May AI Help You? Automatic Settlement Selection for Small-Scale Maps Using Selected Machine Learning Models

I. Karsznia* [a], R. Weibel [b] and S. Leyk[c]

[a] Department of Geoinformatics, Cartography and Remote Sensing, University Of Warsaw, Warsaw, Poland
[b] Department of Geography, University of Zurich, Zurich, Switzerland
[c] Department of Geography, University of Colorado Boulder, Boulder, USA
* i.karsznia@uw.edu.pl

Keywords: artificial intelligence, machine learning, deep learning, settlement, small scales, cartographic generalization

Introduction

Cartographic generalization is a complex decision-making process which aims at creating less detailed, smaller scale maps according to their purpose and taking into account user requirements and constraints. Essential aspects of cartographic generalization constitute modeling and abstracting geographic information while maintaining or even emphasizing important relations and patterns existing in the data (Mackaness et al. 2017). To make this process efficient and automatic the deep knowledge encapsulated in well-designed maps and human cartographer knowledge need to be made explicit and implemented in the software supporting map generalization (Müller and Mouwes 1990; Karsznia and Weibel 2018). Due to the recent advancements in data science and artificial intelligence, research on map generalization has become very complex, with promising applications of machine learning-based frameworks (Feng et al. 2019, Touya et al. 2019, Karsznia and Sielicka 2020, Courtial et al. 2020, Zheng et al. 2021, Courtial et al. 2022). However, most of the research concerns large scales and existing solutions are mainly implemented in raster mode, with machine learning models imported from computer vision, treating maps as images, rather than using the vector format commonly used in cartography to represent maps. Thus, this study aims to contribute to filling this research gap, extending the toolbox for small-scale mapping in vector mode.

The research described in this paper aims at the implementation, verification and comparison of selected machine learning models as potential tools for automated settlement selection to create an optimal design for small scale maps. We test deep learning (DL), random forest (RF), decision tree (DT), and decision tree supported with genetic algorithms (DT-GA) models to automatically classify settlements through selection and omission steps, respectively. We evaluate the different models by comparing the results against the selection status acquired from an atlas reference map. The obtained selection accuracy (i.e., the agreement of the selection status of settlements with the atlas reference map), was very high, ranging from 78% (DT) to nearly 84% (RF).

Method

Source data

The source data used in this study is composed of the settlement layer contained in the General Geographic Object Database (GGOD) at the scale of 1: 250 000, while the target scale for generalization corresponds to 1:500 000. In this research, 16 Polish districts (level-2 administrative division units) were used as test areas (5% of all districts in the country). The districts were differentiated in terms of population density, settlement density and settlement size structure. The structure of the GGOD is relatively similar to the general databases of other countries. Like those, GGOD is rich in geometry but poor in semantics. Thus, to make the database an appropriate source for cartographic generalization, it was decided to enrich the settlement layer with additional semantic and spatial attributes, serving as features in the ML process. The list of considered features included 18 semantic and 15 spatial attributes. The semantic features constituted among others the number of settlement functions, administrative status, features describing relation to other objects (i.e., distance to roads, railroads). The spatial features included among others the area of the settlement (i.e., built-up, residential, industrial area of the settlement), population density, density of settlements calculated using a rectangular and hexagonal grid, Voronoi area as well as the distance to the nearest neighbor. As a reference we used the atlas map at a reduced scale of 1:500,000 (GGK, 1993–1997).

Machine learning

The adopted methodology assumes four main stages: 1) selection of districts; 2) data enrichment with the use of new semantic and spatial features; 3) ML and DL model training; 4) validation and comparison of the research results. To include cartographic knowledge into the automatic selection process, an additional attribute was added to the defined spatial and semantic features. This feature concerns the status of the settlement on the atlas reference map at the 1:500 000 scale, and indicates if an expert map designer would select a particular settlement in the manual map design process. With this feature, cartographic knowledge has been added to the selection process. Using all the features the selection of the settlement was formulated as a binary classification problem with two labels (selected or omitted).

Then, four parallel selection processes — one for each of the selected ML models — were designed. The models were built and executed in RapidMiner Studio v 9.9. The selection processes consisted of developing and applying decision trees supported with genetic algorithms (DT-GA), decision tree (DT), random forest (RF), and deep learning (DL) based on a multi-layer feed-forward artificial neural network trained with stochastic gradient descent using back-propagation models. Evaluation of results was carried out in two steps: validation against the selection status acquired from the atlas map (taken as reference for evaluation), and the comparison of performance statistics across implemented models.

Results

The performance achieved with the various ML models as compared to the atlas map can be found in Table 1. The overall accuracy describes the similarity of the automatic selection results to the manually generalized atlas map. Accuracy is also termed positive predictive value, and it constitutes the fraction of correctly classified settlements (as selected and omitted) among all settlements considered.

Learning model	Overall accuracy in %
Random Forest (RF)	83.27
Deep Learning (DL)	82.76
Decision tree with genetic algorithms (DT_GA)	81.69
Decision tree (DT)	78.67

Table 1: ML models performance for all districts.

