International Cartographic Association
Association Cartographique Internationale

# Exploring essential variables in the settlement selection for small-scale maps using machine learning

Karsznia, Izabela [a,*], Sielicka, Karolina [b]

[a] *Department of Geoinformatics, Cartography and Remote Sensing, Faculty of Geography and Regional Studies, University of Warsaw, Poland; i.karsznia@uw.edu.pl*

[b] *k.sielicka@uw.edu.pl*

* Corresponding author

**Keywords**: cartographic generalization, machine learning, settlement selection, small-scale

## Abstract:

The decision about removing or maintaining an object while changing detail level requires taking into account many features of the object itself and its surrounding. Automatic generalization is the optimal way to obtain maps at various scales, based on a single spatial database, storing up-to-date information with a high level of spatial accuracy. Researchers agree on the need for fully automating the generalization process (Stoter et al., 2016). Numerous research centres, cartographic agencies as well as commercial companies have undertaken successful attempts of implementing certain generalization solutions (Stoter et al., 2009, 2014, 2016; Regnauld, 2015; Burghardt et al., 2008; Chaundhry and Mackaness, 2008). Nevertheless, an effective and consistent methodology for generalizing small-scale maps has not gained enough attention so far, as most of the conducted research has focused on the acquisition of large-scale maps (Stoter et al., 2016). The presented research aims to fulfil this gap by exploring new variables, which are of the key importance in the automatic settlement selection process at small scales. Addressing this issue is an essential step to propose new algorithms for effective and automatic settlement selection that will contribute to enriching, the sparsely filled small-scale generalization toolbox.

The main idea behind this research is using machine learning (ML) for the new variable exploration which can be important in the automatic settlement generalization in small-scales. For automation of the generalization process, cartographic knowledge has to be collected and formalized. So far, a few approaches based on the use of ML have already been proposed. One of the first attempts to determine generalization parameters with the use of ML was performed by Weibel et al. (1995). The learning material was the observation of cartographers manual work. Also, Mustière tried to identify the optimal sequence of the generalization operators for the roads using ML (1998). A different approach was presented by Sester (2000). The goal was to extract the cartographic knowledge from spatial data characteristics, especially from the attributes and geometric properties of objects, regularities and repetitive patterns that govern object selection with the use of decision trees. Lagrange et al. (2000), Balboa and López (2008) also used ML techniques, namely neural networks to generalize line objects. Recently, Sester et al. (2018) proposed the application of deep learning for the task of building generalization. As noticed by Sester et al. (2018), these ideas, although interesting, remained proofs of concepts only. Moreover, they concerned topographic databases and large-scale maps. Promising results of automatic settlement selection in small scales was reported by Karsznia and Weibel (2018). To improve the settlement selection process, they have used data enrichment and ML. Thanks to classification models based on the decision trees, they explored new variables that are decisive in the settlement selection process. However, they have also concluded that there is probably still more “deep knowledge” to be discovered, possibly linked to further variables that were not included in their research. Thus the motivation for this research is to fulfil this research gap and look for additional, essential variables governing settlement selection in small scales.

The first step of this research was to create and verify a list of measurable attributes, named here variables, that are essential in the settlement selection in small scales. Then the use of ML-based models made it possible to assess the importance of the proposed variables by investigating their weights as well as the correlation among them. The scope of this research covered automatic settlements selection from the General Geographic Object Database (GGOD) at the detail level of 1: 250 000 scale to 1:500 000 scale. The sample of 16 Polish districts contained in the GGOD has been used. The settlements have been generalised using two approaches. Firstly, there were selected settlements based on the rules defined in the Polish legal guidelines what was called basic approach (Regulation of the Minister of Interior and Administration, 2011). Then, the source data was enriched with 33 additional variables and the settlement status (selected or omitted by a cartographer) acquired from the atlas map. As a final step, the automatic selection model based on decision trees was built. In both approaches, the settlement status from the atlas map was taken as reference for the evaluation. Finally, 33 variables were considered, where 16 of them had earlier been considered by Karsznia and Weibel (2018). The 17 new variables concerned among others the administrative settlement area, settlement built-up area, the presence of significant facilities within the settlement borders, the existence of important transport connections, settlement population density as well as the density of the settlement network calculated within various enumeration units. Considering a thorough set of variables makes it possible to take into account all the characteristics of the settlements that are decisive

Abstracts of the International Cartographic Association, 1, 2019. © Authors 2019. CC BY 4.0 License.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan | https://doi.org/10.5194/ica-abs-1-162-2019

in the selection process. However, in the case of ML, the number of variables should also be optimized for two reasons. Primarily, the more variables are considered, the more training data the process requires. Secondly, one should also be aware that the information extracted from numeric variables can be redundant. Besides, referring to cartographic knowledge, variables have different importance. Some variables - such as population or area - can be considered with priority. The others – like for instance the number of roads crossing the settlement - are of secondary importance. To evaluate, which variables could be omitted in ML, the assessment of the correlation strength among the variables has also been conducted.

