationship indicator rules with rough set theory mainly contains four procedures from steps (a) to (d) and they are described below.

(a) Expression of spatial relationship indicators with decision table

Before the extraction of the spatial relationship indicator rules, it is necessary to express the various spatial relationship indicators of different spatial relations with a two-dimension decision table. The rows of the decision table correspond to the objects of the geographical issues. The columns of the decision table are divided into two parts: the former is called the condition attribute which contains all the spatial relationship indicators, while the latter usually has only one column and is called the decision attribute which corresponds to the geographical results. The construction process of decision table for spatial relationship indicators mainly includes three steps: selection of related spatial relationship indicators, description of the spatial relationship indicators and expression of spatial relationship indicators using a decision table (Cao *et al*., 2009). For example, the related spatial relationship indicators of a geographical issue, such as urbanization, are selected on the basis of prior knowledge and previous research efforts. They are then described by appropriate quantitative or qualitative values and expressed by a decision table.

(b) Discretization of decision table

The description of the spatial relationship indicators is either nominal (categorical), or continuous (numerical). The term "continuous" is used to indicate both real and integer valued attributes. The attribute selection process in RST assumes that all attributes are nominal, so continuous-valued attributes must, therefore, be discretized prior to attribute selection (Fayyad and Irani, 1992). The discretization process determines how coarsely the world is viewed. Hence, the decision table that is constructed is discretized with an appropriate discretization method. Reliable methods have been forwarded by Fayyad and Irani (1992); Holte (1993), and Dougherty *et al*. (1995).

(c) Reduction of spatial relationship indicators

Spatial relationship indicators selected in the construction of spatial relationship indicator decision table are not all equally important with some even being redundant. Hence, it is necessary to reduce those spatial relationship indicators which are unimportant. The mathe-

matical model that is used for reducing redundant spatial relationship indicators is that proposed by Ohrn (1999). Essentially, the foundation for data representation in RST is an information system. This paper uses the subclass of information system decision system. A decision system A of spatial relationship indicators is defined in terms of a pair $(U, C \cup d)$, where U is a non-empty finite set of various objects belonging to a certain geographical issue called universe, C is a non-empty finite set of spatial relationship indicators called condition attribute and d is the result of certain geographical issue called decision attribute. The decision attribute d induces a partition of the universe of objects U. Without any loss of generality, it is assumed that V_d (decision values) is the set of n result of a certain geographical issue $\{g_1(d), \cdots, g_n(d)\}$, where $g_i(d)$ is said to be one result of the geographical issue of d. The induced partition is therefore the collection of equivalence classes $\{X_{g_1(d)}, \cdots, X_{g_n(d)}\}$, called decision classes, where two objects are said to belong to the same decision class if they have the same result for the decision attribute.

$$X_i = \{x \in U \mid d(x) = g_i(d)\} \tag{1}$$

In a decision system A it is possible that two objects that are indiscernible with respect to attributes C may belong to different decision classes. The decision system A is said to be inconsistent with respect to C if this is the case, and consistent otherwise. The generalized decision attribute ∂_C is a function $\partial_C : U \to 2^{V_d}$ that, when applied to an objects x, produces the set of results that the objects in the indiscernibility set $R_C(x)$ take on for the decision attribute d,

$$\partial_A(x) = \{v \in V_d \mid \exists y \in R_C(x) \; such \; that \; d(y) = v\} \tag{2}$$

$R_C(x)$ consists of those objects that stand in relation to object x by R_C which is an equivalence relation having reflexivity, symmetry and transitivity properties.

$$R_C(x) = \{y \in U \mid xR_C y\} \tag{3}$$

Note that the decision system A is consistent if and only if $\partial_A(x)$ are singletons for all $x \in U$.

The discernibility matrix M_C^d where $|U| \times |U|$ matrix. $M_C^d(x, y) \subseteq C$ is the element of M_C^d, which consists of the set of attributes that can be used to discern objects $x, y \in U$:

$$M_C^d(x, y) = \{a \in C \mid \partial_A(x) \neq \partial_A(y) \; and \; a(x) \neq a(y)\} \tag{4}$$

From a discernibility matrix M_C^d, the discernibility function is constructed relative to an object $x \in U$, as shown below. The function $f_C^d(x)$ is a Boolean product-of-sums (POS) function of $|C|$ Boolean variables, where variable a^* corresponds to attribute a,

$$f_C^d(x) = \prod_{x \in U} \left\{ \sum a^* \,\middle|\, a \in M_C^d(x, y) \; and \; M_C^d(x, y) \neq \phi \right\} \tag{5}$$

The prime implicants of $f_C^d(x)$ reveal the minimal subsets of C that are needed to determine to which decision class object x belongs. The set of all reducts of a decision system A that are relative to an object x is denoted $RED(C, x, d)$.

Reducts in $RED(C,x,d)$ preserve at least the indiscernibility set $R_C(x)$. These reducts are minimal attribute subsets $B \subseteq C$ that permit the identification of the objects in $R_C(x)$, along with some other objects. These other objects, however, are all required to have the same result for the generalized decision attribute as x.

