sciendo

Polish Cartographical Review
Vol. 52, 2020, no. 1, pp. 1–15
DOI: 2478/pcr-2020-0001
Received: 02.03.2020
Accepted: 25.03.2020

ANNA FIEDUKOWICZ
Warsaw University of Technology
Faculty of Geodesy and Cartography
orcid.org/0000-0001-9609-8185
anna.fiedukowicz@pw.edu.pl

The use of rough rules in the selection of topographic objects for generalizing geographical information

Abstract. Selection is a key element of the cartographic generalisation process, often being its first stage. On the other hand it is a component of other generalisation operators, such as simplification. One of the approaches used in generalization is the condition-action approach. The author uses a condition-action approach based on three types of rough logics (Rough Set Theory (RST), Dominance-Based Rough Set Theory (DRST) and Fuzzy-Rough Set Theory (FRST)), checking the possibility of their use in the process of selecting topographic objects (buildings, roads, rivers) and comparing the obtained results. The complexity of the decision system (the number of rules and their conditions) and its effectiveness are assessed, both in terms of quantity and quality – through visual assessment. The conducted research indicates the advantage of the DRST and RST approaches (with the CN2 algorithm) due to the quality of the obtained selection, the greater simplicity of the decision system, and better refined IT tools enabling the use of these systems. At this stage, the FRST approach, which is characterised by the highest complexity of created rules and the worst selection results, is not recommended. Particular approaches have limitations resulting from the need to select appropriate measurement scales for the attributes used in them. Special attention should be paid to the selection of network objects, in which the use of only a condition-action approach, without maintaining consistency of the network, may not produce the desired results. Unlike approaches based on classical logic, rough approaches allow the use of incomplete or contradictory information. The proposed tools can (in their current form) find an auxiliary use in the selection of topographic objects, and potentially also in other generalisation operators.

Keywords: cartographic generalization, Rough Set Theory, ominance-Based Rough Set Theory, Fuzzy-Rough Set Theory, decision rules

1. Introduction

Selection plays a special role in cartographic generalization. It is sometimes included in the basic generalisation operators, e.g. as “elimination” (R. Regnauld et al. 2011). At other times, it is treated as a preliminary step (preprocessing) before the proper generalization of spatial information (K.S. Shea, R.B. McMaster 1989). Nevertheless, it is usually the first operation and as such has a significant impact on the overall result of the generalization process.

The automation of cartographic generalization, as a complex issue, is carried out using various approaches, which can be divided into condition-action modelling, human interaction modelling and constraint-based modelling (L. Harrie, R. Weibel 2007). This work uses a condition-action approach. It allows to directly apply mathematical mechanisms for determining decision rules, existing in rough logics, for the generalization (selection) of geographical information. Currently widely used expert systems based on “IF... THEN...” rules are based on classic rules defined by experts. This article proposes a method of extracting rules based on data, using rough logics.

Classical logic used on a daily basis based on the Aristotelian binary system 0-1 true-false (N. Adamiak 1979) is sometimes insufficient. This is especially the case when there is inconsistent or internally contradictory information, which often occurs when using real data (including spatial data). In response to the need for formal logical operations on such data, many non-classical systems were created (e.g. J. Łukasiewicz 1958, L.A. Zadeh in 1965), e.g. **rough logic** (Z. Pawlak 1982), of which three different types were used in this article. Rough logics provide, among others, the possibility to create decision rules.

Therefore, the author of the article posed the following **research questions**:

- Can rough logic be used to create object selection rules for cartographic generalization?
- What is the quality of selection made using these rules in relation to real topographic data?
- Which of the presented methods works best and what are the limitations of each of them?

2. Rough logics and decision rules

2.1. Rough logics

The article uses rough logic and rough set theory, the foundations of which were created by Professor Z. Pawlak in the 1970s (Z. Pawlak 1982, 1991, Z. Pawlak et al. 1995). Rough Set Theory (RST) assumes, unlike Classical Set Theory, that there can be three (and not two as in classical theory) states of an object:

- An object can, with certainty, belong to the set – it is then in its lower approximation defined as:

$$\underline{P}(X) = \bigcup_{x \in U} \{P(x) : P(x) \subseteq X\}$$

- The object can, with certainty, not belong to the set – it is then outside its upper approximation defined as:

$$\overline{P}(X) = \bigcup_{x \in U} \{P(x) : P(x) \cap X \neq \emptyset\}$$

- The object can belong to the set or not – it is then outside its lower, but inside the upper approximation – within the boundary of the set described by the formula:

$$PN(X) = \overline{P}(X) - \underline{P}(X)$$

In rough logic, the information system is most often represented in the form of a table, whose rows correspond to individual objects, and whose columns correspond to the attributes describing objects. One of the attributes can be distinguished as a decision attribute. Depending on the rough logic type, attributes in the following measurement scales may appear in the attribute table:

- RST – Rough Set Theory – attributes are nominal, their values can differentiate objects, but there is no specific order between them. A special case is the *boolean* attribute, which can have only two values (usually 0 or 1). RST is based on the relationship of indistinguishability between objects (Z. Pawlak 1982, Z. Pawlak et al. 1995);
- DRST – Dominance Based Rough Set Theory – attributes are expressed in an ordinal scale, which ensures a fixed order of the attribute values. DRST is based on the domination relation of one object to another, based on the established order of attribute values (S. Greco et al. 2001, R. Słowiński et.al. 2014);
- FRST – Fuzzy-Rough Set Theory – attributes are expressed numerically (integers or floating-point numbers). It is possible to determine not only the order, but also the distance between the individual attribute values. FRST uses a (non)similarity relation which, in contrast to the binary indistinguishability relation, takes values in the range <0, 1> (D. Dubois, H. Prade 1990; C. Cornelis et al. 2008).

The decision attribute is expressed in a nominal (RST and FRST) or ordinal (DRST) scale. The decision attribute is often binary. This is also the case with this research: 1 – an object selected during selection, 0 – an object not selected. Attempts are also being made for the decision attribute which is a continuous numeric value. However, this requires a change (adaptation) of the existing methodology, e.g. FRST (A. Fiedukowicz 2015a).

