Spatial interpolation of mobile positioning data for population statistics

Anto Aasa*, Erki Saluveer**, Pilleriine Kamenjuk*, Janika Raun*, Jan Šimbera***

* Mobility Lab, department of geography of University of Tartu, anto.aasa@ut.ee
** Positium, Tartu, Estonia, erki.saluveer@positium.com
*** Department of Applied Geoinformatics and Cartography, Faculty of Science, Charles University, Prague, Czechia, simberaj@natur.cuni.cz

Abstract. Mobile positioning data has been mentioned in many agendas as a new input for official statistics. In current paper, we compared four different spatial interpolation methods of mobile positioning data. Best results to describe the population distribution appeared with adaptive Morton grid model where the R^2 was 0.95. Widely used point-in-polygon and areal-weighted interpolation gave much weaker results ($R^2 = 0.42$; $R^2 = 0.35$).

Keywords. Mobile positioning, Population statistics, Spatial interpolation

1. Introduction

Many agendas and strategic plans are mentioning mobile positioning data (MPD) as a potential new smart data source to produce official statistics and enhance data-driven governance. MPD are used more and more to study the placement and mobility of population (e.g. Ahas et al., 2010, Deville et al., 2014). However, MPD also introduce problems, uncertainties and representativeness issues related to sample and spatiotemporal accuracy.

Current paper is focusing on the spatial interpolation of MPD, specifically, how to convert data from discrete antennae locations to meaningful spatial units (i.e. administrative units) or grid. We compare four spatial interpolation methods and extrapolate the data to the general population, which allows to evaluate the goodness of different methods by comparing the results with census data.

LBS 2019
15th Conference on Location Based Services
Vienna, 11–13 November 2019

Published in "Adjunct Proceedings of the 15th International Conference on Location Based Services (LBS 2019)", edited by Georg Gartner and Haosheng Huang, LBS 2019, 11–13 November 2019, Vienna, Austria.

This contribution underwent double-blind peer review based on the paper. https://doi.org/10.34726/lbs2019.64

2. Theoretical Overview

Mobile phone data have been applied to study human mobility behavior since the early 2000s. In the last decade, the number of studies using MPD has increased significantly. Despite the rapid increase in studies using MPD, there are only few papers that focus on the problems associated using MPD in population research (Williams et al., 2015) and how to minimize them.

Main problems and uncertainties of MPD are related with sample, time, space, differences in data types, and spatial interpolation. In current study, we focus on the spatial aspects of interpolation of MPD. Most common location information of MPD are the discrete locations of antennae which radio coverage areas are presented through Thiessen tessellation. This assumes that signal strengths of the antennae are uniform and do not overlap. However, in real life the situation is much more complicated and the statistics produced with Thiessen polygons or other simplistic interpolation methods (i.e. point-in-polygon) tend to create areas where population count is strongly over- or underestimated.

There are several studies that have tried to overcome the bottlenecks of spatial accuracy of MPD and uneven spatial distribution of network using more complex models or including additional contextual data (e.g. Ricciato et al. 2015, Järv et al. 2017). But they focus on densely populated areas. In addition, extrapolation of results from the level of subscribers of MNOs to the whole population is missing in many cases, and researchers only bring out correlations, trends and/or coefficients, but no real numbers (e.g. Järv et al. 2017). In the following chapters we describe the essence of different spatial interpolation techniques and compare the outputs with census data.

3. Methodological Framework

3.1. Data

Mobile positioning data covers a wide range of different datasets. In the current study, CDR (call detail records) data have been used. It is a set of log-files collected by mobile network operators to collect information about billable calling services used by their clients. CDR data covers the year 2011 and is collected by one of the biggest mobile operators in Estonia. Monthly average number of unique ID-s is approximately 405 000 (ca ⅓ of Estonian population). For interpolation we use the meaningful locations (place of residence, daytime location, etc.) detected from anchor point model developed by Ahas et al., (2010). In addition to MPD, census data from the same year is used to compare the population size in municipalities (n=226).

3.2. Spatial Interpolation Methods

Point-In-Polygon (PIP)

Point-in-polygon method has been used surprisingly often for spatial interpolation. The reason for this is obviously the simplicity of the method: statistics calculated for mobile antennae are assigned to the spatial units (e.g. municipalities) within which specific antenna is located. This method can work only for cases where the density of mobile antennae and the size of spatial units are at the same scale (or spatial units are bigger than theoretical coverage areas of antennae), otherwise, the output contains a lot of units where the value is zero because there are no antennae present in the borders of specific unit.

