ANALYSIS OF PREFERENCE MAPS USING DATA MINING METHODS

ANALIZA PREFERENČNIH KART Z METODAMI ODKRIVANJA ZNANJA IZ PODATKOV

Lidija Breskvar Žaucer, Blaž Zupan, Mojca Golobič

UDK:316.653:528.94:711

Klasifikacija prispevka po COBISS-u: 1.01

ABSTRACT

Preference maps allow people to express their opinions about future spatial development in a simple way; they mark the areas that they consider suitable for specific activities on a cartographic base map. The decision on the area is generally intuitive and reflects their views and preferences regarding the solution of spatial problems. In this way, preference maps may hold valuable information and convey hidden knowledge. To make the best of their potential usefulness in spatial planning and to contribute to the transparency of the process, these information and knowledge should be explicit and presented in an interpretable form. In this paper, we report on a case study of preference maps analysis for municipality Komenda in Slovenia, where residents marked areas they considered especially valuable and which should therefore be preserved. In an attempt to discover why specific areas were marked for protection, we used the selected data mining approaches to infer the relations between preferential annotations and spatial characteristics. The inferred patterns were reported in the form of decision rules and in the graphical form of a nomogram. Interpretation of results shows that the methodology proposed in this paper and the explicit decision criteria and rules extracted by data mining can be useful for further applicability in spatial planning.

POVZETEK

Preferenčne karte omogočajo ljudem, da preprosto izrazijo stališča do prostorskega razvoja. Posamezniki neposredno na kartografsko podlogo zarišejo po njihovem primerna območja za izbrano dejavnost. Odločitev o območju je običajno intuitivna. Je rezultat njihove celostne zaznave prostora in iskanja prostorske rešitve. Za obrisi območij se tako skrivajo dragocene informacije in lokalno znanje. Da bi čim bolje izkoristili njihovo potencialno uporabnost pri načrtovanju posegov v prostor in hkrati zagotovili transparentnost postopkov, morajo biti te skrite informacije in znanje podani v eksplicitni obliki. Prispevek opisuje primer analize preferenčnih kart za občino Komenda, na katerih so občani označevali območja, ki se jim zdijo posebej dragocena taka, kot so, in bi jih zato ohranili. Da bi odkrili, zakaj so bila nekatera območja predlagana za ohranjanje, smo uporabili izbrane postopke odkrivanja znanja iz podatkov, s katerimi smo iskali povezave med preferenčnimi območji in prostorskimi značilnostmi. Odkrite zakonitosti so bile predstavljene v obliki odločitvenih pravil in nomograma. Interpretacija rezultatov prikazuje, da je v prispevku predlagana metodologija skupaj z eksplicitnimi odločitvenimi merili in pravili lahko primerna za nadaljnjo uporabo v prostorskem načrtovanju.

KEY WORDS

preference mapping, public participation, land use planning, knowledge extraction, inference of decision making rules

KLJUČNE BESEDE

preferenčno kartiranje, sodelovanje javnosti, prostorsko načrtovanje, pridobivanje znanja, odkrivanje odločitvenih pravil

1 INTRODUCTION

In these times of enhanced democratization of decision-making and, at the same time, of growing diversity and interlacement of interests, consensual spatial planning decisions are increasingly hard to achieve. One of the key steps in this direction is the due identification and acceptance of lay people's goals, interests and values. This has already become one of the main principles of decision-making in modern pluralistic society (Friedman, 1992). Interests and values can be included in the process of resolving conflicts and in building consensual planning proposals. Especially if they are clearly and unambiguously expressed, an improvement in the effectiveness of conflict resolution can be expected (Janssen et al., 2006; Martin et al., 2000).

People commonly use rules of thumb or heuristics when making judgments (Newell and Simon, 1972). According to Jankowski et al. (2001), those who are generally not directly involved in planning reach their decisions in an ad hoc way, without any formal analysis. They have difficulty expressing their own interests and values regarding planned activities in a rational and explicit way. It is therefore appropriate to enable the lay people participating in a planning process to express themselves in a way they are more familiar with. One of these ways is to enable them to express their criteria for planning implicitly by identifying the areas they depict on a cartographic base map.

Maps are generally acknowledged as an effective and practical tool for informative communication between experts and lay people (Ball, 2002; Carver, 2003; Jankowski et al., 2001; Polič et al., 2002; Soini, 2001). Marking out areas of interest enables lay people to express their wishes and needs in a simple, direct way and also allows them to be creative. Their preferences are thereby calibrated with their own rich knowledge of the area they live in (Carver, 2003; Jankowski et al., 2001). The incorporation of local knowledge that is not normally available via ordinary geographic information datasets is clearly a major strength of all participatory approaches (Carver, 2003). It is also known that what are called preference or cognitive maps are an important supplement to traditional, word-based opinion surveys, since they are based on awareness of the spatial implications of a decision problem (Carver, 2003; Linden and Sheehy, 2004) and thus provide a much better indication of whether and where conflicts regarding land use are expected (Golobič and Marušič, 2007).

There are many examples of the use of preference maps in planning processes (e.g. Aravot, 1996; Bartol et al., 1998; Harris and Weiner, 1998; Jankowski et al., 2001; Kingston et al., 2000; Macnab, 1998; McClure, 1997; Polič et al., 1991; Sanoff, 1991). Planners usually include findings resulting from their analyses directly in an intuitive-holistic planning process (e.g. Harris and Weiner, 1998; Macnab, 1998; Polič et al., 1991). The analysis of such maps is therefore manual and runs the risk of overlooking relationships that would otherwise be discovered in some systematic, computer-supported data analysis. The results of such analysis are frequently not clearly presented. They remain in the domain of the planner, making the subsequent process of developing consensual planning proposals less transparent and harder to substantiate.

In this paper we propose a methodology that uses data mining to explicitly infer the relationships between preferential knowledge and spatial characteristics. In this way, we can uncover rules that can provide the basis for argued discussions between planners and lay people, allowing the lay people to better understand the planning decisions and the planners to compare their views with the reasoning of other people involved in the planning process. Because the rules are inferred from available data and they relate planning preferences to an available set of spatial characteristics, the methodology should be regarded as providing a vehicle for argued communication rather than revealing the deeper, tacit knowledge of individuals that may also be based on their living habits, experiences, emotional relations with the environment, interests and wishes (Aravot, 1996; Golobič and Marušič, 2007). Inferred lay people's preferential rules that rely on combinations of spatial characteristics can then be compared to experts' criteria and used to complement the spatial evaluation process, such as the commonly used multi-criteria spatial evaluation (Carver, 1991; Pettit in Pullar, 1999; Voogd, 1983).