2.2. Decision rules

Rough logic allows to create decision rules such as "IF {conditions based on attributes} THEN {decision attribute value}" based on an existing data set that contains a decision attribute. Unlike traditional condition-action approaches, approximate systems can extract rules even from data that is contradictory, e.g. in the case of different values of the decision attribute for objects having the same values of other attributes. Then, two types of rules are created: certain rules – based on data that do not contain internal contradictions, and rough rules – created based on cases from the information system, which are in conflict with other examples (Z. Pawlak 1991). To create such rules, this research uses attributes directly from the topographic database structure (first for model data, then for BDOT10k) as well as the attributes added to this database, which represent the spatial context of the generalised objects. The way they are created and selected is described in more detail in the “ISPRS International Journal of Geo-Information” (A. Fiedukowicz 2020).

3. Methodology and source data

3.1. Research plan

The research presented in this article is part of a study which is the subject of the author’s doctoral dissertation (A. Fiedukowicz 2017). This research involved the following steps, of which those included in this article are in bold:

1. Data enrichment with attributes describing the spatial context
2. Determining significant attributes through reducts (using rough logics)
3. **Determining and analysing decision rules using rough logics**
4. **Applying these rules to the selection of topographic objects**
5. **Evaluation of the selection results**

Points 1 and 2 have been described in the author’s article (A. Fiedukowicz 2020).

3.2. Source data

The research was based on topographic data covering basic classes of objects, such as roads, buildings, and river networks. Classes were chosen that varied both in terms of geometric representation (lines, polygons) and type of objects (natural, anthropogenic).

First, research was conducted on model data prepared for this purpose (the data is described in detail in the works of A. Fiedukowicz 2017, 2020). This data was created on the basis of an analysis of many European topographic databases (A. Fiedukowicz 2017) so that they were as universal as possible and at the same time accurately reflected the structures of attributes in these databases. The advantages

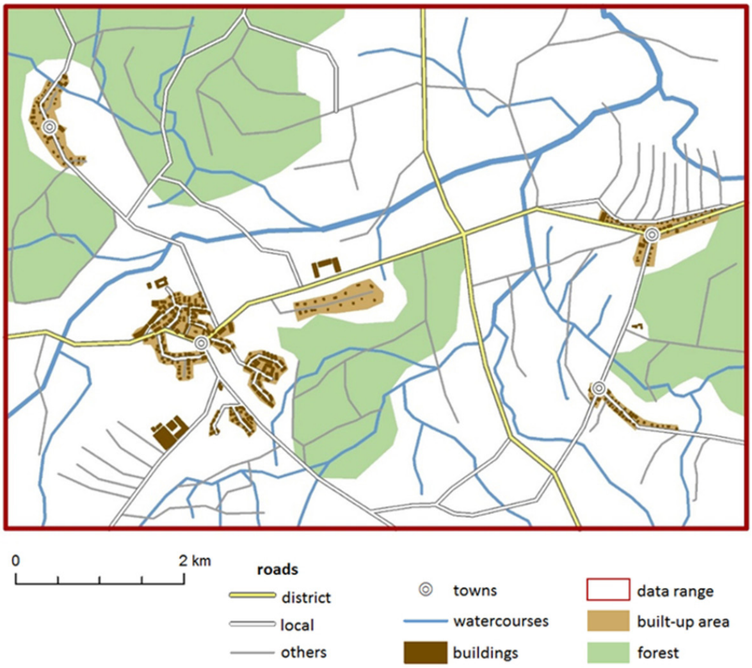

Fig. 1. Model data at the LoD10k (source: A. Fiedukowicz 2020)

of conducting research using model data are discussed in the dissertation by A. Fiedukowicz (2017). Data was analysed at two scale levels:

- 1:10,000 (LoD10k) – data generalised to 1:50,000 (LoD50k), an area of 9.1 × 6.4 km, generalised object classes: buildings, roads, watercourses (fig. 1);
- 1:50,000 – data generalised to 1:250,000 (LoD250k), an area of 24.6 × 22.2 km, generalised object classes: roads, watercourses (buildings were omitted in LoD50k→ LoD250k generalization).

The real data used for research at the LoD10k level came from BDOT10k (Topographic Object Database), which is the basic topographic database in Poland and covers the entire country. The analysed object classes (buildings – BUBD, roads – SKDR, watercourses – SWRS) have attributes similar to the previously analysed model data. However, the larger size of the set, greater complexity and the possibility of unexpected situations or missing values in the data allow to check how the methodology developed on the model data works in relation to real data. The small town of Chocianów and its surroundings, located in Lower Silesia, was chosen as the test area. It gave the opportunity to test methods both in typical rural areas (consisting mainly of farmsteads), as well as in urban areas with more compact, diverse buildings and a system of streets. For this reason, two sub-areas were distinguished: "Town" and "Village". In addition, for sections of watercourses, due to the relatively small number of data for the Chocianów area, the area from the vicinity of the town of Sieniawa was used (fig. 2).

The following object classes were chosen for generalization:

- roads (SKDR) from LoD10k to LoD50k and from LoD50k to LoD250k,
- buildings (BUBD) from LoD10k to LoD50k,
- rivers (SWRS) from LoD50k to LoD250k.

BDOT10k attributes were also enriched with a number of attributes describing the spatial context resulting from the geometry of a given class of objects (geometric attributes) as well as from the neighbouring objects of other classes (relational attributes) (A. Fiedukowicz 2017, 2020).