Formally, there are now the following relationships:

$$B \in RED(C,x,d) \Leftrightarrow \prod_{a \in B} a^* \ is\ a\ prime\ implicant\ of\ f_C^d(x) \tag{6}$$

The problem of computing prime implicants of Boolean POS functions is easily transformed into the problem of computing minimal hitting sets (Skowron and Rauszer 1991).A hitting set of a given bag or multiset S of elements from 2^C is a set $B \subseteq C$ such that the intersection between B and every set in S is non-empty. A bag or multiset is conceptually an unordered collection of elements where the same element may occur more than once. The set $B \in HS(S)$ is a minimal hitting set of S if B ceases to be a hitting set if any of its elements are removed. Let $HS(S)$ and $MHS(S)$ denote the sets of hitting sets and minimal hitting sets, respectively.

$$HS(S) = \{B \subseteq A \mid B \cap S_i = \phi \ for\ all\ S_i\ in\ S\} \tag{7}$$

Let h denotes any Boolean POS function of m Boolean variables $\{a_1^*, \cdots, a_m^*\}$, composed of n Boolean sums $\{s_1, \cdots, s_n\}$. Now h can be interpreted as a bag or multiset $S(h)$ with n elements as follows:

$$S(h) = [S_i \mid S_i = \{a_j \in A \mid a_j^* \ occurs\ in\ s_i\}] \tag{8}$$

The multiset constructor S can interpreted with the following example shows:

$$S((a^* + b^*) \cdot (a^* + b^*) \cdot (c^*)) = [\{a,b\}, \{a,b\}, \{c\}] \tag{9}$$

A hitting set of $S(h)$ obviously defines an implicant of h and subsequently, a minimal hitting set corresponds to a prime implicant. By relating this connection to reducts, we thus can obtain the following relationship:

$$B \in RED(C,x,d) \Leftrightarrow B \in MHS(S(f_C^d(x))) \tag{10}$$

For each x in U, $RED(C,x,d)$ is computed sequentially, and the final reducts set of spatial relationship indicators is the set by overlaying $RED(C,x,d)$.

(d) Extraction of the spatial relationship indicator rules

After computing the reducts set of spatial relationship indicators, the spatial relationship indicator rules are easily constructed by overlaying the reducts over the originating spatial relationship indicator decision table and reading off the values.

Symbolically the spatial relationship indicator rules have been denoted as (Wang, 2001):

$$Des(A) \Rightarrow Des(d) \tag{11}$$

where $Des(A)$ is the conjunction of the values corresponding to the spatial relationship indicators set $A \subseteq C$ and $Des(d)$ is the value of the result for some geographical object.

The coverage of the spatial relationship indicator rules is defined as:

$$\beta = \frac{|X \cap Y|}{|Y|}, \tag{12}$$

while the accuracy of the spatial relationship indicator rules is defined as:

$$\alpha = \frac{|X \cap Y|}{|X|}, \tag{13}$$

where $X = \{x \mid x \in U \wedge A_x\}$ and $Y = \{x \mid x \in U \wedge d_x\}$. Here it should be noted that A_x denotes that object x possesses the values of the spatial relationship indicators A, and d_x denotes that the geographical object x possesses the value of geographical result d. For any set E, $|E|$ denotes the cardinality of set E. The accuracy of the rules provides a measure of how reliable the rule is in drawing geographical result d on the basis of spatial relationship indicators A. The coverage is indicative of how well the rules describe the decision class defined through geographical result d.

5 An empirical study

As mentioned earlier, this paper mainly focuses on the discussion of the use of he spatial relationship indicator rules extracted from one region to identify the different geographical objects in the other region. The study areas are two regions (Figure 1). One region is located in the urban and rural areas of Shenzhen and Hong Kong and the other is located in the urban and rural areas in Zhongshan, Zhuhai and Macao. The extracted spatial relationship indicator rules from the areas of Shenzhen and Hong Kong are used to identify the urban and rural areas in Zhongshan, Zhuhai and Macao. Consequently, the validation is implemented in Section 6.

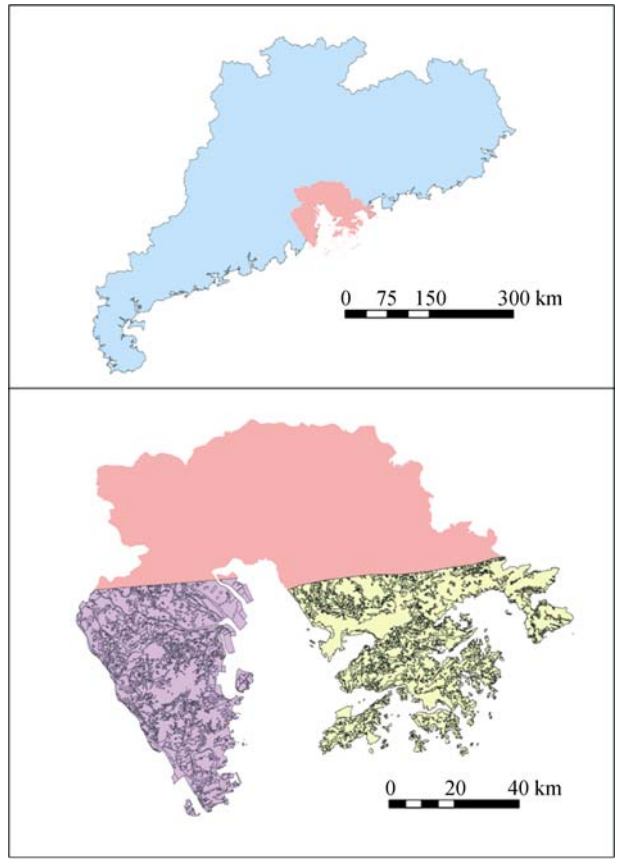

Figure 1 The study area in the Pearl River Delta

5.1 Data description

The land use vector data, mentioned in section 3, of the urban and rural areas in the Pearl River Delta are used to demonstrate how to construct the spatial relationship indicator decision table, and extract the spatial relationship indicator rules of urban and rural areas. The land use vector data (scale of 1:100,000) of the urban and rural areas corresponding to the years 1995 and 2000 have been obtained from the Data Centre for Resources and Environmental Sciences, Chinese Academy of Sciences (DCRES, CAS). Also the vector data of highway, river, etc. (scale of 1:100,000) are from DCRES,

CAS. The rapid structural and spatial transformation started in the late 1970s with government sponsored economic reforms in the Pearl River Delta (Lin, 2001). Two areas of the Pearl River Delta are selected. Area 1 comprises Shenzhen and Hong Kong with Figures 2a and 2b showing the vector data pictures for the years 1995 and 2000, respectively. Area 2 comprising Zhongshan, Zhuhai and Macao is represented by Figures 2c and 2d for the years 1995 and 2000, respectively. The statistical data for the two areas for 1995 and 2000 are provided in Tables 1 and 2, respectively.

