NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

20-13

Constraint Choice for Spatial Microsimulation

Sandy Burden and David Steel

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

Constraint Choice for Spatial Microsimulation

Sandy Burden *[1] and David Steel[2]

[1] *Research Fellow, National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Australia.*
[2] *Director, National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Australia*

Abstract

Spatial microsimulation models are increasingly being used to create realistic microdata for geographical areas, to enable statistical modelling of health, social and economic variables in a wide variety of application areas. The models combine sample records with benchmark data for pre-defined geographic areas, typically by sampling, or re-weighting sample records to fit a set of constraints for each area. The choice of constraints is a key factor in producing microdata that reflect the population structure.

This paper introduces the use of within-area homogeneity for selecting categorical constraint variables for spatial microsimulation. The d-statistic is a measure of within-area homogeneity, that is equivalent to intra-area correlation for areas with equal population. It can be used to identify the spatial autocorrelation exhibited by the categories of constraint variables, or combinations of categories, an important feature to reproduce when modelling local variation in a variable. It may be used to assess the statistical significance of the within-area homogeneity for a given set of categories and can assist in validating spatial microsimulation models.

Keywords: Constraint choice; Health data; Spatial microsimulation; Within-area homogeneity.

1 Introduction

Microsimulation models, are widely used to simulate the effects of policy decisions at the individual level (e.g. Bourguignon and Spadaro, 2006). Greater demand for spatially detailed statistical analyses, due to variability of many phenomena across space (e.g. Tranmer et al., 2005, Getis, 2008, Diez Roux and Mair, 2010, Rosenthal, 2012, and references therein) has led to the development of spatial microsimulation models (SMM) to simulate microdata for small geographic areas. The development of these models is documented in several recent reviews (e.g. Birkin and Clarke, 2011, Hermes and Poulsen, 2012a, Ballas et al., 2013), whilst Holm and Sanders

*Email: sburden@uow.edu.au

(2007) and Tanton and Edwards (2013) provide more conceptual overviews of the field.

SMM are used to model local and/or regional geographic effects for a diverse range of application areas, including labour force participation (Ballas et al., 2006, Morrissey and O'Donoghue, 2011); economic policy analysis (Chin et al., 2005, Campbell and Ballas, 2013); social policy analysis (Miranti et al., 2011, Tanton, 2011, Gong et al., 2012, Rahman et al., 2013); environmental applications (Hynes et al., 2009); education (Kavroudakis et al., 2013); retail market analysis (Hanaoka and Clarke, 2007); and transport modelling (Lovelace et al., 2014, Ma et al., 2014). They have also been used as the basis for dynamic microsimulation models (Ballas et al., 2005); geographical analysis of zoning distributions (Burden and Steel, 2013); behavioural models (van Leeuwen and Dekkers, 2013); health surveillance (Cataife, 2014); econometric analysis (Cullinan, 2011, Cullinan et al., 2011) and planning support (Ballas et al., 2007b).

They are a particularly valuable tool for health studies, because spatially detailed health data are rarely available. SMM have been developed to study a wide range of complex, health-related outcomes, including morbidity (Clark et al., 2014); disability estimates in older Australians (Lymer et al., 2008); obesity (Procter et al., 2008, Edwards and Clarke, 2009, Edwards et al., 2011, Cataife, 2014); depression (Morrissey et al., 2010); psychological distress (Riva and Smith, 2012); smoking (Smith et al., 2011, Hermes and Poulsen, 2012b); and pharmaceutical drug use (Abello et al., 2008).

SMM generally use two sources of information (although some techniques can be applied using a single dataset, e.g. Birkin and Clarke, 1988): one for individuals in the population (i.e. people, households, or businesses), that includes the set of variables of interest; and one for the geographic areas used in the simulation. The aggregate data are used to constrain the SMM, and are typically population counts for a set of categories, obtained from a single variable or a cross-tabulation of several variables, that are available in both the individual and area-level datasets. For each geographic area the individual records are weighted, to optimise their "fit" to the constraints for the area.