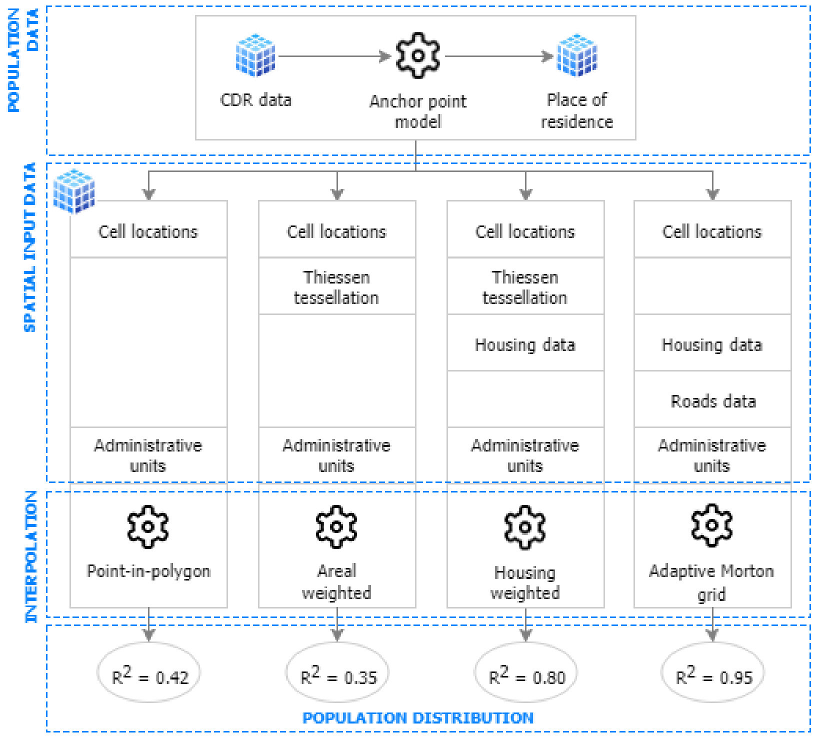

Figure 1. Workflow of spatial interpolation of mobile positioning data.

Areal weighted interpolation (AWI)

Areal weighted interpolation tries to take into account also the radio coverage areas of antennae. Statistics calculated for antennae are assigned to the Thiessen polygons of mobile antennae and those polygons are reaggregated to the level of desired spatial units (e.g. municipalities). The aggregation is based on areal share of theoretical area covering the polygon of municipality (e.g. if 50% of the area of Thiessen polygon covers the polygon of municipality then also 50% of the calculated value is assigned to the specific municipality). Compared with PIP the main advantage of this method is the fact that it avoids white spots and every polygon has a value. At the same time the problem is that in reality the population is not evenly distributed in space.

Housing weighted interpolation

This method works in principle the same as the AWI. The only difference is that the geospatial layers of municipalities and Thiessen polygons are overlaid with the housing layer. The assignment of values is not based on the area, but the number of buildings in specific part of Thiessen polygon.

Adaptive Morton grid interpolation

One possibility to achieve greater accuracy is to use probability surfaces like land use. People are mostly in houses or on roads and much less likely in the forests. In order to use the land use probability surface, it is better to use reference grid by dividing the territory into smaller units. One possibility to achieve a homogeneous layer is to divide the space into grids in the designated coordinate system (Morton, 1966). We are using customized Morton grid system that we call adaptive Morton grid and it is adaptive in a way that the size and the level of the grid is dependent on how many people are probably using it. Probabilities calculations are based on housing and roads densities. In cities, the grid cell size is smaller and in natural areas larger. After we have generated a grid for a country then we will assign all call activities into the grid cells based on indices, and the proportion of coverage area intersecting the specific grid. To estimate the goodness of described interpolation methods (Figure 1), we compared our results with the census data using linear regression models.

4. Results

The results from analysis confirmed initial expectations, explanatory power of point-in-polygon model is rather low (R^2 = 0.42). Biggest problems are in municipalities without any mobile antennae. Areal-weighted model is even worse than point-in-polygon model (R^2 = 0.35). Although the model is able to avoid white spots, the variance of predicted values is even greater than for the PIP method. The results of the housing-weighted model are much better (R^2 = 0.80). Predicted values are considerably closer to the model line and only for a few municipalities, the difference between census and predicted population is strongly biased. In case of adaptive Morton grid, that also takes into account the values of population probability, the results improve even more (Figure 2). In this case, the R^2 is 0.95.

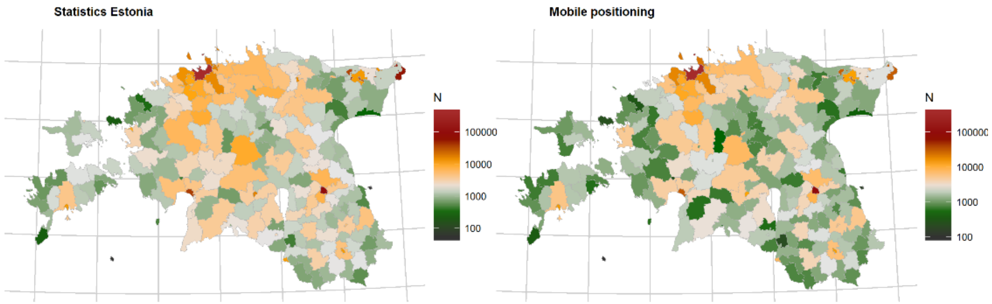

Figure 2. Population distribution according to the Census (left) and mobile positioning data based on adaptive Morton grid method (right).

5. Conclusion

Current analysis demonstrates the weakness of frequently used methods like point-in-polygon and areal-weighted spatial interpolation. We show that using the housing layer improves the output remarkably. Another qualitative leap forward comes with adaptive Morton grid.

6. Acknowledgements

The data used in this study was provided by the Estonian Research Infrastructures Roadmap object „Infotechnological Mobility Observatory" (IMO). Study is funded by Estonian Research Council project PUT PRG306.

References

Ahas, R. et al. (2010) Using mobile positioning data to model locations meaningful to users of mobile phones. Journal of Urban Technology 17(1): 3–27.

Järv, O. et al. (2017) Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation, International Journal of Geographical Information Science, 31:8, 1630-1651.

Ricciato, F. et al. (2015) Estimating population density distribution from network-based mobile phone data. JRC Technical Report. European Commission. Report EUR 27361 EN.

Deville, P. et al. (2014) Dynamic population mapping using mobile phone data. PNAS, vol. 111, no. 45, 15888-15893.

Morton, G. M. (1966), A computer Oriented Geodetic Data Base; and a New Technique in File Sequencing (PDF), Technical Report, Ottawa, Canada: IBM Ltd

Williams, N. E. et al. (2015) Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data. PLoS ONE 10(7): e0133630.