Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

A possible solution for the centroid-to-centroid and intra-zonal trip length problems

Maryam Kordi
Centre for GeoInformatics (CGI)
School of Geography & Geosciences
University of St Andrews
St Andrews, United Kingdom
mk78@st-andrews.ac.uk

Christian Kaiser
Department of Geography
University of Zurich
Zurich, Switzerland
christian.kaiser@geo.uzh.ch

A. Stewart Fotheringham
Centre for GeoInformatics (CGI)
School of Geography & Geosciences
University of St Andrews
St Andrews, United Kingdom
asf7@st-andrews.ac.uk

Abstract

Average trip length estimation for intra-zonal trip is one of ongoing challenges in spatial modelling. The intra-zonal flows are usually ignored in model calibration, but eliminating these flows can bias model estimates of the parameters obtained during calibration. The existing methods for estimating the average trip length of intra-zonal flows are mostly based on various assumptions related to zone shape and population distribution. The assumption-based models are then applicable only in certain circumstances and they might be highly approximate. In this paper, we suggest a method for estimating the average trip lengths of intra-zonal flows by scattering the origins and destinations of the flows within their zones. The distribution of the origins and destinations of the flows can be done randomly or based on an available spatial density distribution. The average trip length is then calculated for all possible trip configurations. The suggested randomly and density based scattering models are not based on any assumptions and can also provide a solution for the problems with centroid-to-centroid flows in spatial models. The scattered-based models and some existing assumption-based models are applied on a Swiss journey-to-work dataset and the results of the models compared. The comparison of the models reveals that the density-based scattering models have a better model fit and less error.
Keywords: Spatial interaction, intra-zonal distance, average trip length

1 Introduction

Spatial interaction is one of the fundamental types of spatial analysis and usually involves estimating the flows between a set of origin and destination regions in geographic space (see [9, 7, 4]). Spatial interaction models are able to estimate a wide variety of flows such as migration, shopping, commuting, airline passenger traffic and even attendance at conferences, cultural exhibitions and sport events [9]. In each case, the objects making the flows, i.e. people, vehicles, goods or information, can freely start their journey from any location within the origin zone and terminate at any arbitrary location within the destination zone.

A general spatial interaction model can be formulated as:

$$T_{ij} = f(\alpha P_i, \gamma P_j, \beta d_{ij}) \quad (1)$$

where T_{ij} represents the flows between m origins and n destinations. P_i and P_j are origin attractiveness and destination propulsiveness variables, and d_{ij} measures the average spatial separation between the origin and destination regions. This is typically the trip length, generally in terms of distance, time, or some other cost. α, γ and β are parameters of the model to be estimated through calibration.

Finding the average trip length between zones is not always straightforward. Usually this distance is approximated through the centroid-to-centroid travel cost between the origin and destination zones, which might be a gross simplification. In such a case, the flows are assumed to have their origins and destinations at one precise location, within both the origin and destination zones. Considering only the centroids of the regions as origins and destinations of the flows might cause poor estimation of the separation variable in the model especially when the population is not concentrated around the centroid of the zone. Additionally, the centroid-to-centroid distance model is problematic for intra-zonal flows, as the spatial separation for these flows would be zero. However, the average trip length for intra-zonal flows is of course always positive in reality (see [7], page 9).

Several approaches for estimating intra-zonal distances exist but estimating the average intra-zonal trip length is still an ongoing challenge in spatial models. A common approach to avoid the intra-zonal distance problem is to simply exclude the internal flows from the analysis. Sometimes, a different formulation is used to calculate the intra-zonal trip length or a modification is applied on the model itself. In many cases, spatial interaction models are applied to short-distance flows, such as journey-to-work flows. In these cases, the intra-zonal flows can make up

a significant percentage of all flows and omitting the intra-zonal flows from the model might be not appropriate. For instance, in a typical urban agglomeration, many people live and work within the main city itself. If the data are only available at the level of the main city, considering only inter-zonal flows will nearly certainly bias the model calibration.

In this paper we suggest a methodology for estimating the average intra-zonal trip length based on scattering origins and destinations of the flows within the zones. The origin and destination of the flows can be scattered based on a random spatial distribution or by considering a spatial density distribution, such as population density. Our method has no limitations regarding the shape of the geometry of origin and destination zones and does not assume a particular population distribution around an arbitrary point such as the centroid.