Reports on attempts to formalize the interpretation of preference maps are at best sparse (Aravot, 1996; Bartol et al., 1998; Golobič and Marušič, 2007; Jankowski et al., 2001). Bartol et al. (1998) used a simple cross-tabulation analysis. The result of the method is a description of the proposed area's spatial characteristics in the form of a list of spatial variables and their values, while their relative importance remains unknown. Our proposed approach is similar to that taken by Aravot (1996) and Golobič and Marušič (2007) which employed classical statistical regression methods but resulted in a more mathematical model, the findings of which are more difficult to use in communication with lay people. Our primary intention was to use techniques that can directly infer rules that are easy to interpret, evaluate the predictive power of an approach using standard procedures from supervised data mining (Witten and Frank, 2000) and, at the same time, identify interesting data subgroups with their own specific spatial characterizations. Similar approaches have already been used to address other spatial and environmental issues (e.g. Bui et al., 2005; Kobler and Adamič, 2000; Naderi and Raman, 2005; Ogris, 2007; Zhang et al., 2005). In the spatial planning field these methods have been tested for reducing the complexity of a multi-criteria suitability analysis by discovery and further consideration of only the most significant criteria (Jankowski et al., 2001).

2 MATERIALS AND METHODS

The study reported in this paper employed data collected through a public survey in the territory of the Komenda Municipality. The analysis used classification trees and the naďve Bayesian classifier, two standard approaches from supervised data mining. Below we provide details on the data and methodological approaches used.

2.1 Preference maps and spatial characteristics

We used data acquired in a 2001 public survey that encompassed the territory of Komenda, a small and fast developing municipality located within the metropolitan region of the Slovenian capital, Ljubljana (Fig. 1). The municipality has about 4,800 inhabitants and an area of some 24 km^2. The survey included a representative sample of 196 participants. It was carried out as part of a land use planning exercise in the framework of undergraduate studies of landscape planning at the University of Ljubljana (Golobič et al., 2001). The survey included a textual part with questions about spatial development and spatial values, and a graphical part where the participants were given base maps in a scale of 1:25,000 and were asked to mark the areas they considered suitable for proposed land uses as well as areas that should be protected from development. We used the information from the base maps representing preferences regarding protected areas in the data mining procedure.

The graphical presentation of all stakeholders' common preferences regarding protected areas is called a map of lay suitability for protection. The maps were digitized, rasterized to a cell size of 25 x 25 meters and overlaid in order to compute the total count of how many survey participants each particular spatial cell in the marked areas included. Each cell was then described using a set of 14 spatial characteristics that include the cells' distances to buildings of cultural and educational importance, distances to rivers and forests, the land use of the cells, and similar (Table 1).

2.2 Data preparation and preprocessing

The number of cells in the entire municipal area is 38,997. For data mining, a standard tabular, attribute-value description was used where each row provided data on a specific cell and where every column recorded a value of each of the 14 spatial characteristics. The latter were manually discretized (see Table 1) to make the later interpretation of the inferred models easier and to enable the use of the simple naďve Bayesian classifier technique with its intuitive presentation in the form of a nomogram. The choice of classification-based supervised modeling techniques also dictated the discretization of a target variable – lay suitability for protection – where the participating planning expert decided to classify the cells as preferential if they had been marked by at least eight participants. The discretization was done so as to preserve the key proposals of

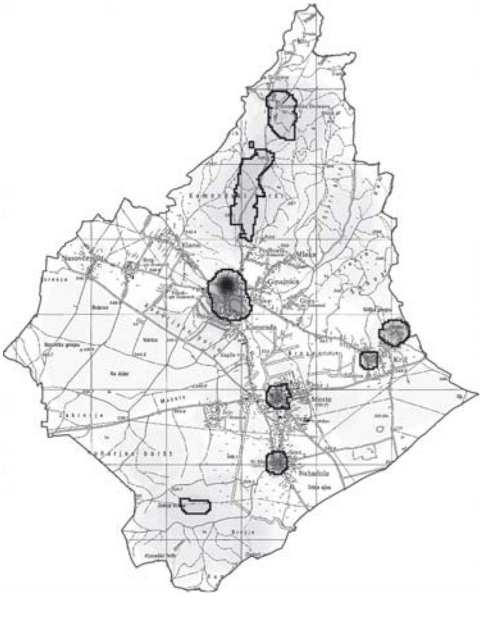


Figure 1: *Lay suitability for protection. The frequency of preferred areas for protection, marked on the preference maps, is higher in the darker areas. Areas marked more than eight times on the preference maps are outlined (one square = 1 km).*

Spatial chara-cteristic	Description	Values	Description of the value	ReliefF score	Info Gain score
dist_church	Distance to churches	-300m 300-750m 750m-	Distance up to 300 m Distance from 300 to 750 m Distance over 750 m	0,061	0,041
dist_pnature	Distance to valuable natural features (objects)	-100m 100-400m 400m-	Distance up to 100 m Distance from 100 to 400 m Distance over 400 m	0,039	0,039
dist_culture	Distance to cultural heritage	-150m 150-400m 400m-	Distance up to 150 m Distance from 150 to 400 m Distance over 400 m	0,039	0,037
land use	Land use	forest agr_mix pasture field urban settlement	Forest Mixed agriculture Pasture Field Urban area Settlement	0,048	0,027
dist_hipo	Distance to hippodrome	-500m 500-1000m 1000m-	Distance up to 500 m Distance from 500 to 1000 m Distance over 1000 m	0,120	0,022
dist_farm	Distance to farms	-150m 150-400m 400m-	Distance up to 150 m Distance from 150 to 400 m Distance over 400 m	0,002	0,017
dist_psata	Distance to river Pšata	-100m 100-400m 400m-	Distance up to 100 m Distance from 100 to 400 m Distance over 400 m	0,059	0,013
dist_hidro	Distance to other hydrological features	-100m 100-300m 300m-	Distance up to 100 m Distance from 100 to 300 m Distance over 300 m	0,014	0,010
dist_school	Distance to schools	-500m 500-1000m 1000m-	Distance up to 500 m Distance from 500 to 1000 m Distance over 1000 m	0,123	0,008
visibility	Visibility	low medium high	Low visibility Medium visibility High visibility	0,007	0,005
anature	Protected areas of nature	no feature prot_area	No feature Protected area of nature	-0,007	0,001
dist_forest	Distance to forests	forest -300m 300m-	Forest Distance up to 300 m Distance over 300 m	0,050	0,001
flood	Flood areas	no feature flood	No feature Flood area	0,030	0,001
DEM	Height above sea level	low medium high	320-360 m above sea level 360-400 m above sea level Over 400 m above sea level	0,052	0,000