Early examples of SMM re-weighted individual records to fit the constraint data in each area using iterative proportional fitting techniques (Birkin and Clarke, 1988, 1989). More recently, deterministic re-weighting (DR) techniques (e.g. Ballas et al., 2005, Tanton et al., 2007, Smith et al., 2009), and probabilistic simulated annealing (SA) (Williamson et al., 1998) have been used. Although these methods both use sample records and constraint data, they use different methods to obtain the sample weights. For each area, the DR approach (e.g. Rahman et al., 2013, Tanton et al., 2011, Harland et al., 2012) calibrates the sample weights to minimise a distance function between the new weights and the existing sample design weights, such that the sum of each constraint equals the known population total for the area. The SA approach (Williamson et al., 1998, Voas and Williamson, 2000, 2001) uses random sampling with replacement to select an appropriate combination of records for each area. Initially, a set of randomly selected sample records are assessed against known benchmark constraints for each area to determine their goodness-of-fit (GOF). Using Markov chain Monte Carlo (MCMC) techniques, records are then stochastically

selected and replaced with other records from the sample when the exchange improves the GOF criterion, subject to simulated annealing constraints. The process is repeated until the GOF criterion reaches a given target or the maximum number of iterations is reached. The final set of selected records form integer weights for the area. SA and DR have been comprehensively described and compared in Williamson (2013) who found the results of the SA algorithm to be superior using several measures. However, a comparison conducted by Tanton et al. (2007) found no appreciable differences in the results from the two methods. Harland et al. (2012) also found, using several measures, that SA was superior, although they recognised that each technique had different strengths and was useful for different applications. Importantly, for a small population, Ryan et al. (2009) found that for both methods, more detailed constraints and, to a lesser extent, a larger sample increased the accuracy of the results.

An implicit assumption made by SMM methods is that the spatial variation in the variables of interest can be adequately reproduced in the simulated data using the selected constraints. SMM are typically fitted using a handful of key variables, due to the limitations of computational time (for SA) or convergence (for DR) (Chin and Harding, 2006, Tanton et al., 2011), although there is ongoing research into the use of a larger number of constraints (Harland and Heppenstall, 2009). Hence, selecting appropriate constraint variables from amongst the set of available variables is a key component of the modelling process.

To select constraint variables, studies use literature reviews (Birkin and Clarke, 2012); consultation with end users (Chin and Harding, 2006); and regression analysis. The latter is used to identify statistically significant variables that contribute explanatory power to the model (Chin and Harding, 2006) and to ensure the distribution of the outcome is represented by the constraints (Harland et al., 2012). Chin and Harding (2006) and Tanton et al. (2011) identify the need for constraints to be correlated with the outcome of interest to maximise information. However, as Birkin and Clarke (2011) discuss, appropriately correlated variables may not be known or the purpose of the microsimulation may not be well defined. Instead, they suggest including a wide range of personal and neighbourhood characteristics in the constraints.

Recently, SMM methods have been extended to include greater population diversity. Smith et al. (2009) used a local approach in which they clustered similar geographic areas in the study region and developed a suite of SMM using different constraint variables. The SMM that best reflected the population was selected for each cluster. Birkin and Clarke (2012) demonstrated the benefit of using geodemographic information from the sample records to fit the SMM in different regions. However, the usefulness of this approach relies on the availability of these data. When a SMM is fitted to data, selection of appropriate constraints that reproduce the spatial variation in the variables of interest is integral to the modelling process. However, measures of spatial correlation for the categorical constraint variables are not presently included in the constraint selection procedure.

In this paper we use the d-statistic introduced by Steel and Tranmer (2011) as a measure of within-area homogeneity for categorical data, to aid the procedure for selecting constraints for SMM. In Section 2 we describe the d-statistic, and outline

some of its properties. In Section 3 we give an example of its use, showing how it can be used in conjunction with established criteria for selecting SMM constraints. In Section 4 we briefly consider model validation and conclude with a discussion in Section 5.

2 The d-Statistic

The socio-economic characteristics of individuals who are located in proximity to one another tend to be more homogeneous than those for the overall population (Tobler (1970); see also the discussion in Steel and Holt (1996b) and Tranmer and Steel (1998)), so understanding the spatial structure of the population is important for policy decisions at the local area level (Tranmer et al., 2005). To reproduce spatial variation in simulated population data, constraint variables must adequately reflect the spatial variability of the variables of interest in the population. When they do not, local and regional variation in the observed relationships will not be reproduced in the analysis. That is, the *diversity* in the population (Smith et al., 2009, Birkin and Clarke, 2012) will not be reproduced in the simulated data. In this case, the data are potentially equivalent to those obtained using random aggregation, (see for example, Steel and Holt, 1996a) or equivalently a microsimulation model with no spatial component. An important aspect of the constraint selection procedure, therefore, is to characterise the spatial correlation in the constraint variables and the variables of interest. When this information is known, constraints that reflect the spatial correlation in the variables of interest can be selected.