The remainder of this paper is structured as follows. Section 2 reviews existing approaches for estimating the average trip length, especially for intra-zonal flows. Section 3 provides some details on relevant existing approaches and introduces our method for computing the average distance between zones. Section 4 shows a case study for journey-to-work flows in the agglomeration of Lausanne, in Switzerland. Computational and implementation issues are discussed. A comparison of our method with existing approaches is made in order to assess the quality of the approach. Section 5 discusses the results and concludes the paper.

2 Previous work

Bharat and Larsen [2] investigated whether ignoring intra-zonal flows in a spatial interaction model represents a feasible solution. Their tests indicate that ignoring the intra-zonal flows biases the parameter estimates of the model since intra-zonal flows are shorter in average and only the longer inter-zonal flows are considered for the model calibration. Ignoring the intra-zonal flows results therefore in a biased data sample.

There are some attempts in the existing literature for estimating the intra-zonal trip length. One of the earliest methods was introduced by U.S. Department of Commerce [14] in which the intra-zonal driving time of a particular zone A is estimated as one-half of the average driving time between the centroid of zone A and the centroids of all neighbouring zones. Venigalla suggested a similar method [15]. In a first step the nearest zone centroid to the centroid of zone A was determined. The intra-zonal trip length for zone A was then computed as half the travel length to the nearest centroid.

A number of suggested models for estimating the average trip length in literature are based on assumptions regarding the geometrical shape of the zone and the internal population distribution. Both Batty [1] and Fotheringham [6] suggested circular-shape based models for deriving the average trip length for zones that are approximately circular. These models are investigated in more detail in section 3.2.

3 Methodology

In this section, we start with a reminder of the spatial interaction model used in this paper and the calibration methodology for estimating the model parameters (section 3.1). In section 3.2, we explain the existing circular-shaped models for intra-zonal distance estimation. Our suggested method for estimating the intra-zonal trip length is described in section 3.3 followed with two subsections explaining both the randomly scattered model and the density-based scattered model.

3.1 Spatial interaction model

Different functional forms of the equation 1 have been formulated for spatial interaction models. In this paper we consider a Poisson gravity model (see [5, 8]) for journey-to-work flows. The probability of a nonnegative integer number of T_{ij} people moving between i and j by our Poisson gravity model is given as:

$$p(T_{ij}) = \frac{e^{-\lambda_{ij}} \cdot \lambda_{ij}^{T_{ij}}}{T_{ij}!} \tag{2}$$

where λ_{ij} is the mean which is logarithmically linked to a linear combination of the logged independent variables. The λ_{ij} can be estimated through following formulation:

$$\lambda_{ij} = exp(k + \alpha \ \ ln\, P_i + \gamma \ \ ln\, J_j + \beta \ \ ln\, d_{ij}) \tag{3}$$

where P_i indicates the active population of the origin i; J_j represents number of jobs in each destination; d_{ij} measures the average Euclidian distance between all possible origins in i and all possible destinations in j; and k, α, γ and β are parameters of the model to be estimated. The Poisson journey-to-work model then is calibrated using maximum likelihood estimation (MLE).

3.2 Circular-shape distance estimates

The intra-zonal distance problem can be seen as a problem in finding the mean trip length within any zone. [1, p. 249]. Batty [1] suggested some models for calculating the mean trip length where one of the simplest is based on the assumption that the zone is roughly circular and the population is spread evenly at a constant density. It is calculated by:

$$d_{ii} = \frac{r_i}{\sqrt{2}} \tag{4}$$

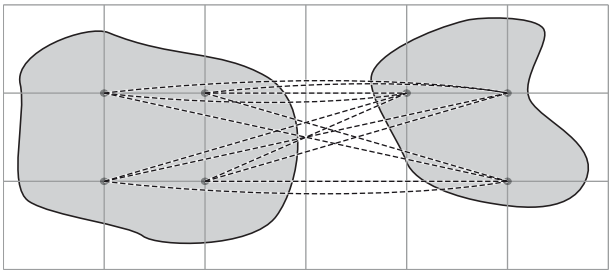
where r_i is the radius of the zone in terms of trip length (travel cost). He also suggested a further variation for this model when the population density varies in a regular way in the zone that can be modelled by a mathematical function, fitted to a particular zone.