Table 1: *List of spatial characteristics and their predictive value (ReliefF and Info gain scores)*

the survey participants (Fig. 1). This resulted in 1,175 preferential cells, i.e. slightly above 3% of the total number of cells on the map.

Out of the 14 selected spatial variables only nine most important ones were used to build the final model. We used two standard scoring techniques that report the degree of relations between the values of the spatial variable and the class. The first one was an entropy-based measure called information gain (Quinlan, 1986), a univariate measure that considers each spatial variable separately and assesses the "purity" of cell subsets characterized by a specific value of the variable. Another variable scoring technique we used was ReliefF, a multivariate measure that assesses the usefulness of the variables on the basis of their ability to distinguish between similar instances belonging to different classes (Kononenko, 1994). A particular advantage of ReliefF is that it can expose variables that are by themselves not very informative, but become useful in the specific contexts determined through the values of other variables. After assessing these two scores, we excluded variables with scores for either of the measures below 0.005 from further analysis. These variables are: height above sea level, distance to farms, protected areas of nature, distance to forests, and flood areas (see Table 1).

2.3 Inference of predictive models

Predictive models were inferred using two distinct yet highly popular supervised data mining approaches: the Bayesian classifier and an inference of classification trees. Data were analyzed and the models were scored in the Orange open-source data mining framework (Demšar et al., 2004).

Compared to other machine learning methods, the Bayesian classifier is perhaps one of the simplest techniques, yet a surprisingly powerful one to construct models that can predict the probability of target variables based on the values of predictive variables (Kononenko, 1993; Možina et al., 2004). The Bayesian rule to assess the probability of target value (also referred to as class) c given a vector of values of predictive variables $X = (a_1, a_2 ... a_n)$ is:

$$P(c|X) \approx P(c) \prod_i P(a_i|c)$$

where $P(c)$ denotes an a priori probability of the class, and $P(a_i|c)$ a conditional probability of the value of the *i*-th predictive feature $a_i|$ given the value of the class. The above equation asserts that the posterior probability of the class given a description of the data instances (predictive feature values) is proportional to the a priori probability of the class times the product of all conditional probabilities of the feature values given the value of the class. Because the same formula holds for instances that do not belong to the class, the odds for the target class *c* can be computed by:

$$\frac{P(c|X)}{P(\bar{c}|X)} = \frac{P(c)}{P(\bar{c})} \prod_i \frac{P(a_i|c)}{P(a_i|\bar{c})}$$

and the probability of $P(c|X)$ can be computed from the above equation using substitution

$$P(\bar{c}|X) = 1 - P(c|X).$$

The main limitations of the Bayesian classifier arise from its assumption of conditional independency between the predictive features. Despite this, the method is frequently used because of its simplicity and often good predictive accuracy. Odds as expressed in the above equation can be turned into log odds, expressed through the sum of the ratios of class/non-class conditional probabilities, and effectively visualized using what is called Bayesian nomogram (Možina et. al., 2004). This visualization device and its utility to reveal the individual effects of predictive factors in an explainable and intuitive form are the main reasons we chose this particular method and why we advocate the use of this approach when mining spatial planning data.

While, in general, the Bayesian model performs well in terms of predictive accuracy, it may prove inferior to more complex modeling techniques when any interactions of predictive factors are present and when their discovery may be crucial for constructing a reliable probability predictor. A popular supervised data mining method that may reveal such feature combinations is inference of classification trees (Quinlan, 1986; Witten and Frank, 2000). The methods employs a divide-and-conquer approach: the data set is split into subsets according to the values of selected predictive feature, repeating the procedure on the resulting subsets until all data instances within a single group are sufficiently “pure”, i.e. they mainly contain instances of a single class. In order to build small trees with a sufficient instance-based representation in the leaves, the feature used for splitting the set is chosen so that it maximizes the “purity” of the resulting set. Using a set of values of predictive features, the class is then predicted by traversing the tree from the root to one of its leaves, where the path taken depends on the values of the predictive features. The predicted class is the majority class of instances in that node. If the class probability is to be predicted, then it is computed from the class distribution of the instances in the leaf node. In our work, we applied a variant of the C4.5 classification tree induction algorithm as implemented in the Orange data mining suite (Demšar et al., 2004) with default attributes, except for the request to build trees with at least 50 instances per leaf.

2.4 Model evaluation

Our evaluation of the model’s quality was initially based on the expert’s comprehension of the planning problem. This is a subjective evaluation of the appropriateness of the selection of variables, of the clarity and comprehension of the model, and of the feasibility of its interpretation. This “qualitative” assessment of the model needs to be complemented with a quantitative evaluation of the model’s predictive properties. Here we used two standard scores: classification accuracy (CA) and the area under the receiver operating curve (AUC). CA reports the probability of a correct classification of a data instance, whereas AUC reports the probability that the model will distinguish between a positive instance (in our case a cell proposed as suitable for protection) and a negative one (in our case a cell that was not proposed as suitable).

The scores are assessed by a standard ten-fold cross-validation procedure (Witten and Frank, 2000) whereby the data are split into 10 subsets of an approximately equal size and class distribution. Classification models are then built from nine subsets, while the remaining one is used for testing. The procedure is repeated ten times, each time using a different subset for testing. Average performance scores as computed in 10 iterations of the cross-validation procedure are reported. Note that the cross-validation assesses the predictive accuracy of the modeling techniques, these then being an approximation of the scores the model would have achieved for data instances provided they were drawn from the same distribution.