Measuring spatial correlation in categorical constraint variables is not straightforward. Common statistics for measuring spatial autocorrelation, such as the Moran I or the Joint Count statistic are not appropriate for categorical data because they do not consider correlation between the categories. Instead, we use the d-statistic (Steel and Tranmer, 2011), to compare and assess the spatial variability, or diversity, of potential constraint variables. The d-statistic takes into consideration the presence of multiple categories and the negative correlation between them and it can be calculated without access to individual level data. It provides a measure of the contribution of each category to the overall statistic and it can be used to compare multiple categorical constraints to identify the within-area homogeneity of different combinations of categories.

For a population divided into M mutually exclusive groups (in this case, areas), so that the gth group contains N_g units, $N = \sum_{g=1}^{M} N_g$, the d-statistic for a categorical variable with $k = 1, ..., K$ categories is given by

$$d = \frac{1}{K-1} \sum_{k=1}^{K} (1 - P_k) d_k, \tag{1}$$

where P_k is the proportion of the population and d_k is the measure of within-area homogeneity in category k, respectively.

If the population variance for category k is given by $S_{kk} = P_k(1 - P_k)$, the

within-area homogeneity for category k, d_k, is given by

$$d_k = \left(\frac{\bar{S}_{kk}}{S_{kk}} - 1 \right) / (\bar{N}^* - 1), \tag{2}$$

where $\bar{S}_{kk} = \frac{1}{M-1} \sum_{g=1}^{M} N_g (P_{kg} - P_k)^2$ is the population weighted between-area variance; $\bar{N}^* = \bar{N} \left(1 - \frac{C^2}{M-1}\right)$; $\bar{N} = N/M$; and $C^2 = \left(\frac{1}{M} \sum_{g=1}^{M} (N_g - \bar{N})^2 \right) / \bar{N}^2$ is the square of the coefficient of variation of the population size in the areas. When all the areas have the same population size, d_k is equivalent to the intra-area correlation (IAC) for category k. For geographically defined groups, the intra-class correlation reflects the average of the within group spatial correlation between individuals. A detailed explanation of the statistic and examples of its use can be found in Steel and Tranmer (2011).

The range of d_k is,

$$\frac{-1}{\bar{N}^* - 1} \leq d_k \leq \frac{(\bar{N} - 1)}{(\bar{N}^* - 1)(1 - M^{-1})}.$$

As C^2 for census demographic variables typically lies between 0.25 and 1.0, and the number of areas is large ($M > 11000$), $\bar{N}^* \approx \bar{N}$ and $\min(d_k) \approx -0.005$, so that effectively $0 < d_k < 1$. When the data are randomly aggregated, the expected value (mean) of d_k is zero i.e. $E[d_k] = 0$.

A statistical test can be performed to identify which, if any, potential constraint variables exhibit significant within-area homogeneity. For the null hypothesis that a variable Y exhibits no within-area homogeneity for the given set of geographic areas, the test statistic, $X^2 = (M-1)(K-1)[(\bar{N}^*-1)d+1]$, is based on the definition of d in Steel and Tranmer (2011). When the null hypothesis is true it has a chi-squared distribution with $(M-1)(K-1)$ degrees of freedom,

$$X^2 = (M-1)(K-1)[(\bar{N}^* - 1)d + 1] \quad \sim \quad \chi^2_{(M-1)(K-1)}.$$

In addition, the statistic $X^2[(M-1)(K-1)]^{-1} = [(\bar{N}^*-1)d+1]$ provides a measure of the difference in the distribution of the categorical variable, Y, between the groups.

By construction, SMM preserves the within-area homogeneity of the constraints. So selecting constraints with high within-area homogeneity preserves this feature in the simulated data. The spatial structure in variables that are associated with the constraints will also be retained, thus enabling the simulation of microdata that reflects the diversity in the population.

An additional application of the d-statistic is for model validation. It can be used to calculate the within-area homogeneity of any simulated variable, or cross-tabulation, from the SMM, so within-area homogeneity of the actual and simulated data can be compared for variables available in both datasets. This facilitates the comparison and assessment of within-area homogeneity for different socio-economic dimensions in the model. It is particularly useful for understanding how well the actual within-area homogeneity is preserved in the simulated population for variables that were not used as constraints. Its applicability to complex cross-tabulations of categories also means that quite detailed comparisons can be made.

3 Constraint Selection Example

This section demonstrates the use of the d-statistic for selecting constraint variables, for a study considering the microsimulation of realistic health outcome and covariate data for the population of New South Wales (NSW), Australia. The data were simulated to study the modifiable areal unit problem for regression parameter estimates of aggregate health data (Burden and Steel, 2013). Individual level health data were simulated using the SA approach and the CO software (Williamson, 2007, Williamson et al., 1998, available at *http://pcwww.liv.ac.uk/~william/microdata/*).