The other model based on the circular shape assumption is suggested by Fotheringham [6] where the intra-zonal trip length is estimated with the following formula:

$$d = 0.846 \cdot (1.693)^{z/r} \cdot r \tag{5}$$

where z is the distance between the zone centroid and the destination point and r_i is the radius of a circle whose area is equal to that of the zone. The coefficients in this formulation are related to the potential minimum and maximum distances in a circular zone (see [3] for further details). In the special case of intra-zonal flows, the origin and destination points are both located at

Figure 1: Estimating the average trip length using a regular grid.



the zone centroid, with z being equal to zero. Equation 5 will in this case simplify to:

$$d = 0.846 \cdot r \tag{6}$$

where the coefficient of the equation is very close to one of the equation 4.

3.3 Scattered intra-zonal distance estimates

The circular-shape based estimates in section 3.2 are only a rough approximation and they are more of analytical interest than practical [1, p. 249]. Also the assumptions of the models are rarely respected in spatial models. The polygon representing the origins and destinations often are not circular and population is unevenly distributed within a zone.

3.3.1 Randomly scattered distance estimates

Ideally, the average intra-zonal trip length should be computed using the trip information for all commuters. Generally, these data are not available and the average separation (distance, time, cost) needs to be approximated. To estimate the average trip length d_{ij} between zone i and j, it is possible to scatter the origins of the flows within zone i and destinations within zone j. In the simplest case, the origins and destinations are scattered randomly within the zones i respectively j. It is then straightforward to compute the average length for all flows between the zones. If $i = j$, we obtain the average trip length for the intra-zonal flows.

Finding a random location for both the origin and destination for each flow inside the respective zones is computationally heavy and inefficient. Additionally, due to the random nature of the scattering, each time the computation is done, the resulting average distance estimate will slightly vary. We now present an equivalent but more efficient way of estimating the average trip length.

Scattering the origins and destinations randomly within a zone assumes that the distribution of these points is homogeneous in space. For this reason, we can approximate the average trip length d_{ij} by using a regular grid. For both zone i and j we select the set of grid points G_i and G_j lying within zone i respectively j. We can then estimate d_{ij} simply by computing the average distance between all possible pairs of points from G_i and G_j. Figure 1 illustrates this approach. This method needs the regular grid to be relatively fine. The grid in figure 1 is very coarse and serves as illustration only.

3.3.2 Density-based scattering model

The assumption of a homogeneous distribution of the origin and destination points in section 3.3.1 does not reflect reality in most cases. In the case of journey-to-work flows, we would expect the probability of a flow origin being at a given location i being proportional to the density of active population at i, and similarly for the destination point with the density of jobs.

Instead of selecting a random location for the flow end points within each zone, it is possible to choose the location randomly according to a given probability density. In the case of journey-to-work, this probability density surface can simply be approximated through a high resolution population density surface. Sometimes, no high resolution population density is available. In this case, it is possible to estimate a high resolution density surface based on smaller scale data, using some kind of areal interpolation (see e.g. [13, 10, 11, 12]).

Similarly to section 3.3.1, we can approximate the average trip length using a fine regular grid. In order to take into account the probability density surface, we need to compute a weighted average between all possible pairs of grid points. The considered weight is $w_{ij} = w_i \cdot w_j$, where w_i is the value of the density surface at origin i, and w_j the density surface value at destination j. The average trip length is then computed by:

$$\hat{d_{ij}} = \frac{\sum w_{ij} \cdot d(g_i, g_j)}{\sum w_{ij}} \tag{7}$$

where $d(g_i, g_j)$ is the distance between a grid point in zone i and another grid point in zone j (with potentially $i = j$). The sum is taken over all possible pairs of grid points inside zone i respectively j.

4 Application & Results

In order to test our approach, we apply the method presented in section 3 to a journey-to-work dataset for the agglomeration of Lausanne, Switzerland. The dataset contains the number of commuting flows between the 70 communes of the agglomeration, which is the smallest administrative level. Additionally, both the active population and the number of jobs is known for all communes. The data are available for year 2000 and have been acquired by the Swiss Federal Statistical Office through the population census. The number of jobs has been acquired during the 2001 corporate census. All data are available online[1]. Additionally, for the density-based scattering model (section 3.3.2), the population data are available as a regular grid with a spatial resolution of 100 metres (hectare-level population data)[2].