The model's quality was also evaluated by visually comparing the modeled lay suitability maps with the original one. To this end, data mining methods were used not only as tools for explaining the data but also as tools for predicting. The same original input data used for inferring the model were used to predict the cell's probability of the suitability for protection. We refer to the resulting map as the modeled lay suitability map. The degree of concurrence with the original lay suitability was assessed subjectively. Note that this comparison was done for illustrative reasons and, since tested on the same data as trained, may be prone to overfitting.

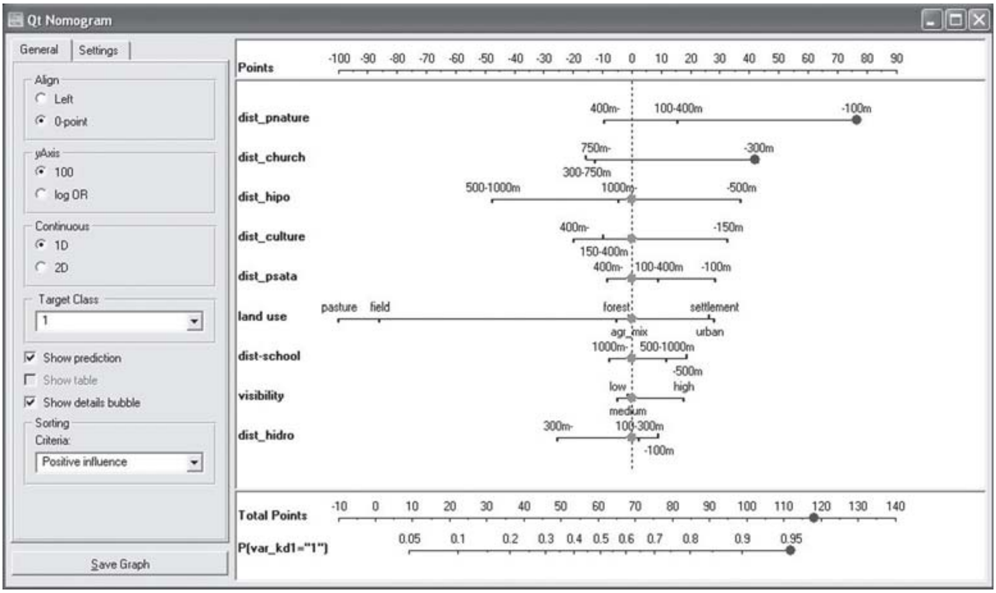
3 RESULTS

We used the methods described in the previous section to infer the naďve Bayesian classifier and classification tree. Below we provide their visualization while also assessing the predictive accuracy of the two methods.

3.1 The Bayesian classifier and nomogram-based visualization

The nomogram for the constructed Bayesian classifier is presented in Fig. 2. The model's structure shows the relative influence of the individual spatial variable values on the probability that the analyzed cells were chosen as being suitable for protection. Feature values on the right side of the vertical dotted line vote in favour of the cells' suitability, while those on the left side of the dotted line vote against it. The distance from the dotted line corresponds to the magnitude of the effect. It is evident from the nomogram that the distance to valuable natural features or objects has the biggest potential influence on lay suitability for protection. Those natural features have previously been recognized as particularly worthy at the municipal or state level and are therefore already under a legal protection regime. The lay suitability for protection areas are also around churches and other cultural heritage objects, in the vicinity of the Pšata river and other hydrologic phenomena as well as in the urban environment and areas of high visibility. Suitable areas are also in the neighbourhood of the hippodrome and schools. Pasture and field land use are very powerful negative indicators of suitability.

The particular implementation of the Bayesian nomogram in the Orange data mining suite enables an interactive what-if analysis and prediction of the cells' suitability for protection. A result of such a prediction can be observed in the snapshot in Fig. 2 where the values of predictive features can be interactively set in the upper part of the nomogram and immediately reflected in the probability target class in the lower part of the nomogram. In the snapshot only the values of



General | Settings

Align
○ Left
● 0-point

yAxis
● 100
○ log OR

Continuous
● 1D
○ 2D

Target Class
1

☑ Show prediction
☐ Show table
☑ Show details bubble

Sorting
Criteria:
Positive influence

Figure 2: *A Bayesian nomogram (a snapshot from Orange's corresponding widget) that visualizes the effect of values of predictive features on the probability of the lay suitability for protection (target class value "1"). See Table 1 for variable descriptions and values.*

two predictive features were set: the distance to valuable natural features (up to 100 meters) and the distance to churches (up to 300 meters) which, according to the Bayesian classifier, result in a 97% probability that cells with these spatial characteristics are also those that are suitable for protection. If the immediate vicinity (up to 150 m) of cultural heritage is also taken into consideration, then the probability rises to 99%.

The nomogram shows that the stakeholders highly value both the natural and cultural qualities of their living area. The interactive quality of the Orange system enables us to research them separately in a fast and simple way. The limitation of the interpretation to just two spatial characteristics determining cultural qualities, i.e. churches and other cultural heritage, shows that cells in their vicinity are, with a 71% probability, also those proposed for protection. In a similar way, the limitation of the interpretation to natural qualities shows that those cells that are close to valuable natural features, to the Pšata river and other hydrological features are, with a 96% probability, also the cells proposed for protection. Proximity to valuable natural features alone assures a 71% probability.

Let us examine the probabilities if certain valuable cultural and natural features appear separately from each other. In areas near to churches and away from valuable natural features the probability is just 18%. In the reverse case, when the areas are near valuable natural features and away from churches, the probability of a proposed protection area is still relatively high at 50%.

It can be assumed that natural qualities are a little more valued among the stakeholders than cultural ones. This information is somewhat less persuasive since the stakeholders did not mark the already protected areas of nature on the preference maps. Due to its low level of importance for the model composition, the variable is even excluded from the analysis (see Table 1). A possible interpretation is that people really perceive natural features along rivers and brooks inside settlements as more valuable maybe just because of their location within the urban environment where natural features play, as a rule, several different functions, including strong social ones. Protected natural areas are, on the other hand, somewhat withdrawn from settlements, and lie in a broader, naturally better preserved hinterland. For a more solid interpretation, additional analyses and contacts with the local inhabitants are necessary.

