Areal interpolation of spatial interaction data

Jan Šimbera*, Anto Aasa**

* Department of Applied Geoinformatics and Cartography, Faculty of Science, Charles University, Prague, Czechia
** Department of Geography, Faculty of Science and Technology, University of Tartu, Estonia

Abstract. Spatial interaction data, such as commuting flows, are important for many purposes but they often come expressed for a different set of spatial units than required. This happens when comparing data from multiple censuses but especially when cell phone positioning data are involved. This study aims to develop a more accurate method to areally interpolate spatial interaction data to a different set of spatial units and test it on cell phone data-derived commuting flows for Estonia.

Keywords. areal interpolation, interactions, flows

1. Introduction

Data about spatial interactions, such as counts of people commuting between given pairs of places, are important for almost any area of spatially related decision making – they are used to assess demand for social events (Calabrese et al., 2010), model transport network utilization (Bolla et al., 2000), improve disaster response (Bengtsson et al., 2011), study social networks and segregation (Silm & Ahas, 2014) or delimit functional regions to inform administrative divisions (Martínez-Bernabeu et al., 2012).

However, the most fruitful source of spatial interaction data today, cell phone networks, where real mobility behavior of their users is recorded through collection of network traffic data, do not yield data in a readily usable form – they are defined on an unsuitable support, namely that of mobile phone network coverage cells. These usually map poorly to the real settlement network, sometimes with multiple cells covering parts of the same central place as well as outlying rural settlements (see *Figure 1*). To make use of the data, we need to solve the *change of support problem* (COSP)

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/lbs2019.61 | © Authors 2019. CC BY 4.0 License.

and transfer to a different set of spatial units depending on the use case – e.g. administrative units of a chosen level or regular grids.

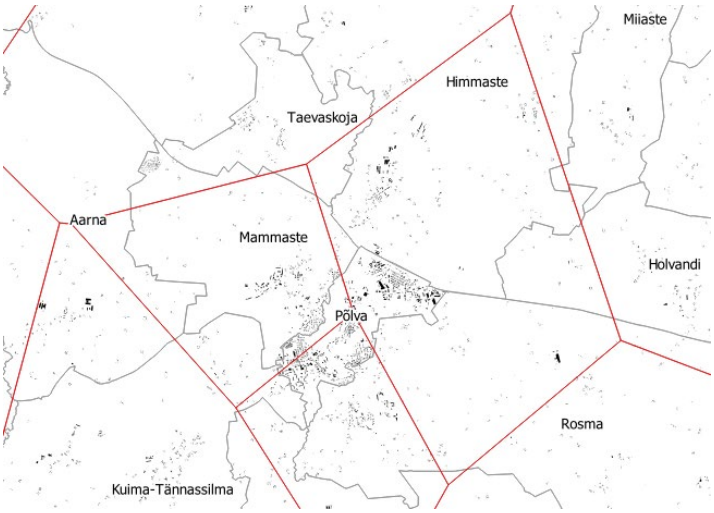

Figure 1. Support mismatch between mobile network cells (red) and territorial communities (kandid; gray) around the Estonian town of Põlva.

COSP is most often encountered for static (non-interaction) data (Gaughan et al., 2015) and solved by *areal interpolation* (Goodchild & Lam, 1980); an example for cell phone stay data was presented by Järv et al. (2017). For interaction data, some limited attempts were made for the purpose of temporal comparison of commuting data across censuses (Boyle & Feng, 2002; Jang & Yao, 2011) due to administrative division changes; however, these methods were primarily suited to small area adjustments, whereas cell phone networks, completely different in both scale and extent of its units, have to the best of our knowledge not been studied in this regard yet.

This article proposes to extend the method from the above studies to account for these differences and presents an example on the Estonian cell phone network.

2. Methodology

2.1. Static areal interpolation

Areal interpolation is a method more widely applied for static quantities such as population densities or various demographic indices. They usually work by computing a *transfer matrix* $T_{i\alpha}$ which determines which fraction

of the value for a given *source* area i (in our case, a mobile phone network cell) is to be transferred to a given *target* area α (in our case, an administrative unit). The matrix is usually zero except for intersecting pairs of units, where the entry is given by a measure of their overlap. The values must be nonnegative and $\sum_\alpha T_{i\alpha} = 1 \; \forall i$ (the *pycnophylactic property* that ensures the sum of all interpolated values remains the same).