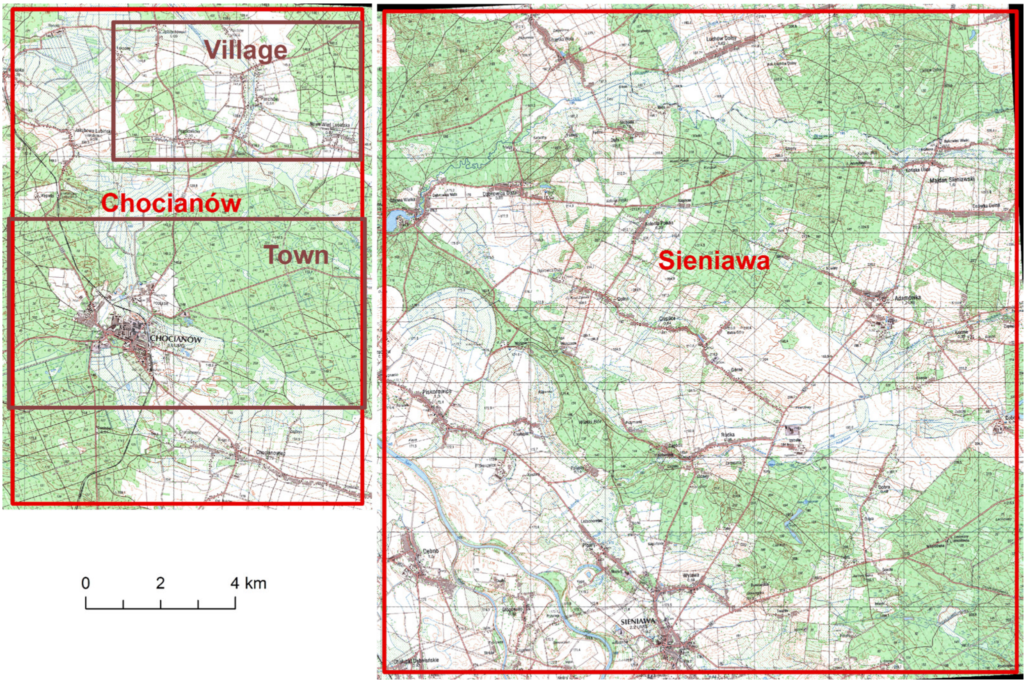

Fig. 2. 1:50,000 analog maps − the reference material with marked research areas (source: A. Fiedukowicz 2017)

The LoD50k level used for generalization was obtained by expert (manual) selection of objects from BDOT10k (supported by existing maps of this scale). Both model and real data have been enriched with geometric and relational attributes representing the spatial context (A. Fiedukowicz 2017, 2020). All objects have also been expertly assigned a decision attribute with the value:

- 1 – if the object was to be selected for the next scale level,
- 0 – if the object was not to be selected.

In the case of real data, the value of the decision attribute was determined based on existing cartographic materials, such as 1:50,000 map scans and BDOO (General Geographic Database – LoD250k). The task of the computed rules was to correctly predict the value of the decision attribute based on the other attributes.

3.3. Detailed research plan

The adopted research methodology was different in relation to model data and real data (fig. 3). First of all, for model data (due to the small size of the set and the preliminary nature of research work at the time), the method of cross-validation (M.W. Browne 2000) and four iterations were used, and the determination of the training and test set was not spatial. In the case of real data, the training and test sets were determined once, by means of a spatial division into two sub-areas. This method of division better reflects real conditions in which the proposed methodology could be used. That is because, in real applications, model training takes place only in a small part of the area, where expert decisions are made, and then the rules can be applied in areas for which the decision is not known. The division into test and training sets was made in such a way as to ensure a comparable share of both decision classes in them. Attempts were also made to maintain the training set at around ⅔ and the test set at around ⅓ of the entire set. The difference in determining learning and test sets for real and model data is illustrated in figure 4.

3.4. Rules assessment

The rules determined for model data were each time evaluated in terms of:

- The numbers of rules
- The number of conditions for each rule
- Support for rules referred to as the number of examples of a training set that meet all the conditions of a rule. Relative support, i.e. the

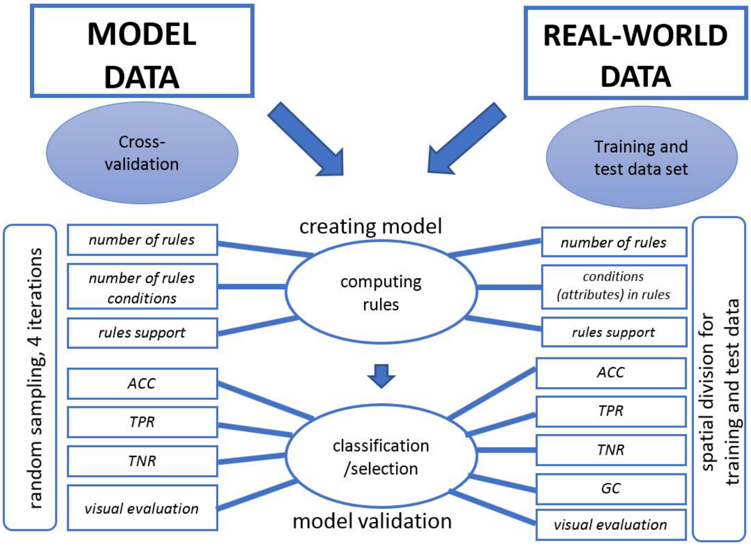

Fig. 3. Scheme of research on model and real data (source: A. Fiedukowicz 2017)

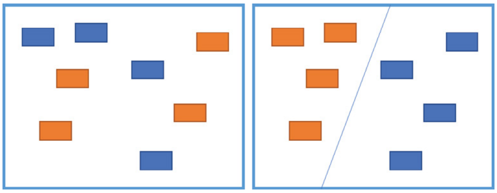

Fig. 4. Division into training and test sets. On the left − a random division used for model data, on the right − a spatial division used for actual data (source: A. Fiedukowicz 2017)

ratio of the number of objects supporting the rule, to all objects of the training set (or to all objects of the training set belonging to the same decision class) can also be considered as support.

The assumption was that simpler models with a fewer number of less complex rules are more desirable because it is much easier to understand the knowledge they represent.

For real data, the assumption was the same, but due to the much greater complexity of the data, and hence the rules as well, it was not possible to analyse the entire decision system. However, selected rules were evaluated in terms of knowledge extracted from the data, i.e. the conditions set in the rules. For RST and DRST rules, the support for each rule was additionally defined. As a result, it was possible to determine the rules with the highest support and their detailed analysis (what was done for real data). For the FRST method, support for rules is not explicitly defined, which is why the number of rule uses on the test set was used analogously to the support (the author's original approach – A. Fiedukowicz 2017).

3.5. The evaluation of results

The effectiveness of the computed rules was checked based on the test set by selecting objects. As the selection can be understood as a specific (binary) type of classification, a confusion matrix was used to evaluate it (T. Fawcett 2006). Based on the confusion matrix it is possible to count several indicators of the classification quality. The following were used:

• Accuracy determines the number of correctly classified examples in relation to all examples:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

• True Positive Rate

• True Negative Rate

• The Gini Coefficient (based on the ROC curve, fig. 5 − Receiver Operating Characteristic):

$$GC = 2 * AUC - 1$$

GC = 1 – the ideal classifier; here: all objects selected or not, as expected,

GC = 0.5 – random classifier, giving results analogous to class randomization,

GC = 0 – reverse classifier; here: all objects selected or not, contrary to predictions.

The use of the GC factor allows to easily compare binary classifiers, assigning them values from 0 to 1 (fig. 5).