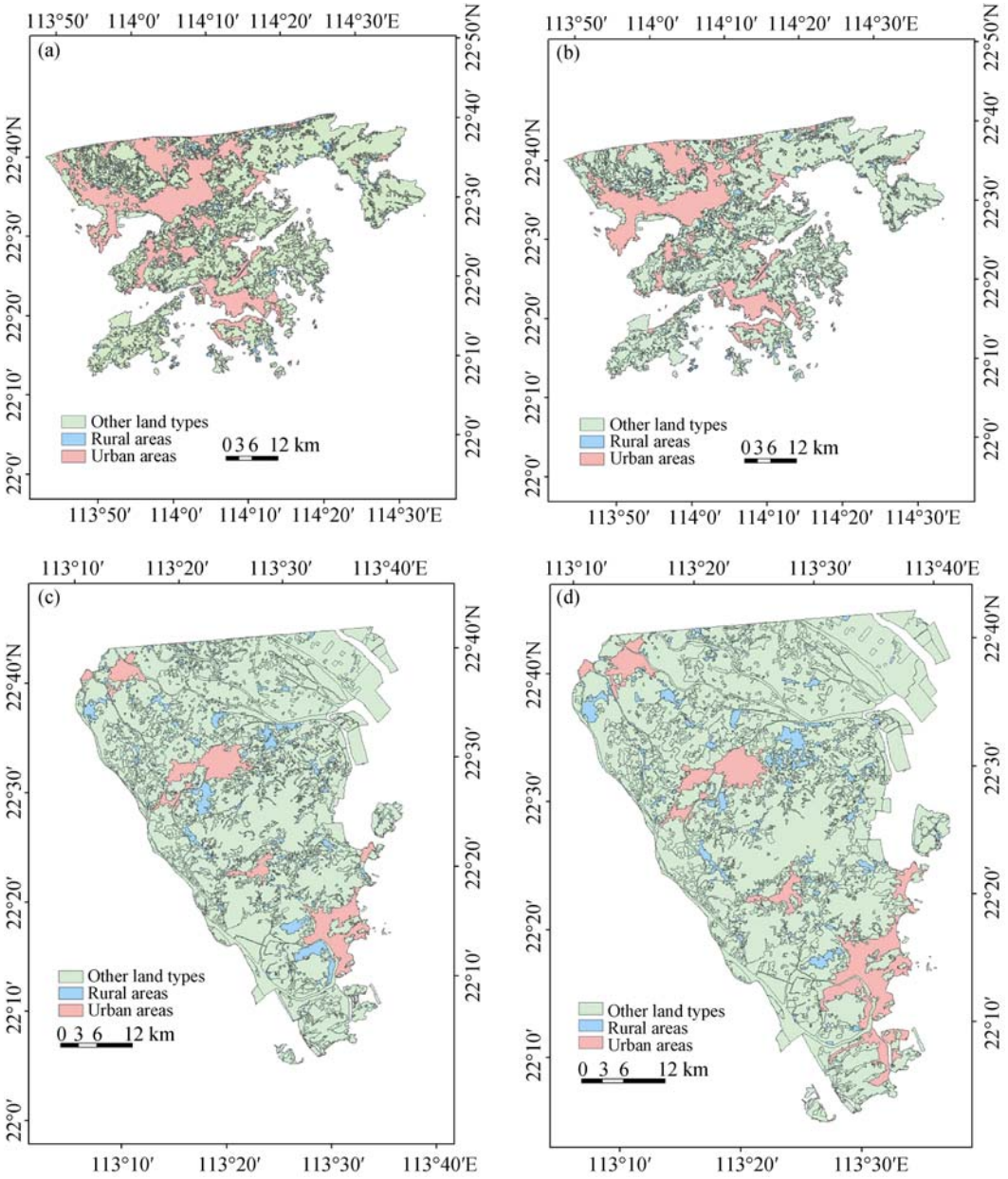

Figure 2 (a) Shenzhen and Hong Kong vector data of 1995; (b) Shenzhen and Hong Kong vector data of 2000; (c) Zhongshan, Zhuhai and Macao vector data of 1995; (d) Zhongshan, Zhuhai and Macao vector data of 2000

Table 1 Statistic data of the parcel number, ratio and area of the land types in the study areas of Shenzhen and Hong Kong (Area 1) and Zhongshan, Zhuhai and Macao(Area 2) located in the Pearl River Delta, China of 1995

Study area	Land type	Number[a]	Ratio (%)[b]	Areas (km^2)
Area 1	Urban areas	19	1.4	145.42
	Rural areas	211	15.7	95.49
	Other land types	1116	82.9	1956.41
Area 2	Urban areas	10	0.7	554.06
	Rural areas	157	11.8	64.81
	Other land types	1167	87.5	1671.14

[a]The parcel number of the land types
[b]The ratio of the parcels in the study areas

Table 2 Statistic data of the parcel number, ratio and area of the land types in the study areas of Shenzhen and Hong Kong (Area 1) and Zhongshan, Zhuhai and Macao (Area 2) located in the Pearl River Delta, China of 2000