As a result of the presented research, the automatic models of settlement selection from 1:250 000 to 1: 500 000 scale for all 16 districts and districts split into four coherent groups have been built. The accuracy of the selection and its visual correctness were compared to the results obtained from the basic approach. Regarding the district group presented in figure 1, the accuracy of the basic approach was 71.6%, while the accuracy based on ML equals 82%. Besides the case presented in figure 1, the selection accuracy has improved by up to several percents comparing to the basic approach in all tested ML models, for all evaluated district groups. The model developed for all 16 districts also gained better accuracy (82%) than the models for individual groups (successively 78%, 81%, 86%, and 82%).

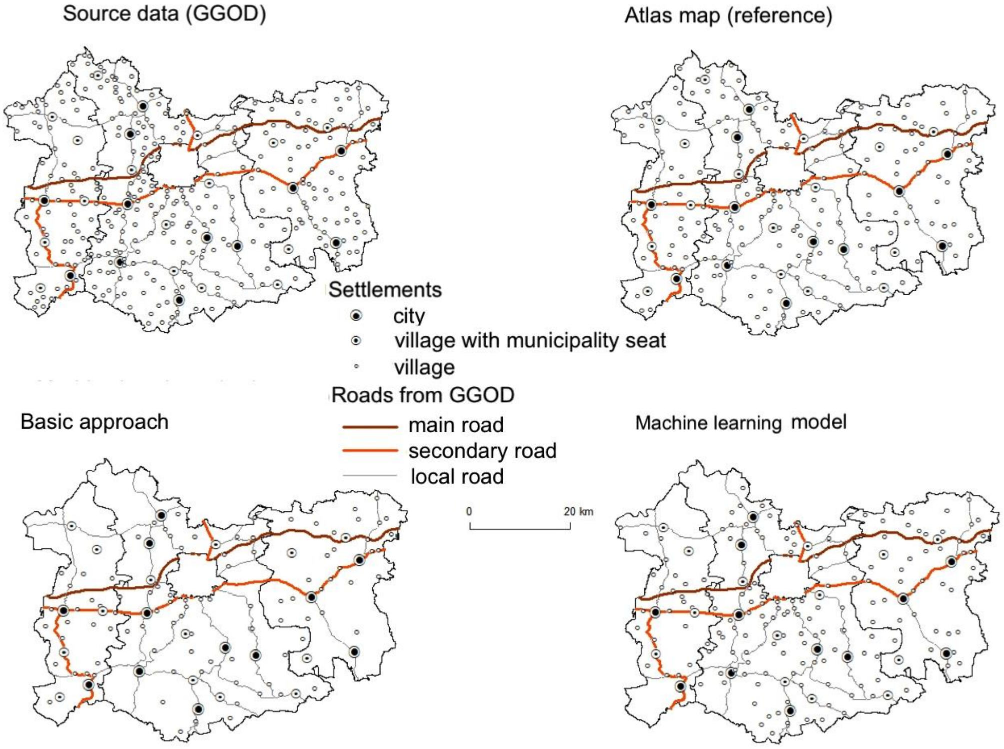

Figure 1. Maps of Brzeski, Tarnowski and Dębicki districts

The evaluation of variables correlation has shown strong correlations of variables that are interrelated (e.g. the presence of industrial facilities and the industrial land area). The strongest correlating variables are the commercial function, built-up area, industrial function and the number of inhabitants. The least-correlated features are the area of Voronoi diagrams, the presence of the airport, the number of roads crossing the settlement. This shows the importance of the variables related to the density of the settlement network and the presence of special objects, like the airports, in the selection process.

The study aimed to propose new variables to fulfil the knowledge gap in the selection algorithms for small-scale maps. The approach assuming data enrichment and ML has been extended to include more significant variables as well as the variable correlation analysis. The ML models built in 4 groups of districts showed that different variables are crucial for selection depending on the region. The obtained selection accuracy in each tested case was better than the selection in the basic approach. The fact that accuracy does not reach 100% means that further work on optimizing the settlement selection ML-based models is recommended. It should also be noted that the goal was not to achieve a complete reconstruction of the manual cartographer’s work, because manual map design process is subjective and may differ according to engaged map designer. The authors’ goal was to automatically achieve the results that would be optimal, acceptable from the cartographic point of view and possibly nearest to the manual map design.

*This work was supported by the Faculty of Geography and Regional Studies at the University of Warsaw.*