International Cartographic Association
Association Cartographique Internationale

Lessons learnt from digitizing a label dataset for AI-based feature extraction of watercourses

Justus Poutanen*, Christian Koski, Juha Oksanen, Pyry Kettunen

Finnish Geospatial Research Institute (FGI) in the National Land Survey of Finland (NLS) – {firstname.lastname}@nls.fi

Keywords: digitization, training data, label, machine learning, hydrography, topographic data

Introduction

As data from improved remote sensing techniques, for example lidar, become more and more accessible, topographic features can be mapped more accurately than with traditional methods, such as stereo mapping. Topographic data available from the National Mapping Agencies (NMAs) is still rarely accurate enough to be used as label data for machine learning tasks due to many reasons. Firstly, the manual mapping process is prone to human errors in addition to technical limitations of the underlying stereo models. Secondly and even more importantly, the intended use of the topographic data was never label data for machine learning. Topographic datasets by the NMAs are typically tied to a specific maximum nominal scale level, and even at that level certain generalisation rules come into play.

The challenge of digitizing natural phenomena, like hydrographic features, is that the classification of nature is inherently arbitrary (Goodchild and Gopal, 1989: 45-46*)*. There are large variations in shape, vegetation, fulfilment of matter and meandering, even within small watercourses. In addition, available vector data in topographic databases is still rarely precise enough to be used for machine learning tasks as training data. In this paper, we present challenges that were identified while digitizing a dataset of watercourses less than five meters wide. This dataset serves as the training data for the ongoing development of the mapping process by machine learning in the National Land Survey of Finland (NLS), whose maximum nominal scale level of data collection can be said to be about 1:10 000.

Data and methods

The study area is typical forested Finnish terrain in central Finland, south of Suonenjoki. All watercourses less than five meters wide were manually digitized from an area of 36 km^2. The watercourses were digitized in QGIS along their centre line as precisely as possible by hand and by the resolution of data. In the digitization, five layers were used to highlight and help the interpretation of the watercourses: 1) a DEM at 0.5 m resolution created from the 5 points/m^2 point cloud by the National Land Survey of Finland (NLS, 2022), a DEM-derived 2) hillshade, 3) relative topographic position (RTP) and 3) flow accumulation (FA), 4) NLS Orthophotos, and 5) the NLS Topographic map.

All digitized watercourses were given a type and a clarity class (CC) attribute (Figure 1). Attributes were given to watercourses to be able to filter the dataset later if necessary. The type of the attribute was determined by the shape and features of the watercourse as well as the context of the area. Five different types were used: ditch, road-side ditch, natural stream, culvert and unknown. A clarity class of one to five (1=most clear, 5=least clear) was given as an indicator of how confident the digitizer was in digitizing along the centre line of the watercourse if the feature was identified as a watercourse. If a watercourse changed significantly in CC during a vector segment, it was divided to separate watercourses with their own CCs. All watercourses were digitized by one researcher and checking was made by another.

During the digitizing, notes were made on challenges that occurred. The challenges were further discussed in a group of three researchers to come up with a way forward. After digitizing the features, the challenges were revisited to reflect on how they impacted the final dataset.

Results and discussion

The digitized dataset contains 4035 vector features with a total length of 365 km. The amount of features by type are 2742 of ditch, 398 of road-side ditch, 68 of natural stream, 233 of culverts, and 594 of unknown. Excluding the culverts, the total lengths by clarity class are 43 558 m of CC1, 65 940 m of CC2, 119 254 m of CC3, 74 888 m of CC4, 59 835 m of CC 5 (Koski et al., 2022). Unknown watercourses are ditches or natural streams and the large number of unknowns reflects the difficulty to identify them apart. The most used background data for the digitizing were RTP, hillshade, FA, orthophotos and the topographic map, in order of usage amount. RTP and hillshade were used clearly the most because it was noticed that they highlight even very shallow watercourses from the terrain. RTP highlighted more than the hillshade because the RTP visually exaggerates even the slight elevation changes. FA, orthophotos and the topographic

Abstracts of the International Cartographic Association, 6, 205, 2023.
31st International Cartographic Conference (ICC 2023), 13–18 August 2023, Cape Town, South Africa.
https://doi.org/10.5194/ica-abs-6-205-2023

map were used mostly for context recognition, such as broader information of the area, if the watercourses were difficult to identify.