Study area	Land type	Number[a]	Ratio (%)[b]	Areas (km^2)
Area 1	Urban areas	23	2.1	196.62
	Rural areas	166	15.0	94.49
	Other land types	916	82.9	2224.39
Area 2	Urban areas	24	2.0	479.58
	Rural areas	162	13.1	50.55
	Other land types	1049	84.9	1773.00

[a]The parcel number of the land types
[b]The ratio of the parcels in the study areas

5.2 Construction of decision table

After analyzing the characteristics of the urban and rural areas in the study areas, the following five spatial relationship indicators of urban and rural areas are considered:

A1: the number of roads in the parcels of urban areas or rural areas

A2: the number of highways in the parcels of urban areas or rural areas

A3: the number of rivers in the parcels of urban areas or rural areas

A4: the number of rural areas in a one-kilometre range around the parcels of an urban area or a rural area

A5: the number of urban areas in a three-kilometres range around the parcels of an urban area or a rural area

Transport patterns tend to reflect development and land use change of both urban and rural areas (Adams, 1970; Kilkenny, 1998; Wee, 2002; Wegener, 2004). Hence, the relationship indicators between road/highway and urban area or rural area are selected, and the extracted spatial relationship indicator rules are descriptions of the spatial features of urban and rural areas. As expected, urban areas contain more highways than rural residential areas. Since urbanization has significant impacts on the rivers of urban and rural areas (Ouyang *et al*., 2006), the relationship indicators between river and urban area or rural area are also selected. The spatial relationship indicator rules that are extracted are other descriptions of

spatial characters of urban and rural areas. To highlight the interactions between urban and rural area (Tacoli, 1998; Belsky, 1990; Oluwasola *et al*., 2008), it is necessary to consider that urban areas and rural areas are closely related to each other. Rural areas are usually clustered spatially and sometimes situated near urban areas. Hence, in this paper the indicators of A4 and A5 are selected.

Once the spatial relationship indicators of urban and rural areas are explicit, they are then described with the appropriate quantitative or qualitative value. In this paper, Visual Basic for Applications, a second development language of ArcGIS software, is used to acquire the value of the five spatial relationship indicators which are then expressed with the decision table. The rows correspond to parcels of urban areas or rural areas. In order to explain the process clearly, Table 3 is prepared. The first five columns (A1-A5) correspond to the five spatial relations mentioned above and the last column d corresponds to the land types of the parcels, including only urban area and rural area.

Table 3 Expression of spatial relations based on rough set with a two-dimension decision table which is a commonly used data presentment format required by rough set theory

Condition attributes					Decision attribute
A1	A2	A3	A4	A5	d
77	15	7	17	2	Urban area
0	0	0	6	0	Urban area
……	……	……	……	……	……
16	0	0	6	4	Urban area
2	0	0	2	0	Rural area
0	0	0	1	4	Rural area
……	……	……	……	……	……
3	0	0	0	0	Rural area

5.3 Extraction of the spatial relationship indicator rules

After expressing the spatial relationship indicators with a decision table, the spatial relationship indicator rules of urban and rural areas in Shenzhen and Hong Kong are extracted. The free ROSETTA(version 1.4.41, www.idi.ntnu.no/~aleks/rosetta/) software is used to execute the extraction process. ROSETTA, first developed by the Department of Computer and Information Science of the Norwegian University of Science and Technology and the Department of Mathematics of Warsaw University, is an analytical tool of data tables based on rough set theory (Ohrn, 1999; Komorowski and Ohrn, 1999; Berger, 2004). The extraction process of the spatial relationship indicator rules is mainly divided into the following four steps (a to d):

(a) Preprocessing

Import the quantitative or qualitative value of the spatial relationship indicators into the ROSETTA, and express the spatial relationship indicators with a decision table.

(b) Discretize the value of spatial relationship indicators

Several discretization algorithms are provided in ROSETTA, such as Boolean reasoning

algorithm, Equal frequency binning, Naive algorithm, Semi-naive algorithm and Entropy/MDL algorithm. In this paper, the Entropy/MDL algorithm is used. The algorithm, described by Dougherty *et al*. (1995), is based on recursively partitioning the value set of each attribute so that a local measure of entropy is optimized. The minimum description length principle defines a stopping criterion for the partitioning process. After the discretization, the values of the decision table become nominal, therefore meeting the requirements of RST.

(c) Reduct of the spatial relationship indicators

There are also several reduction algorithms in ROSETTA, such as Genetic algorithm, Johson's algorithm, and Holte's 1R. In this paper, the Genetic algorithm is used and the main options selected for this algorithm are discernibility equal object related (universe equal all objects), and table interpretation equal Modulo. The Genetic algorithm is used in ROSETTA to compute the minimal hitting set (Vinterbo and Ohrn, 2000).

The minimal hitting set of the spatial relationship indicators is called a reduct set of spatial relationship indicators. The algorithm supports both cost information and approximate solutions. The information about attribute costs is used to steer the algorithm towards finding low-cost solutions. Approximate solutions are found via minimal approximate hitting sets. The algorithm's fitness function f is defined below,

$$f(B) = (1-\alpha) \times \frac{\cos t(A) - \cos t(B)}{\cos t(A)} + \alpha \times \min\left\{\varepsilon, \frac{\left|[S \; in \; \delta | S \cap B \neq \phi]\right|}{\left|\delta\right|}\right\} \tag{14}$$

where δ is the set of sets corresponding to the discernibility function. The parameter α defines a weighting between subset cost and hitting fraction while ε is relevant in the case of approximate solution. Here, α is set at 0.4 in the case of approximate solutions. The subsets B of all attributes set A are found through an evolutionary search driven by the fitness function. They are "good enough" hitting sets if they have a hitting fraction of at least ε, and are collected in a "keep list". Also, $Cost(A) = \left|A\right|$ and $Cost(B) = \left|B\right|$. For a set S, $|S|$ is the cardinality of S. The size of keep list can be specified. Approximate solutions are controlled through two parameters, ε and k. The parameter ε signifies a minimal value for the hitting fraction, while k denotes the number of extra keep lists in use by the algorithm. If k=0, then only minimal hitting sets with a hitting fraction of approximately ε are returned. If k>0, then k+1 groups of minimal hitting sets are returned (Aydogan and Gencer, 2008). Here, the ε and k are set at 0.90 and 1 respectively. Some other parameters of Genetic algorithm include population size, crossover probability, mutation probability and inversion probability, which default values are 70, 0.3, 0.05, and 0.05, respectively.