We have calibrated the Poisson journey-to-work spatial interaction model presented in equation 3, using maximum likelihood estimators (MLE). We have computed different variants for treating the intra-zonal flows and the distance measures:

1. Calibration of the model by excluding the intra-zonal flows, and taking the traditional centroid-to-centroid distance between the zones for the inter-zonal flows.

[1] Journey-to-work flows: `http://www.pendlerstatistik.admin.ch`, active population and number of jobs: `http://www.pxweb.bfs.admin.ch`

[2] See `http://www.geostat.admin.ch` for more information

2. Calibration using all the flows, where the inter-zonal distances are still computed using the centroids, and the intra-zonal distances have been estimated using the circular-shape based model (equation 4).
3. Calibration using all the flows, where both the intra- and inter-zonal distances are computed using the randomly scattering model (section 3.3.1).
4. Calibration using all the flows, where both the intra- and inter-zonal distances are computed using the density-based scattering model (section 3.3.2) with the hectare-level population data as additional density surface.

The parameters of the calibrated model with different average trip lengths are shown in table 1, along with some measures of goodness-of-fit. Parameter α is related to the active population variable and shows a positive influence on the interaction; by increasing the number of active populations, the total trip number or interaction will increase. Parameter γ is associated to the destination propulsiveness variable (in our case the number of jobs in each commune) and has also a positive effect on the interaction; increasing the number of available jobs in a region has a positive effect on the number of incoming flows. Parameter β is the distance-decay parameter and has a negative influence on the number of interactions.
A comparison of these model parameters between the four variants shows that parameters α and γ are similar for all models. The distance-decay parameter β shows a larger variation especially between the centroid-to-centroid model to the other models. The distance-decay parameter in centroid-to-centroid model is -0.668 while in the other three models it is smaller than -1. This shows that ignoring the intra-zonal flows results in a model that is less sensitive to the distance-decay effect, so the model indicates that although the distance-decay parameter has a negative effect on the interaction, people's disutility of distance is underestimated. The circular-shape based model and especially our suggested scattered models, all considering the intra-zonal flows, show a stronger distance-decay effect. This is a clear indication that people see distance as an important criterion for their daily travel to work, and they might prefer to work within their residence region or nearby.

Other comparisons between the different models can be done based on some measurements for goodness-of-fit. We calculated the standard error of estimate (SEE) to measure accuracy of each model with following formulation:

$$SEE = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-k-1}} \tag{8}$$

where $\hat{\varepsilon}$ indicate the estimation error, n is the number of data points and k is the number of independent variables, excluding the intercept. The calculated SEE and coefficient of determination R^2 are listed in the table 1 for all distance measure variants. The R^2 values show a very good model fit in all cases. The reason can be the variable choice and the model type. The variables we have used in this spatial interaction model, i.e. active population and number of jobs, are very precise for a journey-to-work model, and Poisson model gives a better fit in compare with a Gaussian model.

In our journey-to-work model over 45% of the flows are intra-zonal flows. Hence, ignoring these data, the centroid-to-centroid model considers only half the flows compared with the other models. The resulting model is suitable for inter-zonal flows only, while the other models are more general interaction models. A direct comparison of the goodness-of-fit between the centroid-to-centroid model and the other models is not really possible, as the data are not the same. Additionally, the calculation of the SEE is sensitive to total number of flows.

The comparison of the R^2 values of different models considering intra-zonal flows in the table 1 shows that goodness-of-fit in the density-based scattering model is slightly better than the others. Estimating the average trip distance using the population density seems to enhance the spatial interaction model. Between the models taking into account intra-zonal flows, the standard error of estimation is higher for the randomly scattering model compared to the circular-shape based model but it is considerably smaller for the density-based scattering model.

Figure 2 shows the predicted flows versus observed flows for all different model variants. The best fit among the models considering the intra-zonal flows is given by the density-based scattering model.

Considering all the different results and comparison of the models it can be concluded that the density-based scattering model can be a good method for calculating the average trip length and for considering the intra-zonal flows in the model.

5 Discussion & Conclusion

Calculating the average trip length for intra-zonal flows in interaction models is still one of the challenges in spatial modelling. Some of the recent research questioned the procedure of ignoring of the intra-zonal flows in the model as it can bias the model results, especially when a significant amount of flows are intra-zonal. In spatial interaction models where the interaction of objects is modelled in regions, the number of flows within the region can be considerable especially in short-distance flow models, e.g. journey-to-work models. In our Swiss journey-to-work dataset, more than 45% of the flows are intra-zonal showing that a significant number of people work within their commune of residence.

