Report of the Breakout Session of the 24th ICA Workshop on Map Generalisation and Multiple Representation

report by Azelle Courtial and Guillaume Touya

During the breakout session, there were 15 people on site 4-5 people online. The topic of the discussion was quite similar to the one of the workshop: what is next with benchmarks? This report tries to summarize the discussions of the breakout session.

Definition of a benchmark

The discussion started with this question: What is a benchmark and why is it useful for researchers? **A benchmark is a dataset + a reference or target generalisation + an evaluation method**. The arguments to develop their use in map generalisation are:

- the need to enrol more people, for instance with machine learning-oriented benchmarks.
- it highlights important and diverse generalisation problems for researchers to work on.
- it promotes the reproducibility of map generalisation research.
- it adds objectivity in or subjective evaluations by allowing more comparison with the same data, objectives, and evaluation methods.

Who benefits from the use of benchmarks? According to the discussion, researchers do benefit with a better exposition of their work. National mapping agencies (NMA) also benefit from the benchmark as they can compare existing solutions. They can even release their own benchmarks. Finally, software designers also benefit from the use of benchmarks because they can calibrate their tools on the benchmarks.

Regarding comparisons and evaluation, an interesting remark is that there is not one good solution for a generalisation problem, cartographers never agree on the best result. So instead of ranking methods, as benchmarks usually do, our benchmarks could rather exhibit the qualities of each method registered to the benchmark.

Which map/tasks to promote in a benchmark?

It was noted that only topographic generalization benchmarks were proposed to the workshop. Maybe because topographic maps are more formalized / there is an easier evaluation. Some participants asked if benchmarks are relevant for thematic maps? The commission chairs answered yes, as it would be in line with their claim that map generalisation should be more used with thematic maps. And topographic maps often are the background of thematic maps.

Other questions were raised during the discussion: do we need benchmarks specialized for a theme? What is the range of applicability of a benchmark? Once again, the answer is yes after discussion. For instance, three of the five benchmarks proposed to the workshop are specialised in one theme (road selection, relief lines, river simplification).

The participants also discussed the best dataset to put in a benchmark. Some proposed to pick several interesting places in the world and gather them into a benchmark. There is a bias of landscape preference in map generalization evaluation. How do we manage this bias during the benchmark creation? Others proposed the creation of a fictional map, where we can put all the interesting generalisation problems we want. This proposition is interesting but requires a huge data acquisition time.

About the evaluation methods used in benchmarks

The proposed benchmarks often include a reference target to measure a metric of difference with. Is it really a scientific method as the target is subjective? Target is not the absolute best map and similarity is expected, not the exact same result. This is the case with the road selection benchmark proposed during the workshop.

What about computation time and scalability? When it comes to NMA and software designers expectations, these two criteria become prominent when comparing two generalisation algorithms.

When a constraint level of satisfaction is used as the evaluation of a benchmark, the mean is not a good synthesis. => how to give more importance to the worst constraint satisfaction? Is it relevant to count unsatisfied geographic objects? If the evaluation is computed for a sliding window over the test area, as proposed in some of the presentations in the workshop, the size of the windows has to be a balance between computation time and relevance.

The presentation by Barry Kronenfeld (https://www.abstr-int-cartogr-assoc.net/4/2/2022/) proposed a nice visual presentation of the evaluation measures, quite similar to an economics diagram or a doughnuts visualization of constraints. Fuzzy metrics can also be used with this visual synthesis of metrics: "somewhere between here and here is correct".

Conclusion

There are still many questions left for the use of benchmarks, as map generalisation is much more complex to assess than the computer vision problems that benefit a lot from benchmarks right now. However, we think that benchmarks should be tested in practice. To start the process, a webpage will be set up to promote the first benchmarks on the commission website, and we hope that the benchmarks will be improved little by little.