(d) Extract the spatial relationship indicator rules

Once the reducts set of spatial relationship indicators have been computed, the spatial relationship indicator rules are constructed by overlaying the reducts over the originating spatial relationship indicator decision table and reading off the values. The results from ROSETTA are presented in Tables 4 and 5 which are the spatial relationship indicator rules of the urban and rural areas in Shenzhen and Hong Kong for the years 1995 and 2000, respectively.

6 Validation

The extracted spatial relationship indicator rules are tested and analyzed. The land use data

Table 4 Spatial relationship indicator rules of urban and rural areas of Shenzhen and Hong Kong extracted by rough set theory in 1995

Code	Rules	Accuracy (%)	Coverage (%)
1	A1<4 and A3<1 and A4<7=>⇒rural area or urban area	95.9, 4.1	95.9, 47.4
2	2≤A3⇒urban area	100	31.6
3	11≤A1⇒urban area	100	31.6
4	4≤A2⇒urban area	100	26.3
5	4≤A1<8⇒urban area	100	10.5
6	13≤A4 ⇒urban area	100	10.5
7	5≤A5⇒urban area	100	5.2
8	1≤A3<2⇒rural area	100	3.1
9	6≤A1<11⇒rural area	100	0.4
10	7≤A4<13⇒rural area	100	0.4

Table 5 Spatial relationship indicator rules of urban and rural areas of Shenzhen and Hong Kong extracted by rough set theory in 2000

Code	Rules	Accuracy (%)	Coverage (%)
1	A1<3 and A4<2⇒rural area or urban area	91.4, 8.6	66.0, 41.7
2	A1<3 and 2≤A4<5⇒rural area or urban area	98.1, 1.9	32.7, 4.2
3	11≤A1⇒urban area	100	25
4	3≤A2⇒urban area	100	25
5	5≤A4⇒urban area	100	25
6	3≤A1<5⇒urban area	100	12.5
7	6≤A1<9⇒urban area	100	8.3
8	6≤A3⇒urban area	100	8.3
9	5≤A5⇒urban area	100	4.1
10	A1<3 and 2≤A3<6⇒urban area	50	4.1
11	A2<3 and 2≤A3<6⇒rural area	50	0.6
12	9≤A1<11⇒rural area	100	0.6
13	5≤A1<6 ⇒rural area	100	0.6

of Zhongshan, Zhuhai and Macao are used to validate the spatial relationship indicator rules extracted from Shenzhen and Hong Kong. The two areas are very near in distance and have similar geographical environments, and are both located in the Pearl River Delta of China. Both areas have experienced fairly similar processes influencing their development.

6.1 Validation of 1995

The 1995 land use data of the urban and rural areas in Zhongshan, Zhuhai and Macao are used to validate the spatial relationship indicator rules extracted from Shenzhen and Hong Kong. Findings are demonstrated by confusion matrices (Table 6), accuracy statistics (Table 7), and identification results (Figure 3).

Table 6 Confusion matrix about the urban and rural areas of Shenzhen and Hong Kong which can show the identification results with the extracted spatial relationship indicator rules of urban and rural areas of 1995 and 2000

	Actual data					
	1995			2000		
Classified data	Urban area	Rural area	Row total	Urban area	Rural area	Row total
Urban area	4	0	4	10	18	28
Rural area	6	157	163	13	148	161
Column total	10	157	167	23	166	189

Table 7 Producer's accuracy, user's accuracy and overall accuracy of confusion matrix about the urban and rural areas of Shenzhen and Hong Kong of 1995[a] and 2000[b]

	1995		2000	
Classified data	Producer's accuracy (%)	User's accuracy (%)	Producer's accuracy (%)	User's accuracy (%)
Urban area	40	100	43.5	35.7
Rural area	100	96.3	89.2	91.9

[a]Overall accuracy (%)=96.4
[b]Overall accuracy (%)=83.6

The method applied in this paper correctly identified 161 parcels of urban and rural areas. As shown in Table 7 the overall accuracy is 96.4%. It is demonstrated that the urban and rural areas in Shenzhen and Hong Kong have similar spatial relationship features with the urban and rural areas of Zhongshan, Zhuhai and Macao. The extracted spatial relationship indicator rules can reflect, to some extent, those spatial features. In Table 6, six parcels are identified incorrectly, five of which match rule 1, which is non-consistent. The identification accuracy of rural area is 95.9% while the identification accuracy of urban area is 4.1%. Given the much better identification accuracy of the rural areas over that of the identification accuracy of the urban areas, the parcels matching rule 1 are identified as rural areas. Another areal portion which matches rule 8 is identified incorrectly as rural areas.