The sample data used in the study were 20788 unit records from the 2007-2008 National Health Survey (NHS) (Australian Bureau of Statistics, 2008, 2009). They were combined with spatially detailed, aggregate covariate data from the 2006 Australian Census (Australian Bureau of Statistics, 2006a) to simulate a realistic population living in private dwellings. At the finest scale, the 2006 Census provides basic demographic variables for 11879 populated Census Collection districts (CD's) in NSW, which have an average of approximately 550 residents. The data are confidentialised by introducing small random errors in the counts for each area (Australian Bureau of Statistics, 2006a). Cross-tabulations for each area are internally consistent, but minor errors may be observed when data are aggregated.

Previous Australian SMM have used Census data at the Statistical Local Area (SLA) level (e.g. Vidyattama and Tanton, 2010, Chin et al., 2005), which has an average population of approximately 32,000. The advantages of SLA's are that they are not substantially affected by confidentiality issues and more extensive covariate data is available (Chin and Harding, 2006). However, the availability of more detailed data must be balanced against the loss of spatial resolution, which is particularly important for detailed spatial modelling of small area health data. In this study CD level data were used.

The study included two binary response variables: Type 2 diabetes mellitus (diabetes) and angina, that together account for a significant portion of the disease burden in Australia (AIHW, 2009). Statistical models for each response included the covariates age (in approximately 10-year classes), sex and an index of socioeconomic status for the area, plus a set of binary risk factors: Current smoking status (Smoker); a sedentary lifestyle (Sedent); dietary fat (consumption of whole milk/regular/full cream milk with 3% or more fat (DietaryFat)); and obesity (body mass index ≥ 25 (Obesity)). The socioeconomic index was calculated for the simulated data using methods comparable with the Australian Bureau of Statistics socioeconomic indices (Australian Bureau of Statistics, 2006b). Further details can be found in Burden and Steel (2013).

3.1 Constraint Selection Procedure

To formalise the established constraint selection criteria (e.g. Chin and Harding, 2006, Smith et al., 2009, Birkin and Clarke, 2011, 2012, Tanton et al., 2011, Harland et al., 2012) and to incorporate the use of within-area homogeneity into the criteria, we propose to use the following general set of principles for SMM constraint selection:

1. The chosen set of constraints must provide basic demographic information

about individuals in the population;

2. The set of constraints must be associated with the outcome(s) of interest, to maximise information and to ensure the distribution of the outcome is represented by the constraints;

3. The constraints should not be highly collinear, to minimise processing time;

4. The set of constraints must reflect the spatial variation of the population;

5. The set of constraints must represent a broad range of relevant socioeconomic dimensions.

These principles may be applied to a given problem through the use of exploratory data analysis, common statistical diagnostics and professional judgement. For example, the identification of relevant socioeconomic dimensions and demographic variables depends on the proposed use of the microdata and requires professional judgement. The strength of association between the constraints and outcomes may be assessed using regression diagnostics, such as the statistical significance of regression parameter estimates and size of the corresponding coefficient of determination. Multi-collinearity between constraints can be identified using diagnostics such as correlation coefficients and variance inflation factors; and spatial variation in constraint variables can be assessed using the d-statistic proposed in this paper.

As the example below shows, it is unlikely that a constraint variable will satisfy all of these principles. Hence, the final set of constraints should include at least one variable satisfying each principle, and variables which satisfy multiple principles should preferentially be chosen. The relative importance attributed to each principle depends on the proposed use of the simulated data.

For this project, age×sex categories (a well-known predictor of health e.g. Elliott et al., 2000), were included as the basic demographic constraint. Other variables that were available in both the NHS and 2006 Census, that could be used as constraints were social marital status; country of birth; proficiency in spoken English; main language spoken at home; highest year of school; type of educational institution currently attending; highest non-school qualification - level of education or field of study; labour force status; gross weekly individual income (deciles); family versus non-family households; occupation; dwelling structure and tenure type.

To identify constraints with a statistically significant relationship with the response variables, each response was regressed separately against the potential constraint variables, and age×sex categories (with $0-39$ yrs combined), using logistic regression. The d-statistic was calculated for the constraint variables, to provide a measure of spatial variation within the CD's. For variables with many categories, the statistic was also used to identify appropriate categories to combine. The results of the analyses were evaluated in terms of the five selection criteria above and the final set of constraints was chosen to represent a variety of socioeconomic dimensions including housing, education, occupation and ethnicity.

3.2 Regression Analysis Results

The results of individual-level logistic regression analyses using NHS were used to identify statistically significant predictors (at the 5% level) of the response variables angina and diabetes. They are shown in Table 1 along with the base categories used for comparison. After accounting for age and sex, the variables that are generally predictive of the responses are ethnicity; income and labour force status; self assessed health; education; household composition; and housing variables.

