Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques

A.J.Heppenstall[1], K.Harland[1], D.M.Smith[2] and M.H. Birkin[1]

[1]School of Geography, University of Leeds, Leeds, LS2 9JT
Telephone: (+44) 113 343-3392
Fax: (+44) 113 343-3308
Email:[a.j.heppenstall]; [k.harland98]; [m.h.birkin] @leeds.ac.uk

[2] Department of Geography, Queen Mary, University of London
Mile End Road, London E1 4NS
Telephone: (+20) 7882 2750
Fax: (+20) 20 7882 7479
Email: d.smith@qmul.ac.uk

1. Introduction

Recent years have seen a rise in the number of methods and applications which require realistic individual-level data/synthetic populations. This trend can be attributed to a number of factors including increases in computational power and storage, a wealth of individual level data (for example, the British Household Panel Survey) and the development of new computational paradigms, such as cellular automata and agent-based modelling (ABM).

Static spatial microsimulation samples a synthetic population (a population built from anonymous survey data at the individual level) which realistically matches the observed population in a geographical zone for a given set of criteria. There is a diverse set of research and policy applications that use synthetic populations in a spatial setting, including: health (Smith et al, 2009, Tomintz and Clarke, 2008, Brown and Harding, 2002), transportation (see, for example, McFadden et al, 1977; Beckman et al, 1996) and water demand estimation (Williamson and Clarke, 1996).

ABM can also use synthesised data as a base population. There has been a rapid uptake in the use of ABM in Geography with applications ranging from simulating the movement of burglars (Malleson et al, 2009) to replicating dynamics in spatial retail markets (Heppenstall et al, 2006). Although the construction of an ABM does not require a complete individual data set, creating an agent population from a realistic synthesised individual dataset can only improve the realism of these models.

There are several established methodologies for generating synthetic populations. The focus of this paper will be on deterministic reweighting (Smith et al, 2009), conditional probability (Monte Carlo simulation) (Birkin and Clarke, 1988, 1989) and simulated annealing (combinatorial optimisation) (Openshaw, 1995; Williamson, Birkin and Rees 1998; Voas and Williamson, 2000, 2001). These methods were selected due to their common application in geography. Many recent spatial microsimulation studies including Anderson (2007), Ballas *et al.* (2005), Voas and Williamson (2000, 2001), Tomintz *et al.* (2008) Smith *et al.* (2009) and Morrissey *et al.* (2008) have adopted a variation on at least one of the three approaches examined here.

The work within this paper critically compares each approach as they are used to generate a synthetic individual level population at three different spatial scales, extending the initial work reported in Voas and Williamson (2000, 2001).

2. Spatial Microsimulation Algorithms

There are numerous algorithms that have been designed or adapted to produce synthetic populations. Here, three approaches that have commonly been adopted in recent years, each one taking a broadly different methodological approach, are reviewed. The three approaches are deterministic reweighting, a large iterative proportional fitting routine, conditional probabilities, which uses statistical joint probabilities, and simulated annealing, a combinatorial optimisation method.

Table 1 provides a summary comparison of the three algorithms.

	Deterministic Reweighting	**Conditional Probabilities**	**Simulated Annealing**
Easy setup (is there much pre-processing)?	Yes	Yes	No
Sensitive to specification of constraint order?	Yes	Yes	No
Limit to number of constraints that can be used?	Yes	Yes	No
Requires a sample population?	Yes	No	Yes
Can take forward and backward steps to find an appropriate solution?	No	No	Yes
Stochastic?	No	Yes	Yes
Speed of execution	Fastest	Middle	Slowest

Table 1. Summary comparison of the three algorithms.

3. Data and Experiments

Each of the spatial microsimulation methods discussed is used to produce a synthetic population at the Output Area (OA), Lower Layer Super Output Area (LLSOA) and Middle Layer Super Output Area (MLSOA) spatial scales. The synthesised populations are tested against known Census information, produced at all three geographies to evaluate each algorithmic approach. In summary, each population produced will be tested to examine:

(i) Reproduction of variables used to constrain each of the synthetic models at each of the different spatial scales.

(ii) Evaluation of the populations produced against information extracted from the Census of Population 2001 using the constraint variables cross-tabulated against each other.

(iii)Examination of how reliably information from the sample population **not** included in the model constraints can be captured.

(iv)Aggregation of outputs from OA to LLSOA and MLSOA and a subsequent evaluation of the aggregated output against Census of Population 2001 at the appropriate geographical level.

The results of each of these experiments will be presented at the conference.