In the elementary case (so-called *areal weighting*), the entries of the transfer matrix are given by the fraction of area of the source unit covered by the given target unit. However, the method described below works with any valid transfer matrix, such as one generated by more sophisticated *dasymetric mapping* techniques, which weigh areas differently according to their properties, such as land cover (Gallego et al., 2011) or (more appropriately for our purpose) population density (Monteiro et al., 2018).

The values for the target units v_α are computed using the transfer matrix $T_{i\alpha}$ from the source values v_i:

$$v_\alpha = \sum_i T_{i\alpha} v_i$$

2.2. Spatial interpolation of interactions

For interaction values r_{ij} (which represent e.g. the number of people commuting between source areas i and j), the situation is a bit more complicated. The easiest solution would be to apply the transfer matrix twice, once for the *origins* i and once for the *destinations* j of the interactions:

$$r_{\alpha\beta} = \sum_i \sum_j T_{i\alpha} T_{j\beta} r_{ij}$$

This is the form used by Boyle and Feng (2002). Although Jang and Yao (2011) also investigated more sophisticated approaches such as gravity modeling, they found them less accurate than this one, supposedly because additional complexity brought by those models is already embedded in the data itself.

The method is simple and seems to produce good results overall (provided the transfer matrix is accurate) except for the case of self-interactions (where $i = j$ or $\alpha = \beta$). In the case of commuting, self-interactions represent non-commuters, people that live and work in the same spatial unit. Self-interactions tend to be generally underestimated when a source area is split into more target areas because the weighing equation assigns too much of the originally static activity (self-interactions) to interactions between the target areas, generating artificially high interactions between units covered by the same cell (see *Figure 2*).

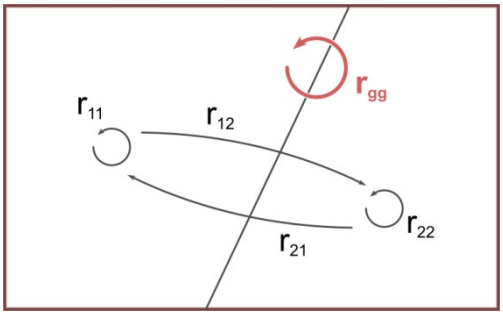

Figure 2. A self-interaction of a source unit r_{gg} is to be redistributed to two target units, both to their self-interactions r_{11}, r_{22} and to interactions between them r_{12}, r_{21}.

2.3. Modification for self-interactions

A slight modification to the above formula can cause it to have more source self-interactions redistributed into target self-interactions. This would be done using a *self-interaction parameter* $\eta \in [0; 1]$ that signifies the fraction of source self-interactions to be a-priori allocated only to target self-interactions:

$$r_{\alpha\beta} = \sum_i \sum_{j \neq i} T_{i\alpha} T_{j\beta} r_{ij} + \sum_i T_{i\alpha} \left[\delta_{\alpha\beta} \eta + (1 - \eta)\, T_{i\beta} \right] r_{ii}$$

where $\delta_{\alpha\beta} = 1 \Leftrightarrow \alpha = \beta$ (self-interaction) and 0 otherwise.

The larger the value of the η parameter, the more the target units will be isolated (more self-interactions of source units will be redistributed to self-interactions of the target units and less to interactions between them). Setting $\eta = 0$ reduces the equation to the simple form from *2.2*, $\eta = 1$ means self-interactions will only be redistributed to self-interactions (not mutual interactions). To satisfy the pycnophylactic property, η has to be constant across any given source area i.

2.4. Estimating the self-interaction parameter

The question is how to get to the value of η that is appropriate for a given settlement system or its individual units (because its value can be varied across different source units according to their characteristics).

A way to get to the value of η using only the data to be interpolated is through simulated aggregation – we merge a few adjacent source units a (two or three as that is generally the amount of units significantly overlapping one source unit) into one (g) and measure what η should be for that breakdown. For this, we use their actual interactions r_{ab} compared to the total aggregated self-interaction $r_{gg} = \sum_{(a)} \sum_{(b)} r_{ab}$, using the relative marginal sums in place of transfer weights from ($T_{ag} = r_{ag}/r_{gg}$). The η_g for the

grouping is then computed as a mean of values generated by the internal interaction matrix, weighted by the absolute interaction values:

$$\eta_g = \sum_{(a)} \sum_{(b)} \frac{r_{ab}}{r_{ag} r_{gg}} \frac{r_{ab} r_{gg} - r_{gb} r_{ag}}{\delta_{ab} r_{gg} - r_{gb}}$$

Then, the question is how to compute η for any set of source areas for which the areal interpolation is attempted. The simple approach undertaken here is to use a single value for η across the whole system. We can obtain its value by a weighted global mean:

$$\eta = \frac{\sum_{(g)} \eta_g r_{gg}}{\sum_{(g)} r_{gg}}$$

A more advanced approach would take into account distinctions across source areas.