Classifying watercourses into CCs affected the CC of the possible nearby watercourses because similar features next to each other could get the same CC: if there is doubt in classifying watercourses the CC will likely be the same as surrounding ones. Determining the CCs was often straightforward even though there were challenges. The scale from 1-5 may have been too fine and a scale from 1-3 might have been sufficient enough as well as more efficient for digitizing.

Figure 1. RTP, hillshade, and hillshade with digitized watercourses coloured according to different CCs. CC1: dark green, CC2: light green, CC3: yellow (a few pixels), CC4: orange, CC5: red.

Challenges for digitization include overgrowth, natural streams, cliffs, roads and marshlands. Overgrowth can seem to fill a watercourse, making identification of type and CC difficult. Natural streams are challenging to digitize due to constant erosion and meandering. Cliffs create challenges in digitizing because they obscure watercourses in RTP and hillshade layers due to elevation change. Roads are also often elevated from the terrain and thus create similar challenges by obscuring the potential watercourses running alongside them. At marshlands, watercourses might not be visible even on-site because of the possible changes in water level and shallowness of watercourses in them. Other challenging aspects are flaws in data, temporality and halts: possible flaws in the point cloud data can distort the visual inspection. Temporality can be a significant factor in the formation of a watercourse; especially natural streams that might seem wide in certain sections. Sudden halts in watercourses with no obvious culprit leave the digitizer without confidence. In general, manually digitizing and classifying natural phenomena is very challenging because every feature is unique with their own context. Digitizing all watercourses in the study area took weeks to complete and output vector features are not without mistakes. Features can be off from the central line or in wrong CCs because of the adaptation of the digitizer's perception to the current areal context: manual digitizing lowers correctness of clarity classification over time between different parts within a large study area. However, the output of manual digitizing became highly sufficient for future studies in the area. There should be some research done on how to better highlight watercourses next to steep but not necessarily major elevation change like roads or ridges.

Conclusions

Digitizing watercourses and classifying them to clarity classes are essential steps for their use as training data in automatic machine learning recognition. However, the digitization has many challenges. Often the classification of a watercourse's clarity class can be difficult and the correct clarity class is rarely obvious. Temporality creates many difficulties in clarity class determination, especially with natural streams. Watercourses next to elevation change are very challenging to notice from the data without field surveys. Even decisions to include some watercourses can be challenging due to fuzziness and field checks show that faint vector features seen in high resolution DEMs can be insignificant. Human error is present in the digitization and without real ground truth data and experience from the field, watercourses are challenging to digitize and classify.

References

Goodchild, M. and Gopal, S. 1989. *The Accuracy of Spatial Databases*. Taylor & Francis, London.

Koski, C., Kettunen, P., Poutanen, J. and Oksanen, J., 2022. Mapping small watercourses with deep learning – impact of training watercourse types separately. *AGILE GIScience Series*, 3, 43. https://doi.org/10.5194/agile-giss-3-43-2022

NLS, 2022. Laser scanning data 5 p. https://www.maanmittauslaitos.fi/en/maps-and-spatial-data/expert-users/product-descriptions/laser-scanning-data-5-p

Abstracts of the International Cartographic Association, 6, 205, 2023.
31st International Cartographic Conference (ICC 2023), 13–18 August 2023, Cape Town, South Africa.
https://doi.org/10.5194/ica-abs-6-205-2023 | © Author(s) 2023. CC BY 4.0 License.