To give a visual impression of the selection results, we show two examples of district groups. The first example represents the districts characterizing high population and settlement density (Figure 1), while the second example represents districts with low population and settlement density (Figure 2). These examples illustrate the range of observed model performances.

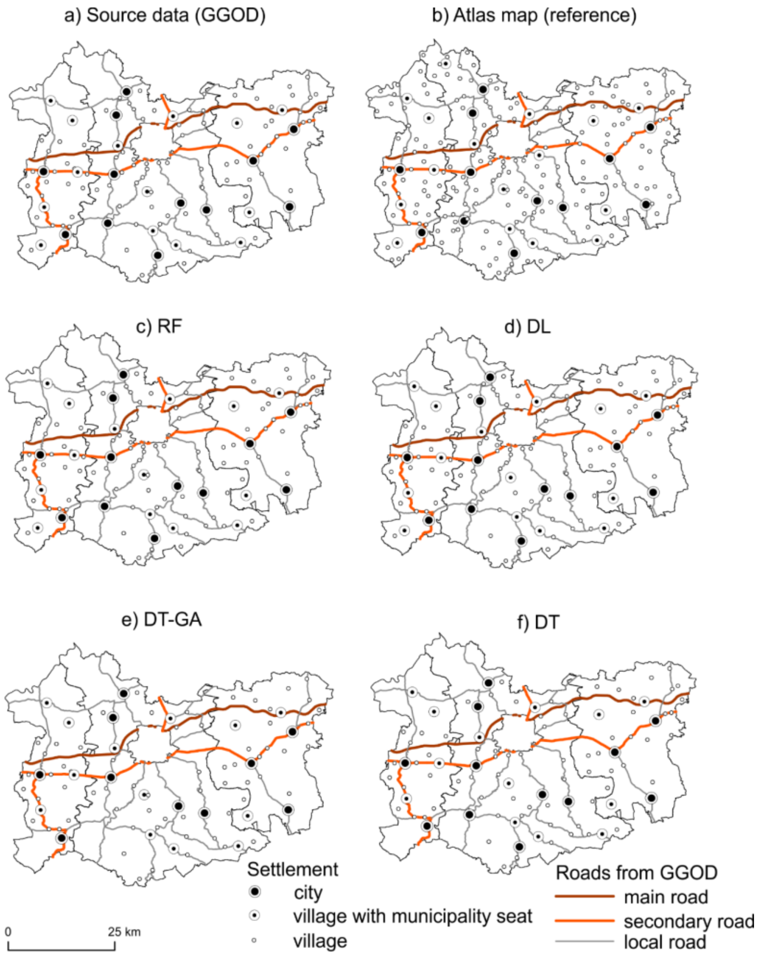

Figure 1: Maps of Tarnowski, Dębicki and Brzeski districts.

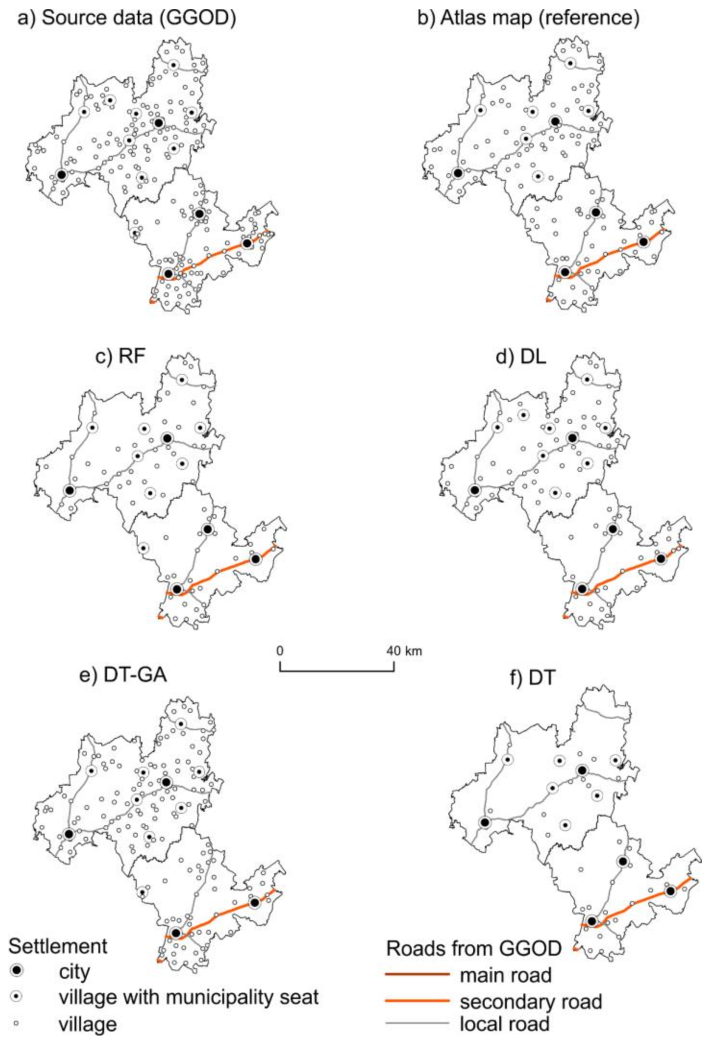

Figure 2: Maps of Chojnicki and Bytowski districts.

