Enriching river network data for multi-scale machine learning based selection: a case study of Poland and France

Bérénice Le Mao [a], Iga Ajdacka [b], Izabela Karsznia [b], Guillaume Touya [a], Albert Adolf [b]

[a] *LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France*
[b] *Department of Geoinformatics, Cartography and Remote Sensing, Faculty of Geography and Regional Studies, University of Warsaw, Poland*

Keywords: cartography, machine learning, map generalization, rivers, enrichment, multi-scale

Abstract:

Cartographic generalization constitutes a challenging decisive process of meaningfully removing details from the map. The first and key task in cartographic generalization is selection, also referred to as elimination, that among other generalization operators affects objects' visual quantity. Selection deals with one or more objects or object classes removal without replacement, thus it is used to reduce map or database content according to the target detail level. However, this abstraction process can be a demanding task, particularly when aiming to automate generalization across multiple scales.

In this respect, we want to show in our research how efficient DT techniques are to reveal additional rules for river selection and question their relevance for automatically generating multi-scale generalizations. Yet, to truly unlock the potential of machine learning, it's crucial to meticulously prepare the data beforehand. This step is sometimes long and time-consuming but is nonetheless essential for the model's optimal performance. This research therefore focuses on the preliminary stages of machine learning-based selection, which includes data improvement and enrichment stages as well as preliminary results of machine learning. The scope of this research covers the use of two diverse data samples to test the replicability of the model: river basins in Poland and France. In Poland, the chosen watersheds in the General Geographic Objects Database (GGOD) are Biebrza and Radomka, diverse in terms of watershed structure characteristics. In France, the chosen watershed in the BD Topo is that of the Adour, covering an area of nearly 17,000 km², to provide ample room for multi-scale generalization. The choice of the datasets is driven by the fairly diverse presence of river patterns, which can allow us to determine if certain patterns are better generalized than others. However, we faced many challenges with the data, requiring significant data improvement efforts to prepare these datasets properly. We can divide the process into four categories, which we will then elaborate on: the data matching step, geometric corrections, stroke creation, and data enrichment.

First of all, it is important to consider the differences between the two data sources, Polish and French, in order to work with the same attributes. Some were indeed discarded because they were too specific to one dataset, while others were transformed to match. In river data, geometry plays a crucial role in automatic selection algorithms, as most need to understand the flow direction to determine the hierarchy between watercourses, distinguishing the main river from its tributaries. However, when the linear features are digitized in databases, like in the GGOD or BD Topo, errors may occur and distort the results, which requires manually correcting the flow directions. Moreover, directional errors stem sometimes from previous generalization algorithms, as is the case with Polish data originating from pre-generalized databases for small scales (GGOD).

This initial step is necessary for data enrichment, particularly for generating river strokes. To improve the quality of the result, we decided to set the selection on strokes, which are a set of arcs that appear to be continuous (Thomson et Brooks, 2000), rather than on segments. Following the principle of good continuation, we employed the stroke creation algorithm implemented in the Cartagen platform (Touya et al., 2019). This process enabled us to enrich the attribute data with values representing the selections on strokes made by expert cartographers, following multi-scale pattern criteria. We determined whether each stroke should be retained (1) or removed (0) for each zoom level, from 7 to 15. This manual step allows us to evaluate the difference in results between models and our multi-scale 'good generalization', which we believe is progressive.

Apart from the basic attributes contained in BD Topo and GGOD, both datasets were enriched with graph-based calculated metrics. We decided to include a set consisting of betweenness, closeness, degree, and load centrality. The next step after the data improvement and enrichment covered machine learning with the use of the selected machine learning models developed by Karsznia et al. (2024), further adjusted to the river network selection process. The machine learning models included decision trees supported with genetic algorithms (DT-GA) and random forest (RF). Our first results seem to favor

Abstracts of the International Cartographic Association, 7, 83, 2024.
European Cartographic Conference – EuroCarto 2024, 9–11 September 2024, TU Wien, Vienna, Austria.
https://doi.org/10.5194/ica-abs-7-83-2024 | © Author(s) 2024. CC BY 4.0 License.

the DT-GA model on Polish data as it tends to remove a little fewer rivers and quite well preserves river continuity, which was a problem with previous attempts based on machine learning (see Figure 1). We still need to extend the models to French data and evaluate whether they perform as well across different scales, thanks to the cross-countries benchmark datasets we curated.

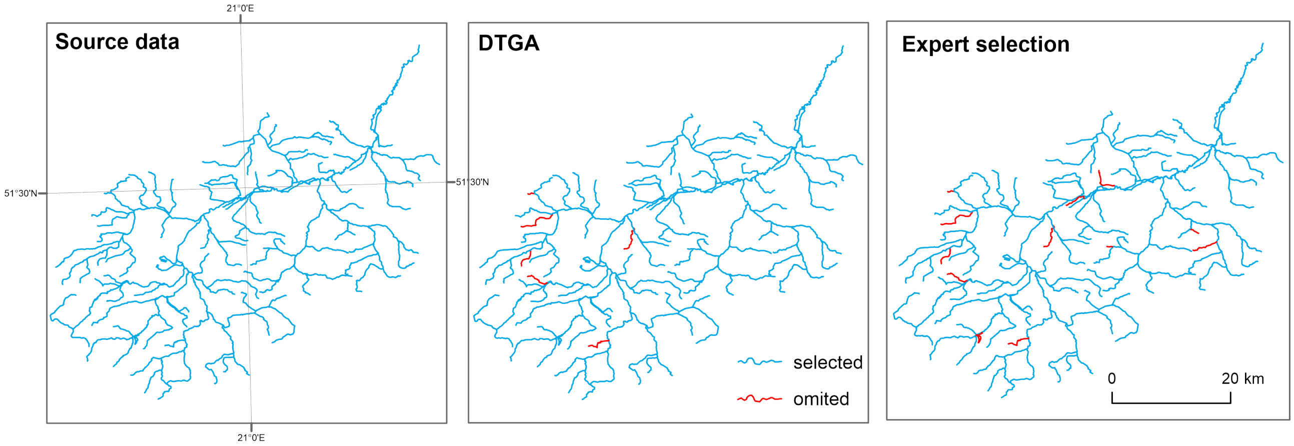

Figure 1. An example of the first automatic selection results on the Radomka watershed in Poland, with the RF model and expert selection compared to source data. River segments marked in red were automatically classified as omitted by the machine learning models.

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.” This research has also received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement N° 101003012).”

References

Karsznia I., Adolf A., Leyk S., Weibel R. 2024. Using machine learning and data enrichment in the selection of roads for small-scale maps. Cartography and Geographic Information Science, 51(1): 60-78. DOI: 10.1080/15230406.2023.2283075.

Thomson, R., Brooks, R.: Efficient generalisation and abstraction of network data using perceptual grouping. Inproceedings of the 5th GeoComputation. University of Greenwich, Kent U.K. (2000)

Touya, Guillaume, Imran Lokhat, et Cécile Duchêne. « CartAGen: An Open Source Research Platform for Map Generalization ». Proceedings of the ICA 2 (10 juillet 2019): 1-9. https://doi.org/10.5194/ica-proc-2-134-2019.