Reducing aggregation error in spatial interaction models by location sampling

A. Hagen-Zanker, Y. Jin

Department of Architecture, University of Cambridge, 1-5 Scroope Terrace, Cambridge CB2 1PX, UK
Telephone: +44(0)1223 330573 / 760112
Email: ahh34@cam.ac.uk / yj242@cam.ac.uk

1. Introduction

Models of spatial interaction such as transport, migration, commuting and trade usually partition space into zones, to represent the receiving and sending end of the interaction. When zones encompass multiple locations, the partitioning causes an aggregation error (Hillsman and Rhoda 1978). The aggregation error increases with the size of zones. Aggregation errors can cause bias (Goodchild 1979; Openshaw 1984) and when zones are larger than a (generally unknown) threshold, models become invalid (Tobler 1989). It therefore seems obvious to make zones smaller whenever possible. In practice, however, zones often remain large for a number of reasons, including data availability, parsimony and computational complexity.

There are different aspects to the aggregation error; there is the information loss associated with averaging variables and the loss of spatial precision – typically by conceptually concentrating all of a zone in its centroid. Both types of error are amplified when non-linear functions are applied on the aggregated variables, which can lead to a further model bias. One domain where non-linear use of aggregated variables causes a risk of bias is Discrete Choice Modelling where the utility of an alternative is typically an exponential function of descriptive variables. It is therefore well-recognized that aggregation of alternatives must account for the effect of size and variability of those alternatives. However size and variability are often imperfectly understood and the analysis has to depend on judgment, experience and proxy variables (Ben-Akiva and Lerman 1985 p. 252-275). In recent years (micro)simulation has been established as a method for aggregation that circumvents many of the complications of analytical solutions (Train 2009). The location variation however, is not usually considered in simulation applications. For instance Train (2009 p. 55) suggests that alternatives with a geographical dimension require utility parameters specified in a log function to facilitate analytical aggregation. This paper intends to follow the simulation approach and extent it to the issue of geographical aggregation.

2. Method

The model that will be used to test the approach is a doubly-constrained model of commuting. The general doubly constrained model has the following form:

$$T_{ij} = a_i b_j P_{ij}, \tag{1}$$

where T_{ij} is interaction between origin zone i and destination j, in this case the number of commuting trips. P_{ij} is the prior distribution of interaction from i to j. a_i and b_j are balancing factors, whose values are determined by the constraints respectively at the origin and destination zone. Balancing factors a_i and b_j are chosen such that:

$$R_i = \sum_j T_{ij} \quad and \quad C_j = \sum_i T_{ij} \ , \tag{2}$$

where R_i is the constraint for the i-th row and C_j is the constraint for the j-th column, which also implies $\sum R_i = \sum C_j$. Balancing factors are typically found by iteratively applying the following equations (Fratar 1954):

$$a_i = \frac{R_i}{\sum_j b_j P_{ij}}, \quad b_j = \frac{C_j}{\sum_i a_i P_{ij}} . \tag{3}$$

The prior distribution expresses the 'gravity' nature of the model, it is defined as follows:

$$P_{ij} = O_i D_j e^{-\beta d_{ij}} \ , \tag{4}$$

where O_i is the size of origin zone i and D_j is the size of destination zone j. In the case of commuting, origin size is the working residents and destination size is the number of workplaces. d_{ij} is the distance between zones i and j and parameter β the sensitivity to distance.

The doubly constrained model is linear except for the exponential function of distance. The simulation approach will therefore focus on that function. In the traditional approach the prior distribution is calculated on the basis of mean distance between zones:

$$P_{ij}^{traditional} = O_i D_j e^{-\beta \overline{d_{ij}}} , \tag{5}$$

where mean distance is the distance between zone centroids, with the intrazonal distance being approximated by the 'internal radius':

$$\overline{d_{ij}} = \begin{cases} \|c_i - c_j\| & if \quad i \neq j \\ \sqrt{A_i / \pi} & if \quad i = j \end{cases} , \tag{6}$$

where c_i is the centroid of zone i and A_i is some measure of the land area of zone i.

This paper proposes the following alternative:

$$P_{ij} = O_i D_j \frac{1}{n} \sum_{i=1}^{n} e^{-\beta d_{ijn}} \ , \tag{7}$$

where d_{ijn} is the n-th random sample of distance between locations in zones i and j:

$$d_{ijn} = \| p_{in} - p_{jn} \| , \tag{8}$$

where p_{in} and p_{jn} are random locations within respectively zones i and j. The random locations are drawn from a uniform spatial distribution: a random location in a zone is found by a series of geometrical operations on the polygon that outlines the zones; First the polygon is decomposed into triangles using a dedicated triangulation library (Shewchuk 1996); Next one triangle is randomly selected using the area of each triangle as the weight; Finally a point is found within the selected triangle by applying the algorithm of Turk (1990).

3. Case study and results

The model is applied on commuting data of England as measured by the U.K. Census of 2001 at the level of Standard Table Wards ('wards' from here) as well as Local Authority Districts ('districts' from here). The data used is available from Centre for Interaction Data Estimation and Research (http://cider.census.ac.uk). Wards form the most detailed geography at which Census commuting data is made available. Districts present a more aggregated geographical level at which practical policy analysis is often carried out. There are 354 districts and 7932 wards in England. The digital boundaries (as polygons) of districts and wards come from UK Borders (http://edina.ac.uk/ukborders/),. The centroids of zones are calculated as their geometric centre. Fig. 1 presents the ward and district geographies.