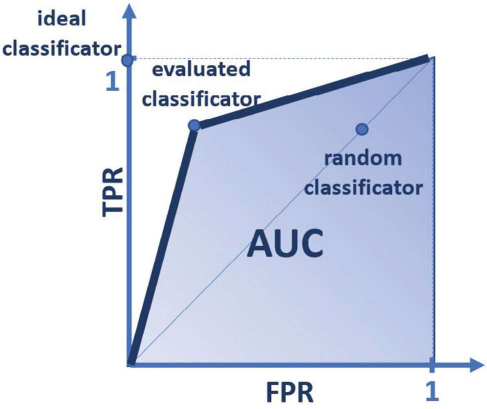

Fig. 5. ROC curve (source: A. Fiedukowicz 2017)

At the same time, due to the spatial nature of the data, quantitative analysis of the results is far from sufficient. Therefore, also visual analysis was performed to detect spatial patterns of correctly/incorrectly classified objects (fig. 6). This type of analysis can sometimes detect patterns which are not visible in quantitative data and make adjustments to the model, e.g. by introducing to the model an additional attribute related to the observed spatial relationship (in the case presented in figure 6 it could be e.g. distance from the forest).

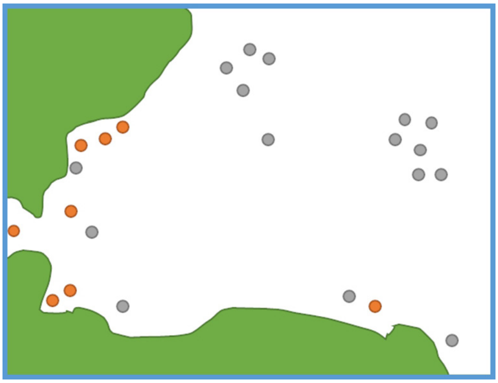

Fig. 6. An example of a visual analysis of classification results: grey – objects classified correctly, orange – objects classified incorrectly. This shows the potential impact of the proximity of another object (here: a forest) on the correctness of classification (source: A. Fiedukowicz 2017)

Visual analysis can also assess whether classification errors are in a given case key from a cartographic point of view or are acceptable. For spatial data, an information table constituting a classic data model in rough theories, is not their full representation.

3.6. Software and algorithms

In order to carry out the research, software (often with the author's modifications) was used to calculate decision rules and predict the value of decisions based on them. The following were used:

- for the Rough Set Theory (RST) – the R language dedicated to statistical calculations (including the RoughSets package),
- for the Dominance Based Rough Set Theory (DRST) – the jMAF program,
- For the Fuzzy-Rough Set Theory (FRST) – the R language dedicated to statistical calculations (including the RoughSets package).

Both programs used (R and jMAF) are available free of charge for non-commercial purposes, and the R environment also for commercial purposes (GNU license). Both jMAF and the RoughSets package were developed by academic environment as implementations of mathematical theories, allowing a wider group of people to apply them in practice.

To determine rules using the RST method, two LEM and CN2 algorithms were used, which have been described in detail by the author (M. Fiedukowicz 2017).

4. Results

A detailed analysis of the obtained results and illustrations can be found in the author's doctoral dissertation (M. Fiedukowicz 2017). The article presents only a summary of quantitative results and selected examples of qualitative analysis along with a description.

4.1. Model data

In terms of the number of rules, the DRST method is best when model data is used (tab. 1).

Table 1. Averages of minimum and maximum numbers most favour and rule lengths for all model data (the smallest, able values are in bold) (source: A. Fiedukowicz 2017)

method	min. no.	max. no.	min. length	max. length
RST CN2	8,0	9,8	**1,0**	**2,0**
RST LEM2	8,8	10,2	1,0	3,4
DRST	**6,2**	**8,8**	1,0	3,2
FRST	64,6	79,4	1,4	10,4

In this method, the average maximum number of rules does not exceed 9 (the maximum number of rules obtained by this method for a set of watercourses in 1:50,000 detail is 18). As for the length of rules, the best results are obtained using the CN2 algorithm in the RST method – in most cases the number of rule conditions does not exceed two (only one exception was noted, where there were three-element rules). The most complex decision systems (both in terms of the number of rules and their length) were obtained using the FRST method. The advantage of this method is the possibility of using continuous data, without the need to downgrade the measuring scale.

As for the effectiveness of applying the created rules on the test set (determined four

Table 2. Averaged results (using a cross-validation) for selecting model data from a scale level of 1:10,000 to a level of 1: 50,000 (the method/algorithm is listed in the first column); red indicates unacceptable results (source: A. Fiedukowicz 2017)

10k	buildings	rivers	roads
ACC CN2	0,98	0,98	0,95
TPR CN2	**0,81**	**0,97**	**0,95**
TNR CN2	0,99	0,98	0,95
ACC LEM2	0,97	0,97	0,97
TPR LEM2	**0,46**	**0,92**	**0,93**
TNR LEM2	1,00	1,00	1,00
ACC DRST	0,98	0,98	0,96
TPR DRST	**0,76**	**0,95**	**0,91**
TNR DRST	0,99	1,00	0,99
ACC FRST	0,78	0,98	0,97
TPR FRST	**0,39**	**1,00**	**1,00**
TNR FRST	0,80	0,97	0,95

Table 3. Averaged results (using a cross-validation) for selecting model data from LoD50k to LoD250k (the method/algorithm is listed in the first column); red indicates unacceptable results (source: A. Fiedukowicz 2017)

50k	rivers	roads
ACC CN2	0,92	1,00
TPR CN2	**0,89**	**0,99**
TNR CN2	0,93	1,00
ACC LEM2	0,91	1,00
TPR LEM2	**0,86**	**0,99**
TNR LEM2	0,94	1,00
ACC DRST	0,93	1,00
TPR DRST	**0,93**	**1,00**
TNR DRST	0,93	1,00
ACC FRST	0,65	0,96
TPR FRST	**0,92**	**0,89**
TNR FRST	0,48	1,00

times as part of cross-validation), it is very high for model data. The accuracy for all methods except the FRST method in all cases is above 90%, sometimes even 100% (especially for the generalisation of LoD50k-> LoD250k (tables 2, 3). At the same time, there is reduced sensitivity (TPR) for building selection using the FRST and RST methods with the LEM2 algorithm (tab. 2) and True Negative Rate (TNR) for the selection of LoD50k→250k watercourses using the FRST method.