3.2 Classification tree

The classification tree model illustrating lay suitability for protection is given in Fig. 3. The most significant variable appearing at the root of the tree is again proximity to valuable natural features.

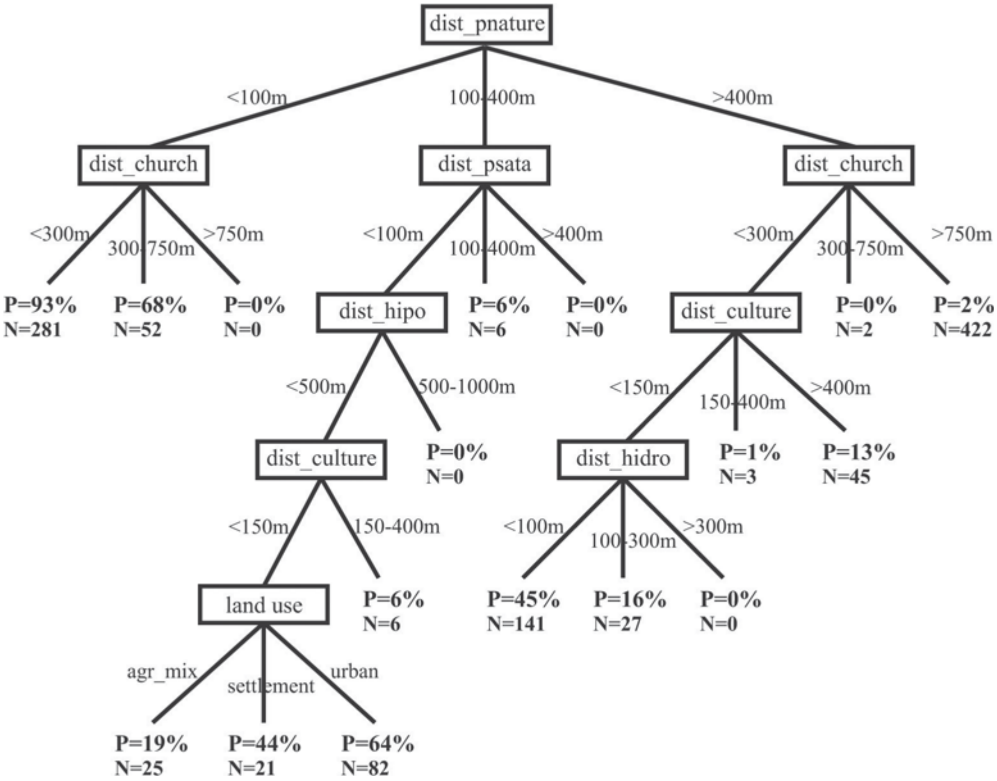


Figure 3: *Classification tree as a prediction model for the lay suitability for protection. P denotes the predicted probability of the lay suitability for protection. N denotes the number of all cells in the leaf. See Table 1 for variable descriptions and values.*

The probability that a cell close to valuable natural features is suitable for protection is 71%. When descending along the same branch of the tree by one level one arrives at two interesting decision rules:

IF a cell is close to valuable natural features (up to 100 m) AND close to churches (up to 300 m)

THEN the probability of its suitability for protection is 93%;

IF a cell is close to valuable natural features (up to 100 m) AND far from churches (over 750 m)

THEN the probability of its suitability for protection is 0%.

Both rules indicate that proximity to valuable natural features is not the only key criterion for people's decisions regarding areas that are suitable for protection.

The second branch of the tree gives evidence of the fact that in areas located 100 – 400 meters from valuable natural features the most suitable parts are those lying nearer to the Pšata river, the hippodrome and cultural heritage as well as in the urban environment.

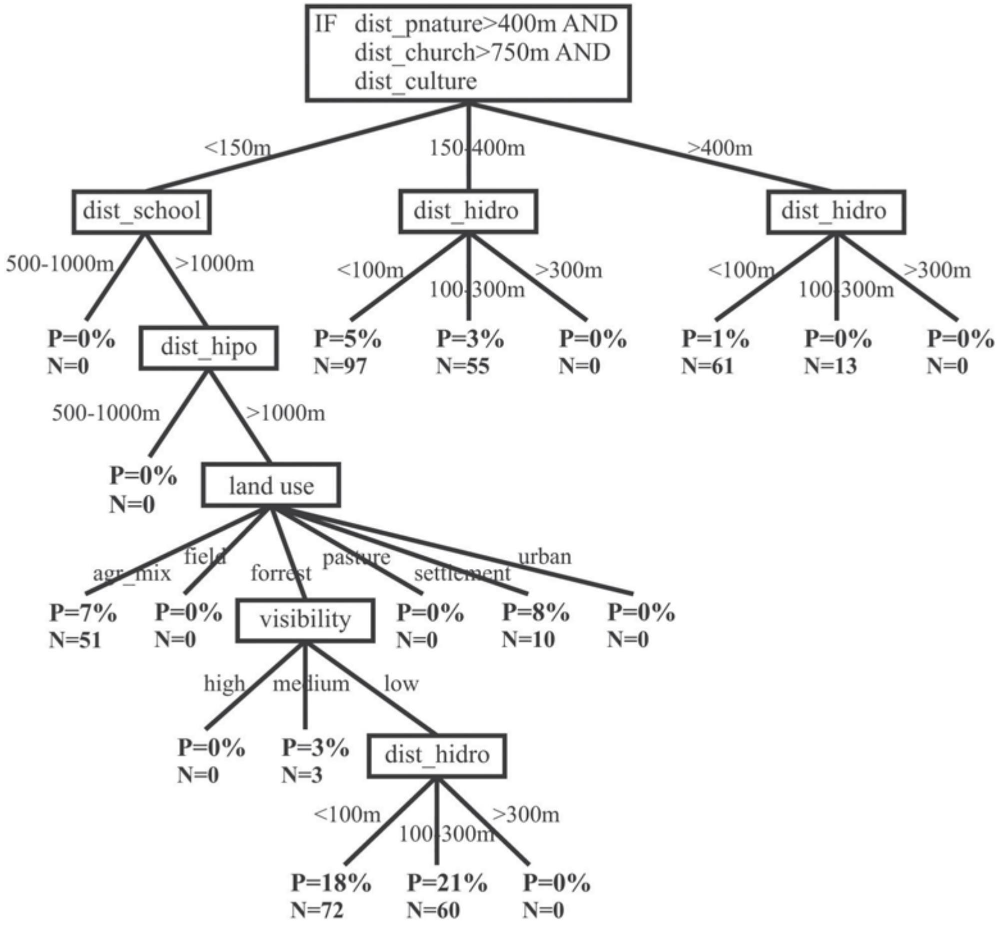


Figure 4: *Classification subtree illustrating the lay suitability for protection of cells that are situated more than 400 meters away from valuable natural features and more than 750 meters from churches. P denotes the predicted probability of the lay suitability for protection. N denotes the number of all cells in the leaf. See Table 1 for variable descriptions and values.*