The model has been calibrated twice, with both versions of priors (i.e. equations 5 and 7). A bracketing approach called Golden Section Search (Press 1992) was followed to find the value of β that minimized the following error:

$$\delta = \sum_{i,j} \left(T_{ij}^{model} - T_{ij}^{census} \right)^2 , \tag{9}$$

where δ is the discrepancy between modelled and actual (Census) commuting matrices.

Table 1 gives estimated values for β and the associated error δ. It shows that for the case of wards it makes little difference which approach is chosen, but for districts there is a marked difference in performance where the simulation based model performs 35% better than the traditional model. The graphs in fig. 2 depict the trip distribution as a function of distance and confirm the difference in performance.

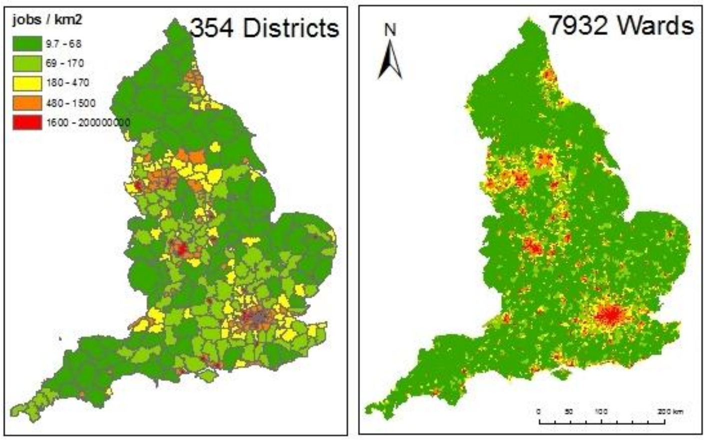

Figure 1. Study area England at district (left) and ward (right) levels of aggregation.

Geography	Model	β	$\delta(*10^9)$
Wards	Traditional	0.34	2.99
Wards	Simulation	0.36	2.87
Districts	Traditional	0.37	90
Districts	Simulation	0.31	58

Table 1. Calibration results and errors. Note that errors are only comparable between models applied at a common geography.

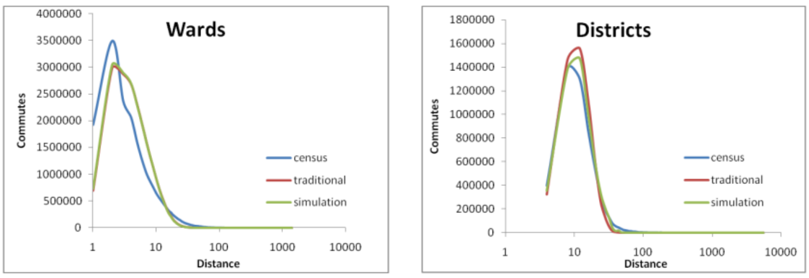

Figure 2. Census and modelled trip distributions. Note zone sizes distort distribution patterns particularly at the district level.

4. Conclusion

This paper follows up on the recommendation of Train (2009) and others to employ simulation when faced with discrete choice models for which analytical models are not feasible or too restrictive. The case study is carried out on the generic doubly-constrained model, which is readily generalisable to more sophisticated random utility models.

By comparing two cases that differ in the level of spatial aggregation it became clear that location sampling does significantly reduce the error caused by using average distances. At the fine scale of wards the effect of error reduction is small although still apparent. At the coarser scale of districts however, simulation would seem essential in future models to contain the aggregation error.

Simulation can be a mechanism for reliable modelling on the basis of coarse scale data when fine scale data is not available. An example of such data is the UK Census commuting data that only offers thematically refined data at coarse spatial scales, for instance commuting patterns specified by industry and socio-economic group which allow segmented modelling of commuter behaviour.

5. Acknowledgments

Geographical boundary data is provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is Crown copyright. Census data (Special Workplace Statistics Levels 1 and 3) is Crown copyright and reproduced with permission of the Controller of HMSO and the Queen's Printer for Scotland. This work is part of the EPSRC Energy Efficient Cities Project.

6. References

Ben-Akiva ME and Lerman SR, 1985, *Discrete choice analysis : theory and application to travel demand*. MIT Press, Cambridge.

Fratar T, 1954, Vehicular trip distribution by successive approximation. *Traffic Quarterly*, 8(1):53-65.

Goodchild M, 1979, The aggregation problem in location-allocation. *Geographical Analysis*, 11(3):240-255.

Hillsman EL and Rhoda R, 1978, Errors in measuring distances from populations to service centers. *The Annals of Regional Science*, 12(3):74-88.

Openshaw S, 1984, Ecological fallacies and the analysis of areal Census-data. *Environment and Planning A*, 16(1):17-31.

Press WH, 1992, *Numerical recipes in C : the art of scientific computing*, 2nd ed. Cambridge University Press, Cambridge ; New York.

Shewchuk J, 1996, Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. In: Lin MC and Manocha D (eds), *Applied Computational Geometry Towards Geometric Engineering*. Springer, Heidelberg; Berlin, 203-222.

Tobler WR, 1989, Frame independent spatial analysis. In: Goodchild MF and Gopal S (eds), *Accuracy of Spatial Databases*. Taylor & Francis, London ; New York, 115-122.

Train K, 2009, *Discrete choice methods with simulation*, 2nd ed. Cambridge University Press, Cambridge ; New York.

Turk G, 1990, Generating random points in triangles. In: Andrew SG (ed) *Graphics gems*. Academic Press Professional, 24-28.