Given the very good results obtained for the model data, it was decided to apply the described methodology to real data. However, even in the case of model data, the FRST method has the lowest potential both in terms of the complexity of the decision system and its effectiveness (quality of classification).

4.2. Real data

The presented tests carried out for various classes of objects using the three tested rough methods allowed to draw conclusions regarding the quality of the performed classification and the factors affecting it. Two variants have been distinguished for buildings: B – selected (k50=1) buildings represented in LoD50k as individual buildings, BZ – also selected buildings represented in LoD50k as a fragment of a built-up area.

The summarized results for real data are presented in table 4. Table 5 presents the values of the Gini Coefficient (GC) for various methods and areas as well as the average and median of this indicator for individual methods. It was also counted (tab. 5) how many times each method achieved first and second place in a specific version of the experiment (for specific data, area, variant, etc.) according to the Gini Coefficient.

It is worth noting (tab. 5) that the average GC for the RST methods with the CN2 algorithm and the DRST method is the same, and the median is only slightly higher for the DRST method. Similarly, each of these two methods was the best four times in some variant of the

Table 4. Collected results for real data (source: A. Fiedukowicz 2017)

	buildings, v. BB, area Town	buildings, v. B, area Town	buildings, v. BB, area Village	buildings, v. BB, area Village	rivers, area Chocianów	rivers, area Sieniawa	roads, LoD10k -> LoD50k	roads, LoD50k -> LoD250k
CN2	**0,82**	**0,74**	**0,60**	**0,68**	**0,90**	**0,66**	**0,72**	**1,00**
CN2_TPR	0,89	0,39	0,50	0,45	0,79	0,42	0,73	1,00
CN2_TNR	0,21	0,81	0,76	0,82	0,95	0,81	0,69	1,00
LEM2	**0,87**	**0,80**	**0,64**	**0,65**	**0,88**	**0,87**	**0,74**	**0,84**
LEM2_TPR	0,95	0,16	0,71	0,28	0,58	0,93	0,86	0,15
LEM2_TNR	0,18	0,93	0,53	0,87	1,00	0,83	0,31	1,00
DRST	**0,83**	**0,85**	**0,67**	**0,72**	**0,94**	**0,88**	**0,62**	**0,88**
DRST_TPR	0,89	0,09	0,74	0,56	0,79	0,88	0,57	0,37
DRST_TNR	0,28	1,00	0,56	0,71	1,00	0,89	0,82	1,00
FRST	**0,91**	**0,84**	**0,61**	**0,59**	**0,91**	**0,69**	**0,76**	**0,81**
FRST_TPR	1,00	0,00	0,99	0,17	0,70	0,96	0,90	0,02
FRST_TNR	0,03	1,00	0,01	0,75	1,00	0,52	0,27	0,99

Table 5. The Gini coefficient for results using real data together with its average and median as well as the first (yellow) and second (blue) place achieved by particular methods (source: A. Fiedukowicz 2017)

	buildings, v. BB, area Town	buildings, v. B, area Town	buildings, v. BB, area Village	buildings, v. BB, area Village	rivers, area Chocianów	rivers, area Sieniawa	roads, LoD10k -> LoD50k	roads, LoD50k -> LoD250k	mean	median	the best (count)	second best (count)
RST_CN2	0,71	1,00	0,55	0,60	0,63	0,64	0,87	0,61	**0,70**	**0,63**	**4**	**2**
RST_LEM2	0,58	0,58	0,56	0,54	0,62	0,57	0,79	0,88	**0,64**	**0,58**	1	2
DRST	0,69	0,68	0,59	0,54	0,65	0,63	0,89	0,88	**0,70**	**0,67**	**4**	**4**
FRST	0,59	0,51	0,51	0,50	0,50	0,46	0,85	0,74	**0,58**	**0,51**	0	1