The third branch of the tree includes all the other cells, i.e. those not situated in proximity to valuable natural features (more than 400 meters). The probability of being suitable for protection there is only 2%, but the number of cells included in this branch is 640, which is more than half of all the cells suitable for protection. These cells' probability of suitability for protection of approximately one-third increases with proximity to other objects of cultural heritage and proximity to water features. 422 cells, i.e. 36% of all cells suitable for protection, are situated away from valuable natural features (more than 400 meters) and from churches (more than 750 meters). From here on the tree is not branched and does not enable interpretation of these cells, so we employed an interactive building of the tree from the chosen node on. The additional subtree is presented in Fig. 4. It shows that the majority of the mentioned cells are situated near to cultural heritage, yet in further branching of the tree the most important factor for their interpretation is the proximity to water features. Its occurrence increases the probability of suitability for protection in all branches of this tree.

A brief interpretation of the classification tree would be that residents designated as valuable and therefore worthy of protection in the first place:

- areas with valuable natural features in the proximity of churches;
- areas with cultural heritage along the Pšata river situated within the settlement; and
- areas around certain other water features outside the settlement.

3.3 Quantitative evaluation and map-based illustration of the models' quality

Quantitative scores that estimate the quality of prediction of the two methods used are shown in Table 2. Note that while both scores are high, it is the relatively high AUC that demonstrates that both methods may be well suited for the task. Namely, while the classification accuracy is high, our problem is that groups of cells belonging to different classes are highly unbalanced: only 3% of the cells are marked for protection, making the baseline classification accuracy 0.970 (for the model that would always classify cells as being "not suitable for protection").

There are some differences between the predictions of the two models, and they are best illustrated in map-based presentations of their modeled suitability for protection. Suitable areas for protection predicted by the Bayesian model (Fig. 5) are much more expansive than the originally proposed suitable areas, and at the same time some of the originally proposed suitable areas with specific spatial characteristics become lost in them (e.g. Komendski Boršt in the northern part of the municipality). In our opinion, the model excessively generalizes the relationships between spatial characteristics and suitability for protection. The map of modeled suitability for protection by

	Evaluation scores	
Modelling technique	CA	AUC
Bayes	0.976	0.905
Classification tree	0.978	0.778

Table 2: *Quantitative evaluation scores for the two modelling techniques used*

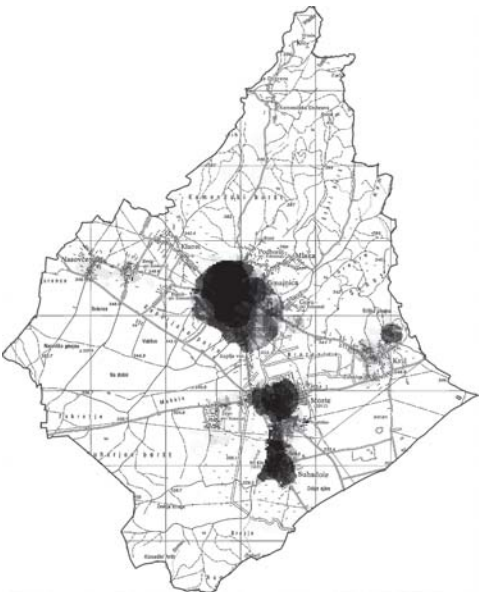


Figure 5: *Modeled suitability for protection by using the Bayesian method. Darker areas are more suitable for protection (one square = 1 km).*

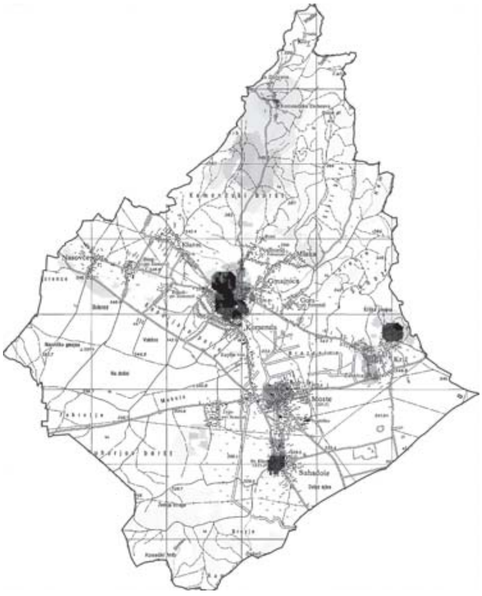


Figure 6: *Modeled suitability for protection by using the classification tree. Darker areas are more suitable for protection (one square = 1 km).*

using classification tree (Fig. 6) concurs more with the original map (Fig. 1). We therefore assume that the inference of classification tree models the non-linear spatial characteristics in the proposed areas for protection more accurately than the naďve Bayesian model.

4 DISCUSSION

We have demonstrated that data from preference maps can be successfully interpreted by employing the selected data mining methods. The goal of data mining in this domain is to infer the relationships between factors that influence lay people's preferences, exposing them and making them available to planning experts. The resulting preferential models, in our study explicitly expressed as a classification tree or a Bayesian nomogram, provide a structured and systematic means for informing planning experts about residents' views. The principal goal of interpreting preference maps is thus to improve the quality and increase the depth of communication between residents and planners in a language that both can understand well.

In our particular application, the finding from both inferred models is that the residents want to protect certain valuable natural features, hydrological features including the Pšata river, and cultural qualities of their local area (churches and other cultural heritage), especially when these are in the area of the settlement. These general spatial characteristics of the analyzed preference areas can be seen more clearly in a nomogram than in a classification tree. The reason is that the structure of the Bayes model is simpler and clearer. Still, this type of model only includes one common general interpretation for all more or less different lay suitability areas. Such an interpretation may be too general, a fact clearly illustrated by the corresponding map of modeled suitability (Fig. 6). Golobič and Marušič (2007) encountered the same problem in their model that used multiple regression. On the other hand, the inference of a classification tree seems to avoid such overgeneralization by being able to model nonlinear relations among spatial characteristics. This model's hierarchical composition enables a structured interpretation that can focus on specific rules – paths from the root to the leaves of the tree – rather than on the model as a whole. The classification tree we constructed suggests that the proposed areas for protection within a settlement are situated either in an area where valuable natural features and churches occur close together, or in the area of the Pšata river with objects of cultural heritage. Protected areas outside the settlement are chiefly characterized by their proximity to water features. In our opinion and at least for the case study presented here, classification trees provide a more detailed interpretation of the preference maps, whereas the interpretation of the naďve Bayesian nomogram is well suited to giving an overall idea of the effects of each of the studied spatial characteristics. This is also confirmed by better concurrence of the map of the modeled lay suitability for protection by using a classification tree with the original lay suitability map.

The utility of data mining approaches for interpreting suitability maps may dictate the development of tools that integrate GIS and machine learning procedures suiting interactive visual exploration of the models. Such a tool would give people the opportunity to not only form their proposals implicitly by selecting preference areas for future development directly on a map, but also to explicitly test and eventually change decision criteria extracted by machine learning methods. The cyclic process of what is called exploratory decision-making with the alternating implicit and explicit expressing of preferences for future development would thus enable experts to acquire people's proposals together with "verified" explicit decision criteria, while at the same time people would not lose the possibility of making their proposals in an ad hoc manner. An example of such an integrated environment is DECADE developed by Jankowski et al. (2001). Its main limitation is that it is suitable for the analysis of small numbers of previously defined larger, naturally homogeneous spatial units. But there is a great deal of uncertainty in the determination of crisp boundaries between naturally homogeneous areas (Burrough and McDonnell, 1998). At the same time, homogeneous areas on which spatial planning are based cannot only be defined by their natural characteristics but also, and primarily, in connection with the planned activity and its geometry (Mejač, 1997). Therefore, the interpretation of preference maps on the basis of the analysis of spatial cells or pixels seems more suitable. The analysis of arbitrarily defined preference areas regarding not only the location but also the size and shape of an area could be a more accurate and thus a more reliable indicator of people's preferences regarding future development or protection in their local area.

Instead of or besides incorporating cell-specific spatial characteristics, the analysis may include more complex, holistic data. For example, the exploration of preference criteria may be based on models of building costs, qualities of the natural environment, attractiveness of the living environment etc., which may provide predictive factors that could be closer to the lay public's reasoning than a set of specific information about the space (Golobič and Marušič, 2007; Ventura et al., 2005). Note, however, that such complex criteria and models would merely change or add more features to our data set, with the analysis procedure proposed in this paper remaining unchanged.

Predictive models like those we have developed for land use planning in the Komenda Municipality can be easily combined with other sources in alternative planning proposals. The interpretation of preference maps has proved to provide an important supplement to the results of the analysis of the textual part of the survey where especially natural qualities (e.g. forests, running water) were evaluated highly by the inhabitants of the Komenda Municipality (Golobič et al., 2001). The textual answers are principled while the preference maps discovered more detailed and spatially defined evaluation criteria. The results of the formalized interpretation of preference maps can be used not only in the stage before the formation of spatial alternatives, but also in other planning stages, e.g. as support for choosing between alternatives or for evaluating a plan proposal, depending on the planning method applied. In any case, the intention of including the formalized interpretation of preference maps in the planning process remains to bring the spatial plan closer to the needs and wishes of users.

5 CONCLUSION

People's deliberations about future development in their living environment should and will be extensively considered in land use planning procedures. Their direct, map-based identification of their preferences facilitates their participation in the planning process and enables a reliable identification of conflict areas (Golobič and Marušič, 2007). The data gathered in this way open up possibilities to infer reasoning patterns used by lay people and which are implicitly present in preference maps. Revealing these patterns in the form of rules can largely support communication between lay people and planners, improving the transparency of the analytical procedures and enabling an argued discussion and easier reconciliation of the different views of everyone involved.

We used two different data mining methods and showed that the proposed procedure can uncover interesting, interpretable rules from preference maps. In our opinion the utility of otherwise established data mining techniques in land use planning and inference of models, such as those reported in the paper, can help build bridges between the public and experts, thereby providing a way to diminish – according to Friedman (1992) and Baxmann (1997) – the still existing and limiting communication gap and contribute to more successful spatial-planning solutions.

References:

Aravot, I. (1996). Integration of future user's evaluations into the process of urban revitalization. Evaluation and Program Planning, 19 (1), 65-78.

Ball, J. (2002). Towards a methodology for mapping 'regions for sustainability' using PPGIS. Progress in Planning, 58, 81-140.

Bartol, B., Golobič, M., Kavčič, I., Logar, J., Marušič, J., Mlakar, A., Simonič, T. (1998). Anketno ocenjevanje kot način pridobivanja meril v postopku prostorskega planiranja. Urbani izziv, 9 (2), 99-103.

Baxmann, M. (1997). Spatial consensus-building through access to web-based GIS: An online planning tool for Leipzig. www.spatial.maine.edu/ucgis/testproc/baxmann/baxmann.html (12. 1. 2005)

Bui, E. N., Henderson, B. L., Viergever, K. (2005). Knowledge discovery from models of soil properties developed through data mining. Ecological Modelling, 191 (3-4), 431-446.

Burrough, P. A., McDonnell, R. A. (1998). Principles of Geographical Information Systems. Oxford: Oxford University Press.

Carver, S. J. (1991). Integrating multi-criteria evaluation with geographic information systems. Int. Jour. Remote Sensing, 5 (3), 321-339.

Carver, S. (2003). The future of participatory approaches using geographic information: developing a research agenda for the 21st century. URISA 15, 61-72.

Demšar, J., Leban, G., Zupan, B. (2004). Orange: From experimental machine learning to interactive data mining (www.ailab.si/orange), Ljubljana: Faculty of computer and information science.

Friedman, J. (1992). Empowerment. Cambridge: Blackwell.

Golobič, M., Marušič, J., Polič, M. et. al. (2001). Prostorski razvoj Komende: stališča prebivalcev občine Komenda do prostorskega razvoja občine: rezultati ankete. Ljubljana: Biotechnical Faculty, Department for Landscape Architecture.

Golobič, M., Marušič, J. (2007). Developing an integrated approach for public participation: a case of land use planning in Slovenia. Environment and Planning B,34 (6), 993-1010.

Harris, T., Weiner, D. (1998). Community-integrated GIS for land reform in Mpumalanga province, South Africa. Department of Geology and Geography, West Virginia University. http://www.ncgia.ucsb.edu/varenius/ppgis/papers/harris.html (12. 1. 2005)

Jankowski, P., Andrienko, N., Andrienko, G. (2001). Map-centred exploratory approach to multiple criteria spatial decision making. Int. J. Geographical Information Science, 15 (2), 101-127.

Janssen, M. A., Goosen, H., Omtzigt, N. (2006). A simple mediation and negotiation support tool for water management in the Netherlands. Landscape and Urban Planning, 78, 71-84.

Kingston, R., Carver, S., Evans, A., Turton, I. (2000). Web-based public participation geographical information systems: an aid to local environmental decision-making. Computers, Environment and Urban Systems, 24, 109-125.

Kobler, A., Adamič, M. (2000). Identifying brown bear habitat by a combined GIS and machine learning method. Ecological Modelling, 135 (2-3), 291-300.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. Applied Artificial Intelligence, 7, 317-337.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. Proceedings of ECML-94, Catania.

Linden, M., Sheehy, N. (2004). Comparison of a verbal questionnaire and map in eliciting environmental perceptions. Environment and Behaviour, 36 (1), 32-40.

Macnab, P. (1998). There must be a catch: Participatory GIS in a Newfoundland fishing community. NCGIA: Marginalization and Public Participation GIS, 15-17 October 1998, Halifax.

Martin, W. E., Wise Bender, H., Shields, D. J. (2000). Stakeholders' objectives for public lands: Rankings of forest management alternatives. Journal of Environmental Management, 58, 21-32.

McClure, W. (1997). The rural town: Designing for growth and sustainability. Moscow: University of Idaho, Center for Business Development and Research.

Mejač, Ž. (1997). Environmental demands in strategic physical planning – the comparative analysis of different approaches. Master of science thesis. Ljubljana: Biotechnical Faculty, Department for Landscape Architecture.

Možina, M., Demšar, J., Kattan, M., Zupan, B. (2004). Nomograms for visualization of naive Bayesian classifier. Proc. of Principles and Practice of Knowledge Discovery in Databases (PKDD-2004), Pisa, 337-348.

Naderi, J. R., Raman, B. (2005). Capturing impressions of pedestrian landscapes used for healing purposes with decision tree learning. Landscape and Urban Planning, 73, 155-166.

Newell, A. and Simon, H. A. (1972). Human Problem Solving. Englewood Cliffs, New York: Prentice-Hall.

Ogris, N. (2007). Model of forest health in Slovenia. Doctoral thesis. Ljubljana: Biotechnical Faculty, Department of Forestry and Renewable Forest Resources.

Pettit, C., Pullar, D. (1999). An Integrated Planning Tool Based Upon Multiple Criteria Evaluation of Spatial Information. Computers, Environment and Urban Systems, 23, 339-357.

Polič, M., Mancin, M., Bartol, B., Marušič, J. (1991). Stališča prebivalcev občine Grosuplje do nekaterih vidikov njenega razvoja. Grosuplje: Grosuplje Municipality.

Polič, M., Klemenčič, M., Kos, D., Kučan, A., Marušič, J., Ule, M., Natek, K., Repovš, G. (2002). Spoznavni zemljevid

Slovenije. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1 (1), 81-106.

Sanoff, H. (1991). Visual research methods in design. New York: Van Nostrand Reinhold.

Soini, K. (2001). Exploring human dimensions of multifunctional landscapes through mapping and map-making. Landscape and Urban Planning, 57, 225-239.

Ventura, S., Niemann, B., Sutphin, T., Chenoweth, R. GIS-enhanced land use planning in Dane county, Wisconsin, Land Information and Computer Graphics Facility, University of Wisconsin-Madison. http://www.ncgia.ucsb.edu/varenius/ppgis/papers/ventura.html (6. 10. 2005)

Voogd, H. (1983). Multi-criteria evaluations for urban and regional planning, London: Princeton University.

Webler, T., Tuler, S., Krueger, R. (2001). What is a Good Public Participation Process? Five Perspectives from the Public. Environmental Management, 27 (3), 435-450.

Witten, I. H., Frank, E. (2000). Data Mining, San Francisco: Morgan Kaufmann Publishers.

Zhang, B., Valentine, I., Kemp, P. (2005). Modelling the productivity of naturalised pasture in the North Island, New Zealand: a decision tree approach. Ecological Modelling, 186, 299-311.

Received to publication: 19 January 2009
Accepted: 26 February 2009

Lidija Breskvar Žaucer
Biotechnical Faculty – Department of Landscape Architecture, Jamnikarjeva 101, SI-1000 Ljubljana
E-mail: lidija.zaucer@bf.uni-lj.si

Ass. Prof. Blaž Zupan, PhD
Faculty of Computer and Information Science, Tržaška 25, SI-1000 Ljubljana and
Baylor College of Medicine, Dept. of Human and Molecular Genetics, 1 Baylor Plaza, Houston, TX 77030, USA
E-mail: blaz.zupan@fri.uni-lj.si

Ass. Prof. Mojca Golobič, PhD
Biotechnical Faculty – Department of Landscape Architecture, Jamnikarjeva 101, SI-1000 Ljubljana and
Urban Planning Institute of the Republic of Slovenia, Trnovski pristan 2, SI-1000 Ljubljana
E-mail: mojca.golobic@uirs.si