The rules which coverage is higher than 25% can reflect the main spatial features in Shenzhen and Hong Kong in Table 4. From rule 1 to 4, we know that the 95.9% rural areas and 47.4% urban areas have the spatial features that the roads, rivers in them are less and the rural areas in a one-kilometer range surrounding them are also less. Also, the 31.6% urban areas have more than two rivers, 31.6% urban areas have more than 11 roads and 26.3% urban areas have more than four highways. Relatively, conditional attributes A_1, A_2 and A_3 when constructing the rules are important in reflecting the spatial features of urban area and rural area in the year of 1995.

Without loss of generality, the spatial relationship indicator rules presented in Table 4, with accuracies higher than 25%, are analyzed. Rule 1 is that if the number of highways in the parcels is less than 4, the number of rivers is less than 1, and the number of rural areas in a one-kilometer range surrounding each parcel is less than 7, then the parcels are identified as rural area with an accuracy of 95.9%. Rural areas have less highways. In some instances some rural areas have spatial aggregation, thereby giving rise to several rural parcels within

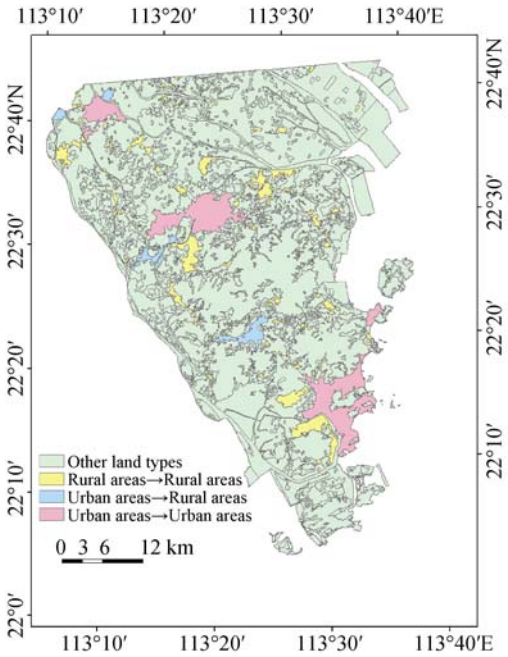

Figure 3 Identification results of Zhongshan, Zhuhai and Macao with the spatial relationship indicator rules of Shenzhen and Hong Kong of 1995

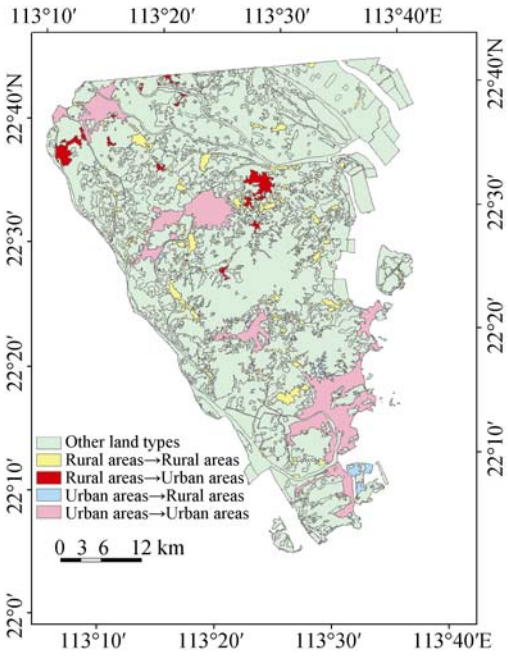

Figure 4 Identification results of Zhongshan, Zhuhai and Macao with the spatial relationship indicator rules of Shenzhen and Hong Kong of 2000

a one-kilometer range in the rural areas. When some parcels of urban areas have similar characteristics with rural areas, they both match rule 1. This is the reason why six parcels are identified incorrectly in Table 6. Statistically, 221 parcels match rule 1 in Shenzhen, and Hong Kong. If only rule 1 is used to identify the urban and rural areas in Zhongshan, Zhuhai and Macao, 155 rural areas are correctly identified. Hence, rule 1 is reasonable in highlighting the spatial features of rural areas. The number of parcels in Shenzhen and Hong Kong which match rules 2, 3 and 4 are 6, 6 and 5 respectively. The ratios of the parcels to all urban areas are approximately 32%, 32% and 26%, respectively. The analysis of rules 2, 3 and 4 support observations that urban areas usually have more rivers, roads and highways than rural areas. Evidently, rules 2, 3 and 4 are reasonable in reflecting the spatial features of urban areas.

6.2 Validation of 2000

The land use data of urban and rural areas in Zhongshan, Zhuhai and Macao of 2000 are also used to validate the spatial relationship indicator rules extracted from Shenzhen and Hong Kong of 2000. Findings are highlighted by the confusion matrices (Table 6), accuracy statistics (Table 7), and identification results (Figure 4).

From Table 6, it can be seen that 158 parcels are correctly identified, with 31 being incorrectly identified. The overall accuracy is 83.6%. The identification accuracy of 2000 is lower than that of 1995, thereby demonstrating that over time the two areas have different spatial features.

As mentioned above, the rules which coverage is higher than 25% is still the main rule in Table 5. It can be seen from the main rules that the spatial relationship indicators A_1, A_2 and A_4 can to some extent reflect the main spatial features in Shenzhen and Hong Kong in the

year of 2000.