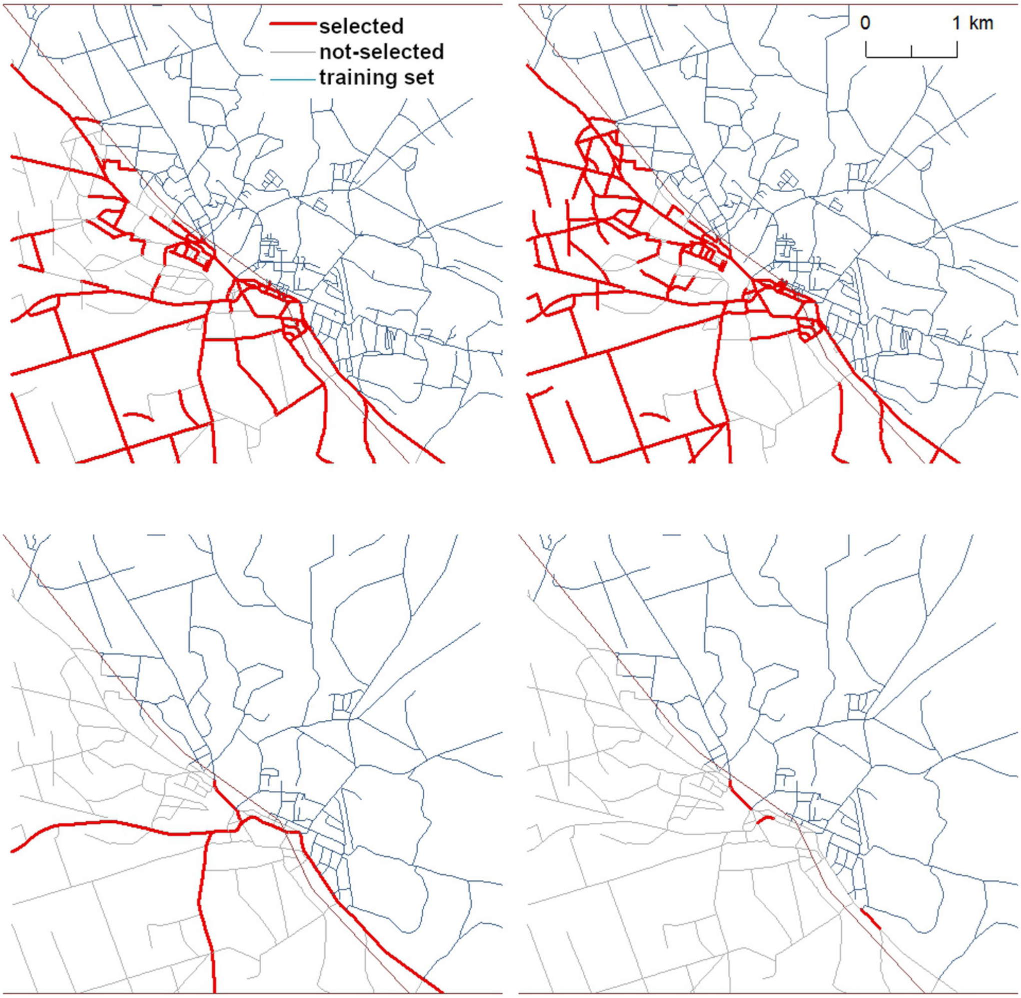

Fig. 7. The results for the road sections obtained by the RST method in the area of Chocianów: on the left − the CN2 algorithm, on the right − the LEM2 algorithm; at the top − 10k→50k selection, at the bottom − 50k→250k selection; the selected objects are marked in red (source: A. Fiedukowicz 2017)

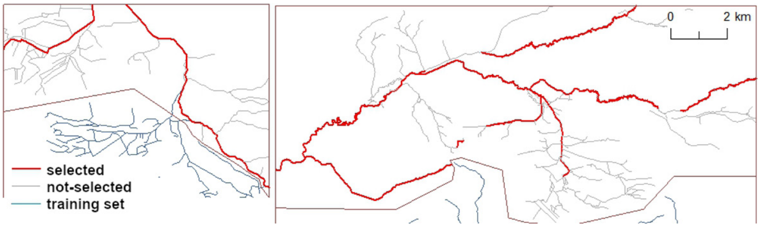

Fig. 8. Results for watercourse sections obtained by the DRST method: on the left − the Chocianów area, on the right − the Sieniawa area; the selected objects are marked in red (source: A. Fiedukowicz 2017)

experiment, while the DRST method was also four times in second place (compared to two cases of the RST method with the CN2 algorithm). This means that the DRST method was the best or second best in each of the experiment variants. Of the methods tested, these two: RST methods with CN2 and DRST algorithms give the best results (with a slight advantage of the DRST method). The worst results by far were obtained for the FRST method.

Quantitative results, however important, do not allow a full evaluation of the results of generalization. The performed qualitative visual analysis indicates shortcomings in many cases concerning in particular the continuity of the road network and river network (fig. 7, 8). The results containing discontinuities of these systems are obviously incorrect from a cartographic point of view. However, the developed methodology can still be used to support the cartographer's work in the tedious, complicated and time-consuming generalization process. On the other hand, in its present form, especially for network objects, it is not suitable to be used as the only and final way of selecting topographic objects. Cartographic editing of pre-selected topographic objects requires both further actions implemented in so-called post-processing and manual editing.

Regarding the complexity of decision rules (examples of rules in table 6), the trends observed in model data were confirmed:

• The simplest in reception and interpretation were the RST rules (especially those created by the CN2 algorithm).

• The rules of the FRST method were the most complex and illegible. It would be advisable to round the values of the attributes, which could partially improve the readability of the rules and increase their level of generalisation.

• The rules of the DRST method were characterised by average complexity. However, due to the characteristic way of their formation, they can be easily simplified (as discussed below).

5. Discussion

Non-classical approaches seem to be an attractive solution from the point of view of automating the generalization process. This is because their unusual, ambiguous nature, using rough or fuzzy information that is typical of cartographic knowledge difficult to formalize in a traditional way. Fuzzy logic together with linguistic variables have been used in the generalisation process (e.g. R. Olszewski 2009). However, the author's preliminary research (A. Fiedukowicz 2013a) showed difficulty in constructing fuzzy rules with a large number of available attributes (including native, relational and geometric attributes). Thus, the initial stages of the study included the possibility of choosing relevant attributes (reducts) using the Rough Set Theory (RST), and then constructing a fuzzy rule system based only on previously selected attributes. A significant obstacle, however, was the inconsistency of the measuring scales used in both methods, which meant that the quality of the generalisation performed in this way was not satisfactory.

Therefore, the next approach used **rough rules,** based on three non-classical logics (approaches, set theories): Rough Set Theory, Dominance-Based Rough Set Theory, Fuzzy-Rough Set Theory. First, they were used to se-

Table 6. Examples of rules for the selection of buildings (source: own work, based on A. Fiedukowicz 2017)

method	supp	rule
RST CN2	8%	IF function is residential and touches_another_building is 1 and area is small THEN k50 is 1
RST LEM2	22%	IF density_in_100m is high THEN k50 is 1
DRST	31%	(touches_another_building >= 1) & (distance_from_centrum_inversed >= small) => (K50 >= 1)
FRST	–	IF function is ~religous and floors~2 and density_in_300m ~513.879 THEN k50 = 1

lect interesting attributes (reducts), including attributes describing the spatial context (A. Fiedukowicz 2020), and then rules were constructed in measurement scales adapted to earlier reducts.

Rules based on rough sets were not previously widely **used in the generalization** of geographical information. J. Zhang (2001) conducted work on this subject, and the author also did some preliminary research (A. Fiedukowicz 2013b, 2015b). The theory of rough sets was used more often in other aspects of spatial data, especially in the non-classical definition of area boundaries (e.g. T. Beaubouef, F.E. Petry 2010).

As part of the research work, some interesting methodological issues were also noticed, the solution of which could help in the application of rough logics, not only in the generalization of geographical information. One of such issues is the construction of rules in the DRST method. As this method assumes a monotonic relationship between attributes and the decision (if the condition $\geq$ occurs on the left and on the right of the rule), the way to artificially determine the inverse relationships is to use inverted attributes (i.e. with changed monotonicity compared to the original). However, this causes interpretation difficulties related to reading such rules. They could be avoided by secondary inverting of the less than and greater than symbols in conditions of inverted attributes, e.g.

$$(pow\_d \geq 2) \leftrightarrow (pow\_d \leq 2)$$

This approach would simplify the resulting rules, and sometimes even aggregate the conditions for the attribute in the basic and inverted versions, if they appear in the same rule, e.g.

$$(odl\_msc \geq 2)\ \&\ (odl\_msc\_d \geq 3) \leftrightarrow$$
$$(odl\_msc \geq 2)\ \&\ (odl\_msc \leq 3) \leftrightarrow (2 \leq odl\_msc \leq 3)$$

In this way, rules would be easier to both read and interpret. It would be possible, also when creating reducts, to just write each attribute only once, regardless of whether it would appear in the reduct in the basic version, inverted version, or in both. The proposed solutions have the advantage of being relatively easy to automate.

