Optimising geospatial vector data matching algorithms for change detection with synthetic data

Paul Guardiola,[1] Juste Raimbault,[1] Ana-Maria Olteanu-Raimond,[1] Julien Perret[1]

1. LASTIG, Univ Gustave Eiffel, IGN-ENSG

In order to plan and manage future territorial systems in a sustainable way, an understanding of past dynamics is necessary, to understand emerging patterns in such complex systems and provide insights to build simulation models [1]. Change detection is a basic way to quantify territorial dynamics between successive time steps. Machine learning is widely used for that purpose [2], generally on raster data, and deep learning provides high performance for example to quantify land-use change [3]. However, in many cases it is important to proceed to change detection on geospatial vector data, such as cases where additional data in attributes has to be used, where the curated dataset holds much more information than remote sensing, or where historical data was consolidated into this format, for example. Change detection in vector data is however not a trivial problem, since the introduction of noise or change in data specification, which occur often in practice, can render a basic method of intersecting features totally inefficient. Specific algorithms for matching geospatial vector datasets are a tool for that purpose [4]. These algorithms have however a certain number of parameters which have to be tuned for an optimal performance given ground truth validation data. The choice of the algorithm itself, and of the many parameters, are left to expert appreciation, since a systematic optimisation of parameters remains rare due to the lack of ground truth data.

This contribution proposes to tackle the lack of such ground truth validation data, which is in practice very time-consume and costly to produce, by using synthetic data generation. Similarly to training data generated synthetically in machine learning [5], our work considers synthetic data for simulation models - the matching algorithms [6]. We consider the case of building change detection, which can be directly applied to the quantification of urban dynamics, and consider in this preliminary work two different matching algorithms which are widely used yet difficult to parametrise, namely the Geometric Matching of Areas algorithm [7], and a multi-criteria matching algorithm based on belief theory [8].

We develop a synthetic data generator, which starts from real-world building data, and perturbs the dataset to generate a virtual future state, for which the knowledge of matching links with the original data will be exact and will act as a ground truth. The generator operates a certain number of geometric operations, that were determined by investigating real data (open building data for France, provided by BDTOPO by the National Geographic Institute), including the deletion of some buildings in the reference and generated dataset (for appeared or destroyed buildings), the splitting or merging of buildings, and some noise in point coordinates or in shapes. The generator is tested on different urban morphologies (dense centre, periurban, rural), and both algorithms are optimised on synthetic ground truth datasets for each. Our main result is that optimal matching algorithm parameters depend on the urban context, and that therefore no absolute best method exists. Furthermore, using ground truth datasets (50 times around 100 buildings, for the cities of Strasbourg and Toulouse in France), we find optimal parameters of the generator to generate realistic ground truth datasets, suggesting that this tool can be used at a larger scale in the future. Future work will include extending this work to other European countries, to see to what extent the behaviour of our generator is not specific to the French urban geography, to test other matching algorithms, and to compare change detection results with machine learning results on remote sensing images.

While our results are quite specific to geospatial data and the case of building change detection, they provide some insights into the broader question of spatial sensitivity analysis for complex systems [9, 10], since we describe a methodology to generate synthetic data by introducing noise into a real world dataset, a research direction that was remaining to be fully explored in the emerging field of spatial sensitivity analysis [11].

[1] Celine Rozenblat and Denise Pumain. Conclusion: Toward a methodology for multi-scalar urban system policies. *International and Transnational Perspectives on Urban Systems*, 385, 2018.

[2] Junye Wang, Michael Bretz, M Ali Akber Dewan, and Mojtaba Aghajani Delavar. Machine learning in modelling land-use and land cover-change (lulcc): Current status, challenges and prospects. *Science of The Total Environment*, 822:153559, 2022.

[3] Lazhar Khelifi and Max Mignotte. Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. *Ieee Access*, 8:126385–126400, 2020.

[4] Emerson MA Xavier, Francisco J Ariza-López, and Manuel A Urena-Camara. A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys (CSUR)*, 49(2):1–34, 2016.

[5] Sergey I Nikolenko. *Synthetic data for deep learning*, volume 174. Springer, 2021.

[6] Juste Raimbault. Second-order control of complex systems with correlated synthetic data. *Complex Adaptive Systems Modeling*, 7(1):1–19, 2019.

[7] Francis Harvey, F Vauglin, and A Bel Hadj Ali. Geometric matching of areas, comparison measures and association links. In *Proceedings of the 8th International Symposium On Spatial Data Handling*, pages 557–568, 1998.

[8] Ana-Maria Olteanu Raimond and Sébastien Mustière. Data matching–a matter of belief. In *Headway in Spatial Data Handling: 13th International Symposium on Spatial Data Handling*, pages 501–519. Springer, 2008.

[9] Juste Raimbault, Clémentine Cottineau, Marion Le Texier, Florent Le Néchet, and Romain Reuillon. Space matters: Extending sensitivity analysis to initial spatial conditions in geosimulation models. *Journal of Artificial Societies & Social Simulation*, 22(4), 2019.

[10] Juste Raimbault, Julien Perret, and Romain Reuillon. A scala library for spatial sensitivity analysis. In *28th GISRUK conference*, pages 32–36, 2020.

[11] Juste Raimbault and Julien Perret. Spatial sensitivity analysis of social simulation models. In *Conference on Complex Systems 2023*, 2023.