The spatial relationship indicator rules presented in Table 5, with accuracies higher than 25%, are also analyzed. Rule 1 shows that if the number of roads in the parcels is less than 3, and the number of rural areas in a one-kilometer range around the parcels is less than 2, the parcels are identified as rural areas. With rule 1 being non-consistent and the identification accuracy of rural areas is 91.4%, then the parcels which match rule 1 are identified as rural areas. According to rule 2, if the number of roads in the parcels is less than 3, and the number of rural areas in a one-kilometer range around the parcels is between 2 and 5, then the parcels are identified as rural areas. Rule 2 is also non-consistent, and the identification accuracy of rural areas is 98.1%. Hence, the parcels which match rule 2 are also identified as rural areas. The numbers of parcels in Shenzhen and Hong Kong which match rule 1 and 2 are 117 and 54, respectively. If rule 1 and rule 2 are used to identify the urban and rural areas in Zhongshan, Zhuhai and Macao, the number of rural areas correctly predicted are 92 and 56, respectively. The ratios of the correctly predicted rural areas to all rural areas are nearly 56.8% and 34.6%, respectively. These results emphasize that rule 1 and rule 2 reflect the spatial features of rural areas. The number of parcels in Shenzhen and Hong Kong which match rules 3, 4 and 5 are 6, 6 and 6, respectively. Each of the ratios of the parcels to all urban areas is nearly 26.1%. When these rules are used to identify the urban and rural areas in Zhongshan, Zhuhai and Macao, the number of urban areas identified correctly is 3, 9 and 5, respectively. The ratios of correctly predicted urban areas to all urban areas are approximately 12.5%, 37.5% and 20.1%, respectively. From this finding it is evident that rules 3, 4 and 5 reflect the spatial features of urban areas.

It can be seen from the result, the consistent spatial relationship indicator rules, which coverage is higher than 25%, reflect the spatial features of urban areas. For example, in the spatial relationship rules of 1995, if the number of rivers in the parcels is more than 2, or the number of roads is more than 11, or the number of highways is more than 4, then the parcels are identified as urban area. And in the spatial relationship rules of 2000, if the number of roads in the parcels is more than 11, or the number of highways is more than 3, or the number of rural areas in a one-kilometre range around the parcels is more than 5, then the parcels are identified as urban area. But, the consistent spatial relationship indicator rules, which coverage is lower, reflect the spatial features of rural areas. For example, in the spatial relationship rules of 1995, if the number of rivers in the parcels is between 6 and 11, or the number of rural areas in a one-kilometre range around the parcels is between 7 and 13, then the parcels are identified as rural areas. The coverage of these rules is only 0.4%. Also, in the spatial relationship rules of 2000, if the number of highways is less than 3 and the number of rivers is between 2 and 6, or the number of roads is between 9 and 11, or the number of rivers is between 5 and 6, then the parcels are identified as rural areas. The coverage of these rules is 0.6%. Obviously, the spatial relationship indicator rules of urban areas are different from the spatial relationship indicator rules of rural areas. The reasons lie in the difference between the spatial features of urban areas and the spatial features of rural areas.

7 Conclusions and future works

Spatial relationship indicators of spatial relations reflect the spatial characteristics of geo-

graphical phenomena, and are beneficial for the solution of geographical issues and problems. Although geographical issues are affected by many interrelated factors, the spatial relationship indicator rules could be used to identify, isolate and analyze specific factors governing geographical issues. This paper presents a comprehensive account on the use of RST to extract spatial relationship indicator rules. In addition to utilizing a decision table, the known discretization and reduction algorithms in RST have been used to extract the spatial indicator rules. To illustrate the extraction process empirical land use vector data from the Pearl River Delta have been used.

While success has been obtained in extracting spatial indicator rules to identify, with some degree of accuracy, rural and urban areas it is, nevertheless, worthwhile to note that the discretization and reduction algorithms may have influenced the spatial relationship indicator rules. Further research, therefore, remains to be done not only on the selection of the best discretization and reduction algorithms but also on the use of techniques which provide an indication of spatial indicator rules which have a major influence on land use changes. By knowing the spatial relationship indicator rules which govern land use changes it will be possible to predict future land uses. This would certainly be beneficial to land use managers and policy planners who have to solve emerging geographical issues and problems. In addition to the use of RST, Du *et al*. (2002) emphasized that contemporary avenues of research can now pursue geo-case based reasoning whereby artificial intelligence methods are employed to account for spatial relationship indicator rules. Since geo-case based reasoning uses the method of case-based reasoning, it will be possible to have enhanced solutions to specific geographical issues.

References

Adams J S, 1970. Residential structure of midwestern cities. *Annals of the Association of America Geographers*, 60(1): 37–62.

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

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

Aydogan E K, Gencer C, 2008. Mining classification rules with reduced MEPAR-miner algorithm. *Applied Mathematics and Computation*, 195(2): 786–798.

Bai H X, Ge Y, Liao Y L *et al*., 2010. Using rough set theory to identify villages affected by birth defects: The example of Heshun, Shanxi, China. *International Journal of Geographical Information Science*, 24(4): 559–576.

Batty M, Xie Y, 1994. From cells to cities. *Environment and Planning B: Planning and Design*, 21(7): s31–s48.

Beaubouef T, Petry F E, Ladner R, 2007. Spatial data methods and vague regions: A rough set approach. *Applied Soft Computing*, 7(1): 425–440.

Belsky E S, 1990. Approaches to locating urban functions in developing rural areas. *International Regional Science Review*, 13(3): 225–240.

Berger P A, 2004. Rough set rule induction for suitability assessment. *Environmental Management*, 34(4): 546–558.

Bittner T, 2001. Rough sets in spatio-temporal data mining. In: Roddick J F, Hornsby K (eds.). Temporal, Spatial, and Spatio-temporal Data Mining. Berlin: Springer-Verlag, 89–104.

Bittner T, Stell J G, 2001. Rough sets in approximate spatial reasoning. In: Ziarko W, Yao Y (eds.). Rough Sets

and Current Trends in Computing. Berlin: Springer-Verlag, 445–453.