4. Selected Results

4.1 Representing Constraint Variables

Voas and Williamson (2000) stated that all constraint attributes should be well represented in a synthetic population. The purpose of this test is to evaluate how well the constraint attributes are reproduced in each of the algorithms. Populations are synthesised using each algorithm at each spatial scale OA, LLSOA and MLSOA, making a total of nine different synthetic populations being evaluated. The evaluation statistic used was classification error (CE); this is the total absolute error/ 2.

Table 2 shows that only simulated annealing has successfully recreated all of the constraint attributes at all three spatial scales with zero misclassification. The conditional probabilities algorithm produces a reasonable fit for all of the constraints over each scale. However, the classification error almost doubles for each constraint as the geographical scale becomes finer. The deterministic reweighting method produced the worst fit. With the exception of Highest Qualification (which shows a slight decrease in CE, but overall this constraint has a very poor fit to the observed data) all of the constraints show a slight increase in CE as geographical scale becomes finer.

Constraint	DR		CP		SA	
	CE	**% CE**	**CE**	**% CE**	**CE**	**% CE**
	Middle Layer Super Output Area					
Gender	29,510	4.12	102	0.01	0	0.00
Ethnic Group	14,897	2.08	2,290	0.32	0	0.00
Age	128,999	18.03	144	0.02	0	0.00
Marital Status	95,335	13.33	478	0.07	0	0.00
NSSEC	84,731	11.84	4,378	0.61	0	0.00
Highest Qualification	229,407	32.07	2,569	0.36	0	0.00
	Lower Layer Super Output Area					
Gender	30,297	4.23	176	0.02	0	0.00
Ethnic Group	15,631	2.18	4,010	0.56	0	0.00
Age	131,230	18.34	245	0.03	0	0.00

Marital Status	96,453	13.48	842	0.12	0	0.00
NSSEC	88,282	12.34	9,659	1.35	0	0.00
Highest Qualification	228,425	31.93	5,219	0.73	0	0.00
	Output Area					
Gender	33,430	4.67	245	0.03	0	0.00
Ethnic Group	16,707	2.34	5,292	0.74	0	0.00
Age	135,673	18.96	418	0.06	0	0.00
Marital Status	98,696	13.80	1,828	0.26	0	0.00
NSSEC	95,117	13.30	21,939	3.07	0	0.00
Highest Qualification	227,720	31.83	11,385	1.59	0	0.00

DR = deterministic reweighting, CP = conditional probabilities, SA = simulated annealing

Table 2. Representation of the model constraints in the synthesised populations.

To investigate the poor fit of the deterministic reweighting algorithm, the number of misclassified people per zone is plotted for the Ethnic Group, Gender and Marital Status constraints at the MLSOA geography (fig. 1 - 3). The Ethnic Group scatter plot (fig. 1) shows that, despite having almost 15,000 classification errors, the spread of error tracks the line of perfect fit (where each point would reside if the synthesised population matched the observed population exactly). Only small discrepancies exist, but the discrepancies are evident in many geographical zones.

Fig. 2 shows a scatter plot of gender classification errors which are grouped very tightly together. The lack of spread along the line of perfect fit is a reflection that most geographical zones have a relatively balanced population between male and female and do not display the extremes that can be observed in other constraint attributes. Despite the relatively ubiquitous nature of the attribute, many of the geographical zones are some distance away from the perfect fit line; this is reflected in the 29,510 classification errors observed at the MLSOA geography. This high level of error may be due to the constraint being last in the processing order and the attempt of the algorithm to smooth towards the global mean.

The marital status constraint (fig. 3) is particularly poorly fit by the deterministic reweighting routine. Although this constraint does not have the highest level of associated classification error, it does display a distinct pattern. Most MLSOA zones have the married category over represented and the single category underrepresented in the synthetic population. This suggests that the algorithm is smoothing towards the distribution of the sample population rather than preserving the distribution observed in the constraint information for each geographical area.

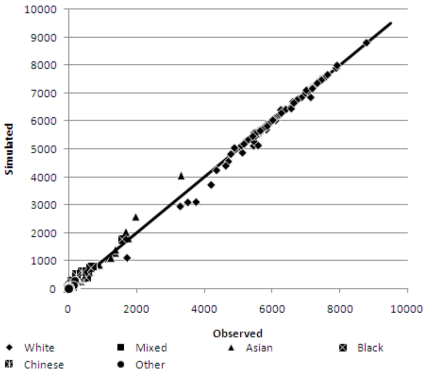

Figure 1. Deterministic reweighting - Ethnic Group misclassification error at MLSOA geography.