Furthermore, a way for determining support for rules in the FRST method should be proposed. With the proposed averaging of the value of many objects, this could be the number of objects used to average the rule conditions, which would correspond to the traditional definition of support. It would also be possible to define the support of the rule in a fuzzy manner, taking into account the distribution of the values of individual attributes around the obtained average. For example, if 10 objects used to create a rule would have exactly the same attribute value (equal to the average used in the rule), then support would be 10. The more diverse the values of the attribute around the calculated average, the lower the support is. Such a mechanism for determining support could also be helpful when creating and reducing FRST rules. Unfortunately, the number of rule uses in this paper for a given rule in the training set do not fulfil the role of a criterion for evaluating the rules well enough. It is not known whether their use was correct (or if it allowed to determine the appropriate decision class for the rule). This leads to a situation where the number of times the rule is used is more than the given decision class in the test set (which means the rule is being used incorrectly). Therefore, it cannot be concluded that the greater the number of uses of the rule, the better its quality. This number cannot be directly compared with the support value determined for the other methods.

Similarly, for the FRST method, it seems reasonable to try to use fuzzy and classic relationship of indistinguishability together. Thanks to this, the original nature of the data could be maintained: continuous for continuous data (fuzzy similarity relationship), nominal for nominal data (classic indistinguishability relationship). Only an ordinal relationship cannot be represented in this method.

It should be remembered that the obtained results cannot be evaluated in isolation not only from the mathematical approach used, but also from the algorithms used and their implementation. While mathematical theories are strict (and deterministic), the algorithms used in the research are most often heuristic algorithms, and therefore do not necessarily allow to determine optimal solutions. Moreover, the results of the experiment, and in particular the effectiveness of algorithms, is significantly influenced by the manner and correctness of implementation in the used software. The ease and intuitiveness of using this software are

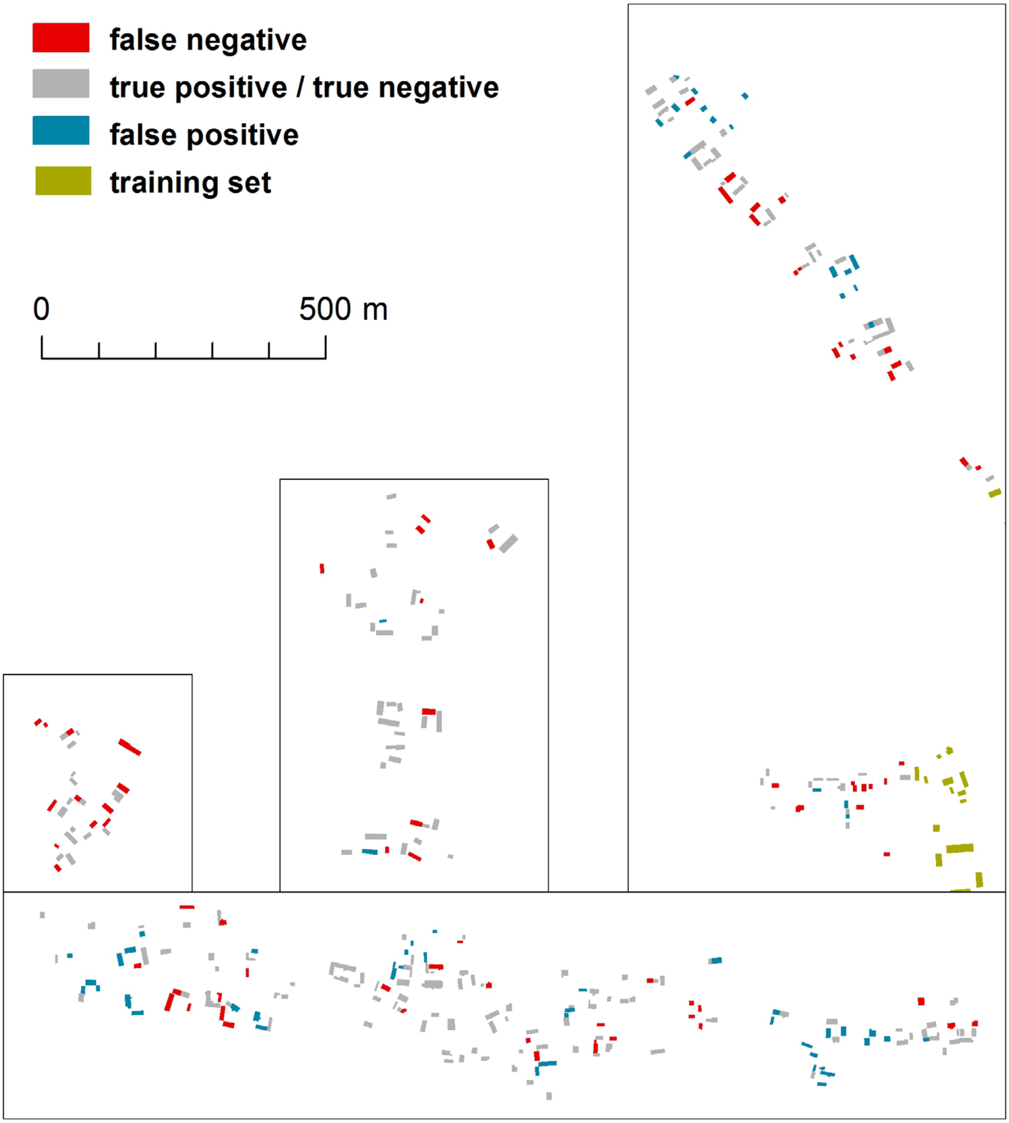

Fig. 9. Results of the selection of buildings using the FRST method in relation to the expected value; Variant B – parts of the “Village” area containing buildings (source: A. Fiedukowicz 2017)

also significant factors. The lack of a graphical user interface when using the R language can be an obstacle to the use of this tool by a wider group of people without programming skills. On the other hand, the open nature of the R language provides great flexibility both at the level of script creation and in the modification of existing packages. In turn, the GUI available in jMAF is quite easy to use, but both the input formats and the way the output is combined are strictly defined and cannot be modified.