Cao F, Du Y Y, Ge Y *et al.*, 2009. Extraction of geo-spatial relationship rules based on rough set theory: Exemplified by land use. *Journal of Geographical Information Science*, 11: 139–144. (in Chinese)

Clementini E, Di Felice P, Koperski K, 2000. Mining multiple-level spatial association rules for objects with a broad boundary. *Data & Knowledge Engineering*, 34(3): 251–270.

Dong G J, Zhang Y S, Fan Y H, 2007. Remote sensing image classification algorithm based on rough set theory. In: Cao B Y (ed.), Fuzzy Information and Engineering. Berlin/Heidelberg: Springer-Verlag, 846–851.

Dougherty J, Kohavi R, Sahami M, 1995. Supervised and unsupervised discretization of continuous features. In: Priedits A, Russell S (eds.). Proceeding of the Twelfth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 194–202.

Du Y Y, Zhou C H, Shao Q Q *et al*., 2002. Theoretic and application of geo-case based reasoning. *Acta Geographica Sinica*, 57(2): 151–158. (in Chinese)

Egenhofer M, 1991. Reasoning about binary topological relations. In: Gunther O, Schek H J (eds.). Advances in Spatial Databases. Berlin: Springer-Verlag, 141–161.

Egenhofer M, 1997. Query processing in spatial-query-by sketch. *Journal of Visual Languages and Computing*, 8(4): 403–424.

Fayyad U M, Irani K B, 1992. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8(1): 87–102.

Frank A U, 1992. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages and Computing*, 3: 343–371.

Guo L, Du S H, 2009. Deriving topological relations between regions from direction relations. *Journal of Visual Languages and Computing*, 20: 368–384.

Holte R C, 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1): 63–90.

Kilkenny M, 1998. Transport costs and rural development. *Journal of Regional Science*, 38(2): 293–312.

Komorowski J, Ohrn A, 1999. Modelling prognostic power of cardiac tests using rough sets. *Artificial Intelligence in Medicine*, 15: 167–191.

Leung Y, Fung T, Mi J S *et al*., 2007. A rough set approach to the discovery of classification rules in spatial data. *International Journal of Geographical Information Science*, 21(9): 1033–1038.

Leung Y, Li D Y, 2003. Maximal consistent block technique for rule acquisition in incomplete information systems. *Information Sciences*, 153(1): 85–106.

Li X, Yeh A G O, 2000. Modelling sustainable urban development by the integration of constrained cellular automata and GIS. *International Journal of Geographical Information Science*, 14(2): 131–152.

Li X, Yeh A G O, 2002. Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science*, 16(4): 323–343.

Lin G C S, 2001. Evolving spatial form of urban-rural interaction in the Pearl River Delta, China. *Professional Geographer*, 53(1): 56–70.

Ohrn A, 1999. Discernibility and Rough Sets in Medicine: Tools and Applications, Computer and Information Science. Trondheim: Norwegian University of Science and Technology.

Oluwasola O, Ldowu E O, Osuntogun D A, 2008. Increasing agricultural household incomes through rural-urban linkages in Nigeria. *African Journal of Agricultural Research*, 3(8): 566–573.

Ouyang T P, Zhu Z Y, Kuang Y Q, 2006. Assessing impact of urbanization on river water quality in the Pearl River Delta economic zone, China. *Environmental Monitoring and Assessment*, 120: 313–325.

Papadias D, Theodoridis Y, 1997. Spatial relations, minimum bounding rectangles, and spatial data structures. *International Journal of Geographical Information Science*, 11(2): 111–138.

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

Polkowski L, Skowron A, 1998. Rough Sets in Knowledge Discovery 1: Methodology and Applications, 2: Applications. Heidelberg: Physica-Verlag.

Polkowski L, Tsumoto S, Lin T Y, 2000. Rough Set Methods and Applications. Heidelberg: Physica-Verlag.

Potter R B, Unwin T, 1995. Urban-rural interaction: Physical form and political process in the third world. *Cities*, 12(1): 67–73.

Skowron A, 2001. Rough sets and Boolean reasoning. In: Pedrycz W (ed.), Granular Computing: An Emerging Paradigm. Heidelberg: Physica-Verlag, 95–124.

Skowron A, Rauszer C, 1991. The discernibility matrices and functions in information systems. In: Slowinski R (ed.). Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer Academic Publisher, 331–362.

Tacoli C, 1998. Rural-urban interactions: A guide to the literature. *Environment and Urbanization*, 10(1): 147–166.

Verburg P H, de Nijs T C M, van Eck J R *et al*., 2004a. A method to analyse neighbourhood characteristics of land use patterns. *Computers*, *Environment and Urban Systems*, 28(6): 667–690.

Verburg P H, van Eck J R R, de Nijs T C M *et al.*, 2004b. Determinants of land-use change patterns in the Netherlands, *Environment and Planning B*: *Planning and Design*, 31(1): 125–150.

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

Wang G Y, 2001. Rough Set Theory and Knowledge Acquisition. Xi'an: Xi'an Jiaotong University Press. (in Chinese)

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

Wee B V, 2002. Land use and transport: research and policy challenges. *Journal of Transport Geography*, 10: 259–271.

Wegener M, 2004. Overview of Land-use Transport Models: Transport Geography and Spatial Systems. Kidlington: Pergamon/Elsevier Science.

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

Yeh A G O, Li X, 2001. A constrained CA model for the simulation and planning of sustainable urban forms by using GIS. *Environment and Planning B*: *Planning and Design*, 28(5): 733–753.

Yeh A G O, Li X, 2003. Simulation of development alternatives using neural networks, cellular automata and GIS for urban planning. *Photogrammetric Engineering and Remote Sensing*, 69(9): 1043–1052.

Yeh A G O, Li X, 2006. Errors and uncertainties in urban cellular automata. *Computers, Environment and Urban Systems*, 30(1): 10–28.