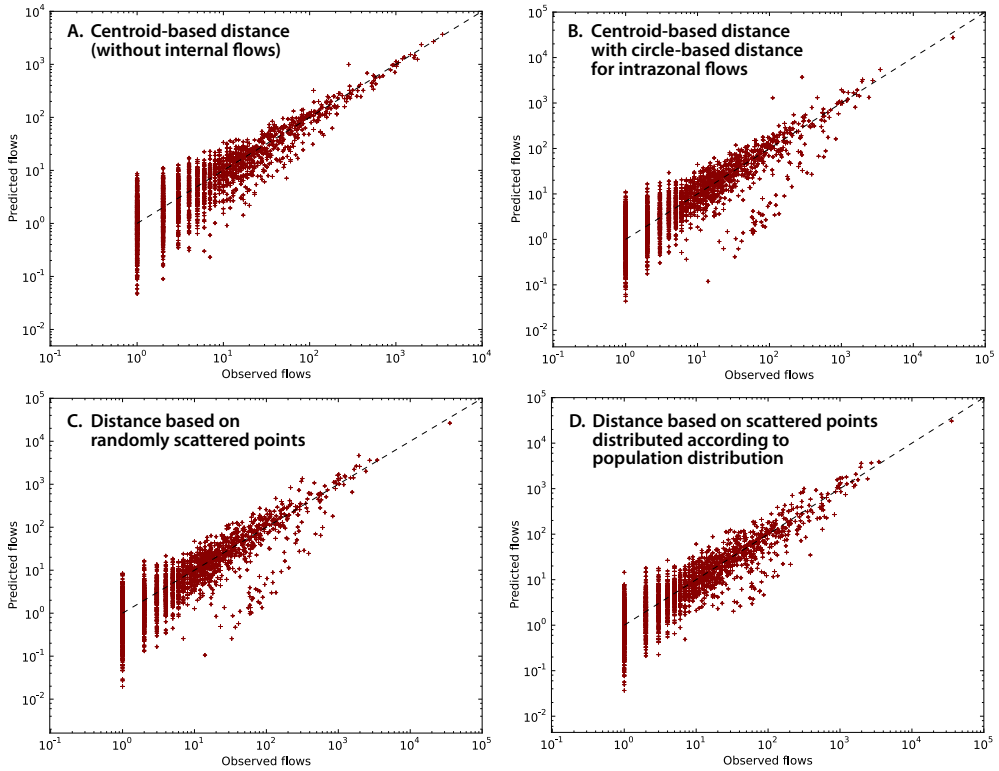
However, although the consideration of intra-zonal flows in the model seems to be necessary, calculating the average trip length for intra-zonal flows can be a problem. There are some attempts in literature to estimate the intra-zonal trip length based on algebraic methods. These models are mostly based on various assumptions, they are mainly highly approximate and more of analytical interest. In this paper, we show a methodology for estimating the average distance for intra-zonal flows. The origins and destination of the flows are distributed randomly or based on an available density surface within the origin and destination zones. This method is not based on any pre-defined assumptions or conditions and it considers most information possible in the model by potentially using the available density such as population density.

The scattered methods based on random distribution or a density surface suggested in this paper are named randomly scattering and density-based scattering method respectively. Application of the methods on the Swiss journey-to-work data and comparison between the models with existing circular-shape

Table 1: Comparison of different models of calculating average trip length (distance).

Distance measure model	SEE	R^2	α	γ	β	$k(intercept)$
Centroids-to-centroid model	19.352	0.965	0.846	1.011	-0.668	-5.700
Circular-shape distance estimates	138.754	0.959	0.862	0.969	-1.017	-2.274
Randomly scattered distance estimates	144.081	0.965	0.945	1.044	-1.317	-0.881
Density-based scattering model	87.688	0.985	0.791	0.949	-1.297	0.818

Figure 2: Predicted flows vs. observed flows for different distance measures.



based methods shows that notably the density-based scattering methods gives a better fit in compare with circular-shape based method. It is interesting to see what proportion of the intra-zonal flows are "real interaction", and how many people work at their home. While in our model the proportion of intra-zonal flows including home workers is 45% of all flows, 18% of all "flows" are from people working at home. It is interesting to note that the latter proportion varies highly across the 70 communes. The city of Lausanne and the other population-rich communes show small proportions of below 10% of home workers, while small villages have typically a relatively high proportion from 30% up to 80% in some cases. The analysis of reasons for this variation can be interesting but exceeds the purpose of this paper. However considering the zero-distance intra-zonal flows in the model and comparing the results with other models might be an interesting topic for future research.