3. Validation

3.1. Data

We tested the proposed approach on areal interpolation of commuting interaction data generated from the Estonian cell phone network, which provides information about the home and work anchor points (Ahas et al., 2010) of each mobile network user; commuting interactions were then derived by summing users having home and work anchors respectively in the given pair of mobile network cells.

The interpolation was performed from the level of mobile network cells to that of Estonian municipalities and the results were compared to the census-derived dataset for the comparable period, 2011, which were considered ground truth. For comparison, the interpolated commuting interactions were multiplied by a coefficient to match their overall sums to the census figures; this is necessitated by the fact that the mobile network data only capture a segment of the population according to the network operator's market share.

3.2. Estimating the self-interaction parameter

Using the approach in *2.4*, we computed the η_g values for all pairs of neighboring cells. There seem to be significant differences between the η_g in different areas as depicted in *Figure 3*.

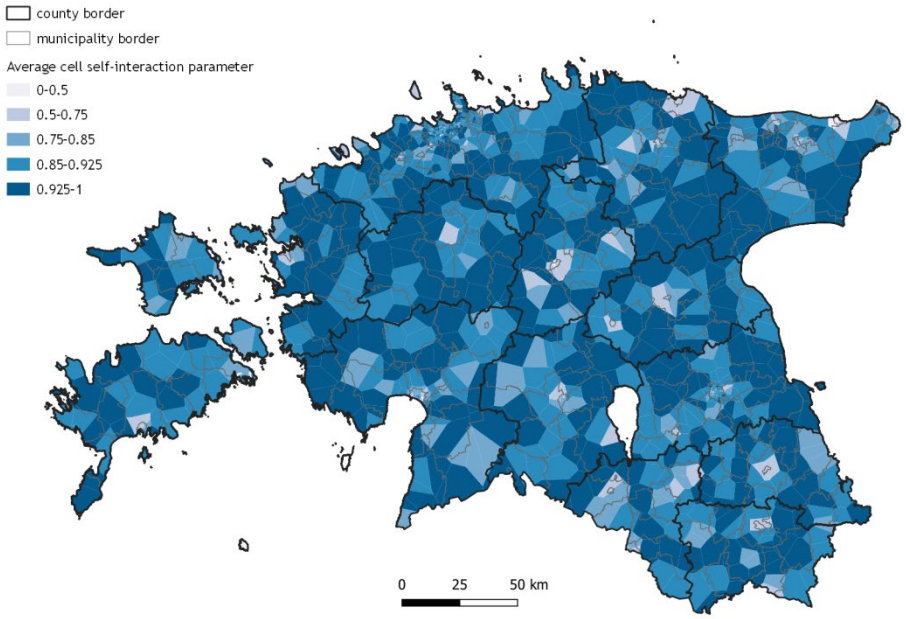

Figure 3. Self-interaction parameter η_g values across Estonian mobile network cells. The value for a given cell is a weighted mean of all its neighbor pairs.

η_g is usually significantly larger in rural Estonian areas than in urban ones, hinting at a lower degree of interaction there. Lower η_g can be found in cells with non-compact shapes that share long borders with neighboring cells whose centroids are close by. Also, lower η_g is observed around mid-sized settlements. Using the global mean estimation method yielded $\eta = 0.865$ with a mean absolute error (MAE) of 0.102.

3.3. Areal interpolation

We examined the effect of different values of η on the interpolated municipal interactions, performing the interpolation with values varying across the $[0; 1]$ range and comparing the result with the census-derived interactions. The share of self-interaction volume in the result increased linearly from 70.2% at $\eta = 0$ to 75.1% at $\eta = 1$. *Figure 4* shows the effect on the correspondence of the interpolated interactions with the census-derived interactions as measured by relative total absolute error (RTAE):

$$RTAE = \frac{\sum_{(\alpha,\beta)} |r_{\alpha\beta} - c_{\alpha\beta}|}{\sum_{(\alpha,\beta)} c_{\alpha\beta}}$$