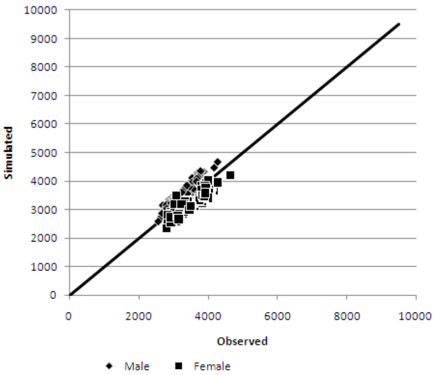

Figure 2. Deterministic reweighting - Gender misclassification error at MLSOA geography.

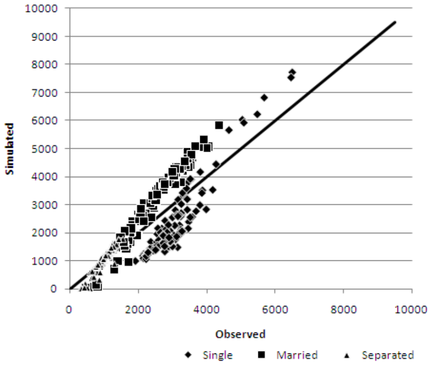

Figure 3. Deterministic reweighting - Marital Status misclassification error at MLSOA geography.

5. Conclusion

The work in this paper has briefly presented selected results of deterministic reweighting, conditional probabilities and simulated annealing spatial microsimulation methods for representing constraint variables at varying spatial scales. Of the three methods assessed, simulated annealing was found to consistently produce the best outcome when fitting constraints. Further conclusions and analysis drawn from the other experiments will be presented at the conference.

6. Acknowledgements

This work forms part of the ESRC funded Modelling Individual Consumer Behaviour project (RES-061-25-0030). Part of this work was funded by a Royal Geographical Society small grants award (SRG 04/09).

7. References

Anderson B, 2007, *Creating small-area Income Estimates: spatial microsimulation modelling*, Department for Communities and Local Government, Communities and Local Government Publications, London

Ballas D, Clarke G, Dorling D, Eyre H, Thomas B, Rossiter D, 2005, "SimBritain: a spatial microsimulation approach to population dynamics "*Population, Space and Place,* 11 13-34

Beckman R J, Baggerly K A and McKay M D, 1996 Creating synthetic baseline populations. *Transportation Research* 30 (6), 415-429

Birkin M, Clarke M, 1988, ``SYNTHESIS - a synthetic spatial information system for urban and regional analysis: methods and examples" *Environment and Planning A* 20 1645 -1671

Birkin M, Clarke M, 1989, ``The generation of individual and household incomes at the small area level using synthesis" *Regional Studies* 23 535 - 548

Brown L, Harding A, 2002, "Social modelling and public policy: Application of microsimulation modelling in Australia." *Jasss-the Journal of Artificial Societies and Social Simulation* 5(4)

Heppenstall AJ, Evans AJ, Birkin MH, 2006, “Application of Multi-Agent Systems to Modelling a Dynamic, Locally Interacting Retail Market” *Jasss-the Journal of Artificial Societies and Social Simulation*. 9(3)

Malleson NS, Heppenstall AJ, See LM, “Simulating Burglary with an Agent-Based Model”. *Computers, Environment and Urban Systems*. In review

McFadden D, Cosslett S, Duguay G and Jung W, 1977 *Demographic Data for Policy Analysis.* Urban Travel Demand Forecasting Project, Final Report Series, Vol VIII. Institute of Transportation Studies, University of California, Berkeley and Irvine

Morrissey K, Clarke G, Ballas D, Hynes S, O'Donoghue C, 2008 “Examining access to GP services in rural Ireland using microsimulation analysis” *Area*, 40(3) 354-364

Openshaw S, Rao L, 1995 “Algorithms for reengineering 1991 Census geography” Environment and Planning A 27 425-446

Smith DM, Clarke GP, Harland K, 2009, Improving the synthetic data generation process in spatial microsimulation models *Environment and Planning A* 41 1251 – 1268

Tomintz MN, GP Clarke, 2008 "The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services." *Area* 40(3): 341-353

Voas D, Williamson P, 2000, ``An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata" *International Journal of Population Geography* 6 349 - 366

Voas D, Williamson P, 2001, ``Evaluating goodness-of-fit measures for synthetic microdata" *Geographical and Environmental Modelling* 5 177 - 200

Williamson P, Birkin M, Rees P, 1998, ``The estimation of population microdata by using data from small area statistics and samples of anonymised records" *Environment and Planning A* 30 785 – 816

Williamson P, Clarke GP, 1996, Estimating small-area demands for water with the use of microsimulation. *Microsimulation for urban and regional policy analysis.* Ed: Clarke, GP. London, Pion 117-148