6. Conclusions

Rough logic can be used to select objects when generalizing geographical information. The conducted research indicates the advantage of the DRST and RST approaches (with the CN2 algorithm) due to the quality of the obtained selection, the greater simplicity of the decision system, and better refined IT tools enabling the use of these systems. RST CN2 and DRST approaches obtained an average accuracy of actual data classification (measured using the Gini coefficient) of 0.70.

At this stage, the FRST approach, which is characterised by the highest complexity of created rules and the worst selection results, is not recommended. The average accuracy of the actual data classification determined for this method is 0.58. In addition, the result obtained using the FRST method was not the best of the results for neither any object class nor the test area. This method achieved second place just once. In the case of this method, not only the approach itself but also the quality of implementation could have an impact on the obtained results – the functions of the R package regarding the FRST method are rarely used.

Individual approaches are limited by the need to select appropriate measurement scales for the attributes used in them. From this point of view, the DRST method seems to be the most promising. It allows the use of continuous numerical attributes after discretization without losing information about the order of classes, attributes on the ordinal scale – directly, binary attributes – treating them as attributes on the ordinal scale. Only attributes on the nominal scale must be omitted or artificially ordered.

Particular attention should be paid to the selection of network objects, in which the use of only a condition-action approach, without maintaining consistency of the network, may not produce the desired results. Therefore, when evaluating the quality of the obtained results, it is necessary, in addition to quantitative indicators, to use expert visual evaluation.

Unlike approaches based on classical logic, rough approaches allow the use of incomplete or contradictory information. The proposed tools can (in their current form) find an auxiliary use in the selection of topographic objects, and also in other generalisation operators.

Literature

Adamiak N., 1979, *Logika*. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

Beaubouef T., Petry F.E., 2010, *Fuzzy and rough set approaches for uncertainty in spatial data*. Berlin – Heidelberg: Springer, pp. 103–129.

Browne, M.W., 2000, *Cross-validation methods*. “Journal of Mathematical Psychology” Vol. 44, no. 1, pp. 108–132.

Cornelis C., Martín G.H., Jensen R., Ślęzak D., 2008, *Feature selection with fuzzy decision reducts*. In: *Proceedings of the International Conference on Rough Sets and Knowledge Technology*, Chengdu, China, 17–18 May 2008, Berlin – Heidelberg: Springer, pp. 284–291.

Dubois D., Prade H., 1990, *Rough fuzzy sets and fuzzy rough sets*. “Intern. Journal of General Systems” Vol, 17, no. 2/3, pp. 191–209.

Fawcett T., 2006, *An introduction to ROC analysis*. “Pattern Recognition Letters” Vol. 27, no. 8, pp. 861–874.

Fiedukowicz A., 2013a, *Construction of fuzzy interference system for generalization of geographic information – selection of roads segments*. In: “Geoinformatica Polonica” Vol. 12, pp. 53–62.

Fiedukowicz A., 2013b, *Wykorzystanie zbiorów przybliżonych do pozyskiwania wiedzy i budowy reguł systemu generalizacji informacji geograficznej*. „Roczniki Geomatyki” T. 11, nr 2(59), pp. 33–46.

Fiedukowicz A., 2015a, *Fuzzy rough sets theory reducts for quantitative decisions – Approach for spatial data generalization*. In: *Pattern Recognition and Machine Intelligence. Proceedings*. Eds. M. Kryszkiewicz end al. „Lecture Notes in Computer Science” Vol. 9124, pp. 314–323.

Fiedukowicz A., 2015b, *Redukcja wymiarowości problemu – ograniczenie liczby cech*. In: *Wybrane metody eksploracyjnej analizy danych przestrzennych (Spatial Data Mining)*. Eds. A. Fiedukowicz, J. Gąsiorowski, R. Olszewski. Warszawa: Wydział Geodezji i Kartografii Politechniki Warszawskiej.

Fiedukowicz A., 2017, *Metodyka wykorzystania reduktów i reguł przybliżonych w procesie generalizacji informacji geograficznej*. PhD. dissertation, Warsaw University of Technology, Faculty of Geodesy and Cartography.

Fiedukowicz A., 2020, *The role of spatial context information in the generalization of geographic information: Using reducts to indicate relevant attributes*. “ISPRS International Journal of Geo-Information” Vol. 9, no. 1, 37.

Greco S., Matarazzo B., Słowiński R., 2001, *Rough sets theory for multicriteria decision analysis*. “European Journal of Operational Research” Vol. 129, pp. 1–47.

Harrie L., Weibel R., 2007, *Modelling the overall process of generalization.* In: *Generalization of Geo-*

graphic Information. Amsterdam: Elsevier Science BV, pp. 67–87.

Łukasiewicz J., 1958, *Elementy logiki matematycznej*. Warszawa: Państwowe Wydawnictwo Naukowe.

Olszewski R., 2009, *Kartograficzne modelowanie rzeźby terenu metodami inteligencji obliczeniowej*. "Prace Naukowe Politechniki Warszawskiej. Geodezja" No. 46.

Pawlak Z., 1982, *Rough sets*. "Intern. Journal of Comput. Information Science" Vol. 11, no. 5, pp. 341–356.

Pawlak Z., 1991, *Rough sets: Theoretical aspects of reasoning about data*. Dordrecht: Kluwer Academic Publishing.

Pawlak Z., Grzymala-Busse J., Słowiński R., Ziarko W., 1995, *Rough sets*. "Communication of the ACM" Vol. 38, pp. 88–95.

Regnauld N., McMaster R.B., 2007, *A synoptic view of generalization operators*. In: *Generalisation of Geographic Information*. Amsterdam: Elsevier Science BV, pp. 37–66.

Roth R.E., Brewer C.A., Stryker M.S., 2011, *A typology of operators for maintaining legible map designs at multiple scales*. "Cartographic Perspective" Vol. 68, pp. 29–64.

Shea K.S., McMaster R.B., 1989, *Cartographic generalization in a digital environment: When and how to generalize*. In: *Proceedings of the Auto-Carto, Baltimore, MD, USA*, Vol. 9, pp. 56–67.

Słowiński R., Greco S., Matarazzo B., 2014, *Rough-set-based decision support*. In: *Search Methodologies*. Boston MA, pp. 557–609.

Zadeh L. A., 1965, *Fuzzy sets. "Information and Control"* Vol. 8, no. 3, pp. 338–353.

Zhang J., 2001, *Using rough set represent the uncertainty in GIS spatial data*. In: *Proceedings of ICA Conference Beijing*, China.