Discussion and Conclusion

In this study we aimed to examine the performance of various learning-based models that have the potential to make settlement selection more effective, holistic, and aware of semantic and spatial context. The goal was to find the solutions that would be optimal in the sense that automatically produced results will be as close as possible to the manual map design of a reference atlas map.

Looking at Table 1 we see that the best performing models are RF (83.27%) and DL (82.76%), followed by DT_GA (81.69%) and DT (78.67%). The difference between the best and least performing model is 4.6%, which means that we can get a performance improvement of 4.6% with the use of different learning models. RF and DL models provide the results that are closest to the manual atlas map. Moreover, when we analyse the maps presenting the selection results of all source data and the reference atlas map for the high density (Tarnowski, Dębicki and Brzeski) districts, we can see that RF and DL models performed better than DT_GA and DT, especially in terms of maintaining settlement density (Figure 1). However, the differences among the particular models were not as evident as in the case of the results obtained for low density (Chojnicki and Bytowski) districts (Figure 2). In these cases, we can see that RF and DL results are very similar to each other and also closer to the reference atlas map than the DT and DT-GA results, where DT selects too many settlements and DT_GA selects too few settlements compared to the reference atlas map.

However, the drawback of DL and RF models in comparison to DT and DT_GA models is that when we use DL, we do not get an intuitive and straightforward description of the decision process. Instead, with DL we just get the decision itself with no explanation, and for RF we may get dozens of trees which makes it difficult to understand the features which most heavily influence the decision process or the stages of the decision process. On the other hand, using DT or DT_GA we do not achieve high performance results, but we do generate holistic decision trees for each model.

The goal of this research was to automatically achieve results that would be nearly equivalent to the manual map design. With the use of RF, DL and DT_GA models, this goal has been achieved. Moreover, in our experiments DT in combination with optimized feature selection with the use of GA, namely the DT_GA model, performed quite similar to the RF and DL models. The solutions presented in this study are a further step towards full automation of the selection process for small-scale maps.

Acknowledgements:

This research was funded by the National Science Centre, Poland, grant number UMO-2020/37/B/HS4/02605, “Improving Settlement and Road Network Design for Maps of Small Scales Using Artificial Intelligence and Graph Theory”.

References

Courtial A., El Ayedi A., Touya G. & Zhang X. (2020) Exploring the potential of deep learning segmentation for mountain roads generalisation. ISPRS Int J Geo-Inf 9(5):338. https://doi.org/10.3390/ijgi9050338. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute

Courtial A., Touya G. & Zhang X. (2022). Constraint-Based Evaluation of Map Images Generalized by Deep Learning. Journal of Geovisualization and Spatial Analysis volume 6(13). doi: https://doi.org/10.1007/s41651-022-00104-2

Feng, Y., Thiemann, F., & Sester, M., (2019). Learning Cartographic Building Generalization with Deep Convolutional Neural Network. International Journal of Geo-Information, 8(6), 258. doi: 10.3390/ijgi8060258

GGK. (1993–1997). Atlas Rzeczypospolitej Polskiej. Mapa 1:500 000, Główny Geodeta Kraju, Warszawa [Atlas of the Republic Poland, 1: 500 000 map, The Surveyor General of Poland, Warsaw, 1993–1997]. Warsaw: Head Office of Geodesy and Cartography.

Karsznia, I., & Weibel R. (2018). Improving Settlement Selection for Small-scale Maps Using Data Enrichment and Machine Learning, Cartography and Geographic Information Science, 1-17. doi: 10.1080/15230406.2016.1274237

Karsznia, I., & Sielicka, K. (2020). When Traditional Selection Fails: How to Improve Settlement Selection for Small-Scale Maps Using Machine Learning. ISPRS International Journal of Geo-Information 9(4), 230; doi: 10.3390/ijgi9040230

Mackaness W. A., Burghardt D., & Duchêne, C. (2017). Map generalisation. In: Richardson D., Castree N., Goodchild M.F., Kobayashi A., Liu W., and Marston R.A. (Eds) The International Encyclopedia of Geography, John Wiley & Sons, doi: 10.1002/9781118786352.wbieg1015

Müller, J. C., & Mouwes, P. J. (1990). Knowledge acquisition and representation for rule based map generalization: An example from the Netherlands. In Proceedings of GIS/LIS 90, (pp. 58–67). Falls Church, VA: American Congress on Surveying and Mapping.

Touya G., Zhang X., Lokhat I. (2019). Is deep learning the new agent for map generalisation? International Journal of Cartography, doi: 10.1080/23729333.2019.1613071.

Zheng, J. Gao, Z. Ma, J. Shen, J. & Zhang, K. (2021) Deep Graph Convolutional Networks for Accurate Automatic Road Network Selection. ISPRS Int. J. Geo-Inf., 10, 768. doi: 10.3390/ijgi10110768