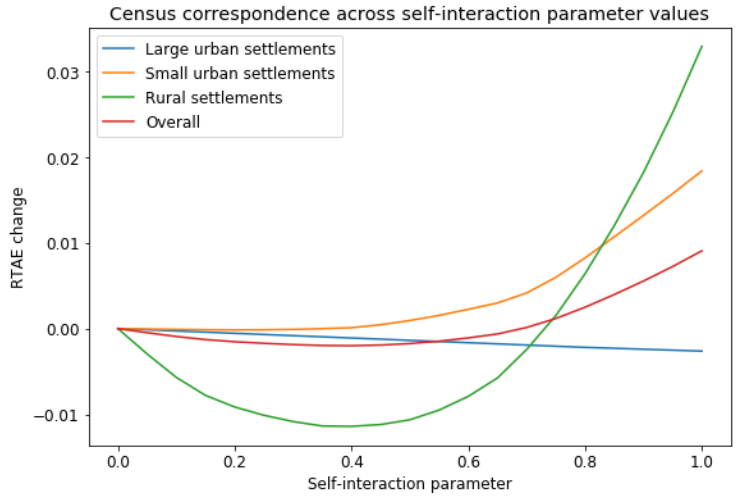

Figure 4. Correspondence of interactions interpolated from source self-interactions with census data, across different settlement classes and self-interaction parameter values.

It seems that overall, greater correspondence with census interactions is achieved at a lower but non-zero value of η in this case, although the differences are rather low. However, within different categories of settlements (rural settlements below 1 000 census commuters, large urban settlements over 20 000 and small urban between these bounds), very different values of η might be appropriate. Therefore, the development of a model that would be able to estimate η for each source area independently could increase the interpolation accuracy further. The global mean value shows itself not to be a suitable estimator in this case, perhaps because it is computed on a higher level (source area aggregations) and therefore tends to minimize the error for larger areas.

4. Conclusion

We suggested an improvement to a commonly used method for areal interpolation of interaction data by Boyle and Feng (2002) with respect to self-interactions that normally tend to produce interpolation artifacts. The method does not require additional data. Using a globally calibrated self-interaction preference parameter η to control self-interaction assignment, a small improvement in accuracy is achieved. An option to increase accuracy further by calibrating the parameter locally is proposed as a further research direction. The method presented here works for spatially extensive variables such as counts of commuters, but it can be easily adapted to spatially intensive variables such as modal split fractions by switching from sums to weighted means.

References

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. Journal of urban technology, 17(1), 3–27.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS medicine, 8(8), e1001083.

Bolla, R., Davoli, F., & Giordano, F. (2000). Estimating road traffic parameters from mobile communications. In Proceedings of the 7th World Congress on ITS, Turin, Italy.

Boyle, P., & Feng, Z. (2002). A method for integrating the 1981 and 1991 British census interaction data. Computers, Environment and Urban Systems, 26(2-3), 241–256.

Calabrese, F.,Pereira, F.C., Di Lorenzo,G., Liu, L.,& Ratti,C. (2010). The geography of taste: analyzing cell-phone mobility and social events. In International conference on pervasive computing (pp. 22–37).

Gallego, F. J., Batista, F., Rocha, C., & Mubareka, S. (2011). Disaggregating population density of the European Union with CORINE land cover. International Journal of Geographical Information Science, 25(12), 2051–2069.

Gaughan, A., Stevens, F. R., Linard, C., Patel, N. N., & Tatem, A. J. (2015). Exploring nationally and regionally defined models for large area population mapping. International Journal of Digital Earth, 8(12), 989–1006.

Goodchild, M. F., & Lam, N. S. N. (1980). Areal interpolation: a variant of the traditional spatial problem. Department of Geography, University of Western Ontario.

Jang, W., & Yao, X. (2011). Interpolating spatial interaction data. Transactions in GIS, 15(4), 541–555.

Järv, O., Tenkanen, H., & Toivonen, T. (2017). Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. International Journal of Geographical Information Science, 31(8), 1630–1651.

Martínez-Bernabeu, L., Flórez-Revuelta, F., & Casado-Díaz, J. M. (2012). Grouping genetic operators for the delineation of functional areas based on spatial interaction. Expert Systems with Applications, 39(8), 6754–6766.

Monteiro, J., Martins, B., & Pires, J. M. (2018). A hybrid approach for the spatial disaggregation of socio-economic indicators. International Journal of Data Science and Analytics, 5(2–3), 189–211. Retrieved from http://dx.doi.org/10.1007/s41060-017-0080-z

Silm, S., & Ahas, R. (2014). Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data. Annals of the Association of American Geographers, 104(3), 542–559.