The scattered methods suggested in this paper can also be used for calculating inter-zonal distances, where traditionally centroid-based distances have been considered. Depending on the geometry of the origin and destination regions, and the population distribution within the zones, the resulting inter-zonal distance can be considerably different compared to centroid-to-centroid distance. This is especially true for adjacent zones, for example in agglomerations where the built-up zone stretches over several administrative units. Obstacles, such as rivers, lakes or hills also affect the distance between zones. Currently, the method suggested in this paper does not address this problem directly. However, it could be easily modified to consider travel distance or time instead of Euclidean distance between the set of points of the origin an destination zones. This issue will be considered in future work.

Several other variants for computing the average trip length could also be studied in future work. Sometimes, typically in routing datasets, we might have address points rather than the population density. The distance between zones could then be calculated as the average length of all possible pairs of address

points lying within two zones. Another approach would be to scatter the origins and destinations around the centroid according to some kernel surface, instead of the random scattering approach. Further, the effect of not having a high resolution density surface and taking into account a surface computed using areal interpolation should also be studied. It would be interesting to evaluate to what extent remote sensed imagery and volunteered geographic information, such as OpenStreetMap data could be used for estimating a high resolution density surface.

Further, it is interesting to note that the model considering only inter-zonal flows shows a notably different distance-decay parameter value than the ones taking into account all flows. This might be an indication that the distance-decay parameter might vary in space, or at least among different population groups, and will be studied in future research. The spatial interaction model used in this paper shows a very good fit for our journey-to-work dataset. In future research, it will be needed to consider other datasets commonly used in spatial interaction models to validate our approach in different circumstances.

References

[1] M. Batty. *Urban Modeling: Algorithms, Calibrations, Predictions*. Cambridge University Press, London, 1976.

[2] P.B. Bharat and O.I. Larsen. Are intrazonal trips ignorable? *Transport Policy*, 18:13–22, 2011.

[3] S. Eilon, C.D.T. Watson-Gandy, and N. Christofides. *Distribution management: mathematical modelling and practical analysis*. Griffin, London, 1971.

[4] M.M. Fischer. Spatial interaction models and role of geographic information systems. In A.S. Fotheringham and M. Wegener, editors, *Spatial Models and GIS: New Potential and New Models*, pages 33–43. Taylor and Francis, Philadelphia, 2000.

[5] R. Flowerdew and M. Aitkin. A method of fitting the gravity model based on the Poisson distribution. *Journal of Regional Science*, 22:191–202, 1982.

[6] A.S. Fotheringham. Market share analysis techniques: a review and illustration of current U.S. practice. In N. Wrigley, editor, *Store Choice, Store Location and Market Analysis*, pages 120–159. Routledge, London, 1988.

[7] A.S. Fotheringham and M.E. O'Kelly. *Spatial interaction models: Formulations and applications*. Kluwer Academic Publishers, Dordrecht, 1989.

[8] A.S. Fotheringham and A. Williams. Further discussion on the Poisson interaction model. *Geographical Analysis*, 15(4):343–347, 1983.

[9] K.E. Haynes and A.S. Fotheringham. *Gravity and spatial interaction models*. Number 2 in Scientific Geography Series. Sage, London, 1984.

[10] P.C. Kyriakidis. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3):259–289, 2004.

[11] X.H. Liu, P.C. Kyriakidis, and M. Goodchild. Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science*, 22(4):431–447, 2008.

[12] A. Pozdnoukhov and C. Kaiser. Area-to-point kernel regression on streaming data. In *2nd ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS)*. Association for Computing Machinery ACM, Chicago, 2011.

[13] W. Tobler. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530, 1979.

[14] U.S. Department of Commerce. *Calibrating and treating a gravity model for any size urban area*. U.S. Government Printing Office, Washington, D.C., 1965.

[15] M.M. Venigalla, A. Chatterjee, and M.S. Bronzini. A specialized equilibrium assignment algorithm for air quality modeling. *Transportation Research Part D: Transport and Environment*, 4(1):29–44, 1999.

Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke