Formal Interpretation of Preference Maps –
A Data Mining Approach

Lidija BRESKVAR ŽAUCER, Blaž ZUPAN and Mojca GOLOBIČ

1 Introduction

People commonly use rules of thumb or heuristics when making judgments (NEWELL & SIMON, 1972). This holds true also for non-experts involved in a spatial planning process (JANKOWSKI ET AL., 2001). It is therefore appropriate to enable them to express their opinions about future spatial development in a simpler and less formal way. One of these ways is to enable them to express their planning preferences implicitly by identifying the areas they depict on a cartographic base map. Such maps of preferences regarding future spatial development are also called preference maps.

Maps are generally acknowledged as an effective and practical tool for informative communication between planning experts and non-experts (BALL, 2002; CARVER, 2003; JANKOWSKI ET AL., 2001). It is also known that preference 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 & SHEENY, 2004) and thus provide a much better indication of whether and where conflicts regarding land use are expected (GOLOBIČ & MARUŠIČ, 2007). If information and knowledge hidden behind preference areas are made explicit and presented in an interpretable form, an improvement in the effectiveness of developing consensual planning solution can be expected (JANSSEN ET AL., 2006; MARTIN ET AL., 2000).

There are many examples of the use of preference maps in planning processes (e.g. HARRIS AND WEINER, 1998; JANKOWSKI ET AL., 2001; KINGSTON ET AL., 2000; MACNAB, 1998). However, reports on attempts to formalize their interpretation are at best sparse. ARVOT (1996) and GOLOBIČ & MARUŠIČ (2007) employed classical statistical regression methods and resulted in a more mathematical model whose findings are more difficult to use in communication with non-experts.

In this paper we propose an alternative methodology that uses supervised data mining techniques (WITTEN AND FRANK, 2000) to explicitly infer the relationships between graphic responses, or so-called preference areas, and spatial characteristics in a form of preferential rules that are easy to interpret. The advantage of the proposed methodology is also its ability to identify more or less different areas or interesting data subgroups with their own specific spatial characteristics, as opposed to the statistical regression methodology which results in only one general interpretation. The method relates planning preferences to a limited set of spatial characteristics and does not consider other factors influencing people’s attitudes, such as the living habits, experiences, emotional relations with the environment, interests and wishes (ARVOT, 1996; GOLOBIČ & MARUŠIČ, 2007). It therefore does not attempt to reveal the deeper, tacit knowledge of individuals, but rather to provide a vehicle for argued discussion between planners and non-experts. In the spatial
planning field these methods have already 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 Material and Methods

2.1 Study area and preference maps

We used data acquired in a 2001 public survey (GOLABIČ ET AL., 2001) 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 24 km². The survey that included a representative sample of 196 participants consisted of 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 maps representing preferences regarding protected areas in the data mining procedure.

Graphic responses of all survey participants were digitized, rasterized to a cell size of 25 × 25 meters and overlaid to get a collective map of residents’ suitability for protection. Each cell was given 15 attributes: 14 of these describing its spatial characteristics (Tab. 1) and one attribute disclosing how many respondents included it in a proposed protection area.

Fig. 1: Location of the case study area with residents’ suitability for protection
2.2 Data preparation and preprocessing

Values of spatial characteristics were manually discretized to make the later interpretation of the inferred models easier and enable the use of the simple naïve 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 — residents’ suitability for protection — where the cells were classified as preferential if they had been marked by at least eight participants. The discretization was done so as to preserve the key proposals of the survey participants (Fig. 1). This resulted in 1,175 preferential cells out of 38,997 cells in the entire municipal area.

We tested the relevance of the chosen spatial variables by two standard scoring techniques: Relief, 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), and 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. Out of the 14 selected spatial variables only nine of the most important ones were used to build the model. Variables with scores for either of the measures below 0.005 were excluded from further analysis (see Tab. 1).

<table>
<thead>
<tr>
<th>Spatial characteristic</th>
<th>Description</th>
<th>ReliefF score</th>
<th>Info Gain score</th>
<th>Selection for model building</th>
</tr>
</thead>
<tbody>
<tr>
<td>dist_church</td>
<td>Distance to churches</td>
<td>0.061</td>
<td>0.041</td>
<td>√</td>
</tr>
<tr>
<td>dist_pnature</td>
<td>Distance to valuable natural features (objects)</td>
<td>0.039</td>
<td>0.039</td>
<td>√</td>
</tr>
<tr>
<td>dist_culture</td>
<td>Distance to cultural heritage</td>
<td>0.039</td>
<td>0.037</td>
<td>√</td>
</tr>
<tr>
<td>land use</td>
<td>Land use</td>
<td>0.048</td>
<td>0.027</td>
<td>√</td>
</tr>
<tr>
<td>dist_hipo</td>
<td>Distance to hippodrome</td>
<td>0.120</td>
<td>0.022</td>
<td>√</td>
</tr>
<tr>
<td>dist_farm</td>
<td>Distance to farms</td>
<td>0.002</td>
<td>0.017</td>
<td>x</td>
</tr>
<tr>
<td>dist_psata</td>
<td>Distance to river Pšata</td>
<td>0.059</td>
<td>0.013</td>
<td>√</td>
</tr>
<tr>
<td>dist_hidro</td>
<td>Distance to other hydrological features</td>
<td>0.014</td>
<td>0.010</td>
<td>√</td>
</tr>
<tr>
<td>dist_school</td>
<td>Distance to schools</td>
<td>0.123</td>
<td>0.008</td>
<td>√</td>
</tr>
<tr>
<td>visibility</td>
<td>Visibility</td>
<td>0.007</td>
<td>0.005</td>
<td>√</td>
</tr>
<tr>
<td>anature</td>
<td>Protected areas of nature</td>
<td>-0.007</td>
<td>0.001</td>
<td>x</td>
</tr>
<tr>
<td>dist_forest</td>
<td>Distance to forests</td>
<td>0.050</td>
<td>0.001</td>
<td>x</td>
</tr>
<tr>
<td>flood</td>
<td>Flood areas</td>
<td>0.030</td>
<td>0.001</td>
<td>x</td>
</tr>
<tr>
<td>DEM</td>
<td>Height above sea level</td>
<td>0.052</td>
<td>0.000</td>
<td>x</td>
</tr>
</tbody>
</table>
2.3 Inference of predictive models

Predictive models were inferred using two distinct yet widely used supervised data mining approaches: the naïve 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 naïve Bayesian classifier is perhaps one of the simplest techniques yet a surprisingly powerful one (KONONENKO, 1993; MOŽINA ET AL., 2004). This probabilistic classifier is based on applying Bayes’ theorem. Besides often good predictive accuracy, it can also provide a valuable insight into the structure of the training data and the effects of the independent variables (in our case spatial characteristics) on the output class (in our case suitability for protection) probabilities. Orange data mining suit enables effective visualization of the naïve Bayesian model in a form of a so-called nomogram (MOŽINA ET AL., 2004). This visualization device and its utility to reveal the individual effects of factors in an explainable and intuitive form is the main reason we chose this particular method and why we advocate this approach for use when mining spatial planning data.

The main limitations of the naïve Bayesian classifier arise from its assumption of conditional independency between the predictive features. It may also 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 & FRANK, 2000). Classification tree recursively splits the data set according to the values of the independent variables into subsets that are as homogenous as possible. The variable used for splitting the set is chosen so that it maximizes the “purity” of the resulting set. The class of an instance is then predicted through traversing the tree from the root to one of its leaves, where the path taken depends on the values of the independent variables. We applied a variant of the C4.5 classification tree induction algorithm as implemented in the Orange data mining suite (DEMŠAR ET AL., 2004). The tree was pruned to a minimum of 50 instances per leaf.

2.4 Model evaluation

After subjective evaluation of the clarity and comprehension of the model, a quantitative evaluation of the model’s predictive properties followed. 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 10-fold cross-validation procedure (WITTEN & FRANK, 2000).

For illustrative reasons, the model’s quality was also evaluated by a visual comparison of the modeled residents’ suitability maps with the original one. For the purpose of this evaluation, the same original input data used for inferring the model were used to predict the cell’s probability of the suitability of protection.
3 Results

3.1 The naïve Bayesian classifier and nomogram-based visualization

The naïve Bayesian nomogram (Fig. 2) shows the relative influence of the individual spatial variable values on the probability that the analyzed cells were chosen as being suitable for protection. Variable values on the right side of the vertical dotted line vote in favour of the cells’ suitability, while those on the left side against it. The distance from the dotted line corresponds to the magnitude of the effect. It is evident from the nomogram that distance to valuable natural features has the biggest potential influence on residents’ suitability for protection. The residents’ 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.

The particular implementation of the naïve Bayesian nomogram in the Orange data mining suite enables an interactive what-if analysis. The values of spatial variables can be interactively set in the upper part of the nomogram and immediately reflected in a probability output class in the lower part of the nomogram. By way of illustration, only the values of two spatial variables were set on the snapshot: the distance to valuable natural features (up to 100 meters) and the distance to churches (up to 300 meters) which result in a 97% probability that cells with these spatial characteristics are also those that are suitable for protection. The nomogram shows that the stakeholders highly value both the natural and cultural qualities of their living area. But with interactive what-if analysis we can see that in areas near to churches and away from valuable natural features the probability is just

Fig. 2: A naïve Bayesian nomogram (a snapshot from Orange’s corresponding widget)
18%. In the reverse case, the probability is still relatively high at 50%. It can be assumed that natural qualities are more valued among the stakeholders than cultural ones. For a more solid interpretation, additional analyses and contacts with the local inhabitants are necessary.

3.2 Classification tree

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

Fig. 3: Classification tree as a prediction model for the residents’ suitability for protection. P denotes the predicted probability of the residents’ suitability for protection. N denotes the number of cells in the leaf.
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.

A brief interpretation of the classification tree would be that residents designated as being 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;
- areas with cultural heritage along certain other water features.

3.3 Evaluation and map-based illustration of the models’ quality

Quantitative scores that estimate the quality of the prediction of the two methods used are shown in Table 2.

<table>
<thead>
<tr>
<th>Modelling technique</th>
<th>CA</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.976</td>
<td>0.905</td>
</tr>
<tr>
<td>Classification tree</td>
<td>0.978</td>
<td>0.778</td>
</tr>
</tbody>
</table>

The differences between the predictions of the two models are best illustrated in map-based presentations of their modeled suitability for protection. Suitable areas for protection predicted by the naïve Bayesian model (Fig. 4.a) are much more expansive than the ones originally proposed by the residents (Fig. 1). At the same time some of the originally proposed suitable areas with specific spatial characteristics (e.g. Komendski boršt in the northern part of the municipality) are not included in the model. In our opinion, the model excessively generalizes the relationships between spatial characteristics and suitability for protection. The map of modeled suitability for protection by use of classification tree (Fig. 4.b) concurs more with the original map. We therefore assume that the inference of classification tree models the spatial characteristics in preference areas more accurately than the naïve Bayesian classifier.
Fig. 4a: Modeled suitability for protection, by the naïve Bayesian model
Fig. 4b: Modeled suitability for protection, by use of the classification tree. Darker areas are more suitable for protection.
4 Conclusion

People’s deliberations about future development in their living environment should and will be extensively considered in land use planning procedures. Direct, map-based identification of their preferences facilitates their participation in the planning process and enables a reliable identification of conflict areas. The goal of data mining in this domain is to infer the relationships between factors that influence 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 naïve 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 and also to improve the transparency of the analytical planning procedures.

Both data mining methods used in our case study proved to be applicable for deducing informative and interpretable rules from preference maps. In our opinion classification trees provide a more detailed interpretation of the preference maps, whereas the interpretation of the naïve Bayesian nomogram is well suited to giving an overall idea of the effects of each of the spatial characteristics studied. This is also confirmed by a visual comparison of the modeled residents’ suitability maps with the original one. The applicability of both methods and their results in the planning process should be verified also by obtaining the feedback from (a sample of) the residents who participated in the study and from the planners who were engaged in the planning process. Since the material for our experiment was collected some time ago, this remains a task for further research. The results of formal interpretation of preference maps in the form of preferential rules expressed as combinations of spatial characteristics can be used in the planning process not only as a basis for communication but also in some other ways. For example, they can be compared to experts’ criteria and used to complement the spatial evaluation process, such as the commonly used multi-criteria spatial evaluation (VOOGD, 1983). They can be also used 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 and contribute to more successful spatial-planning solutions. 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 also 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.
References


Ball, J. (2002): Towards a methodology for mapping 'regions for sustainability' using PPGIS. Prog. Plann., 58, 81-140.