The results highlight two issues in the selection of constraint variables for microsimulation. First, several different demographic variables are significant predictors of each response, but there is little overlap between them, so the selection of constraints that efficiently reproduce the complex relationships in the data is not simple (Birkin and Clarke, 2012). A second complexity arises because many variables that are useful for predicting each response, such as self assessed health, are not available in the Census and so cannot be used as constraints.

The regression results identified the main socioeconomic dimensions that are associated with the incidence of disease. In this case, housing variables, household structure and composition, education and income/employment variables all have some association with the response variables. Hence, the spatial distribution of these variables should, to an extent, reflect the spatial distribution of the response variables.

3.3 Within-Area Homogeneity Results

The within-area homogeneity of the potential constraint variables, calculated using the d-statistic, is shown in Table 2 for each variable, summarised into the given number of categories. The socioeconomic dimensions with the highest within-area homogeneity are the housing characteristics and language skills. In order of homogeneity, the variables are: dwelling structure by dwellings and then by persons; main language spoken at home; language and proficiency in English; dwelling structure $\times$ tenure type; and tenure type.

Occupation- and education- related variables generally had a low to medium within-area homogeneity. Available variables included occupation; type of institution currently attending; highest year of school; level of education (non-school qualification); social marital status; and non-school qualification: field of study. The general demographic variables such as labour force status and age$\times$sex showed low levels of within-area homogeneity.

These results identified that the inclusion of housing characteristics and language skills amongst the constraint variables was important to ensure that the within-area homogeneity of the data was maintained at the CD level. Moreover, demographic variables that are highly predictive of the response variables at the individual level, do not necessarily exhibit high within-area homogeneity at the area level.

3.4 Constraint Selection

The final set of constraints was chosen using the principles outlined in Section 3.1. Variables were selected to represent different socioeconomic dimensions such as hous-

Table 1: Statistically significant estimated logistic regression coefficients, at $\alpha = 0.05$, for responses regressed at person-level on age×sex and each covariate separately using the NHS data. Base category in brackets below.

Variable	Level	Angina	Diabetes
Tenure type	Renter	0.59 (0.153)***	
	(Owner without mortgage)		
Country of birth	Other		0.43 (0.116)**
	(Australia)		
Labour force status	Not in labour force	1.09 (0.232)***	0.92 (0.153)***
Highest year of school completed	Yr 9 or equiv		0.57 (0.176)**
	Yr 8 or below		0.59 (0.182)**
	(Yr 12)		
Main language spoken at home	Other language		0.65 (0.152)***
	(English)		
Number of bedrooms	Four or more		-0.39 (0.171)*
	(One)		
Self assessed health	Very good	1.20 (0.449)*	0.72 (0.355)*
	Good	1.92 (0.449)***	2.00 (0.263)***
	Fair	2.54 (0.446)***	2.69 (0.286)***
	Poor	3.29 (0.427)***	3.06 (0.295)***
	(Excellent)		
Landlord type	Private	0.49 (0.156)**	
	Public	0.79 (0.266)**	0.73 (0.246)**
	Other		0.89 (0.392)*
	(Not applicable)		
Household structure	Single adult + child(ren)		0.67 (0.241)**
	All other households		0.74 (0.293)*
	(Couple and child(ren))		
Equivalent income of the household	Fourth decile	-0.58 (0.288)*	
	Fifth decile		-0.52 (0.236)*
	Seventh decile	-1.51 (0.517)**	-0.60 (0.260)*
	Eighth decile	-1.20 (0.525)*	-0.55 (0.243)*
	Ninth decile	-1.03 (0.484)*	-0.74 (0.335)*
	Tenth decile	-1.50 (0.580)*	-0.86 (0.293)**
	(First Decile)		
Household composition	≥ 2 family only		0.95 (0.369)*
	≥ 1 family + non-fam	1.41 (0.697)*	
	(One family household with only family members present)		

Standard errors in parentheses
*($p < 0.05$), **($p < 0.01$), ***($p < 0.001$)

Table 2: Comparison of within-area homogeneity for the census and simulated data.

	2006 Census Data				Simulated Data		
	d	k	Min(d_k)	Max(d_k)	d	Min(d_k)	Max(d_k)
Age by sex	0.008	18	0.002	0.020	0.007	0.002	0.018
Country of birth	0.110	3	0.024	0.180	0.102	0.024	0.179
Year left school	0.030	5	0.003	0.077	0.030	0.003	0.081
Dwelling structure × tenure type	0.108	12	0.043	0.305	0.107	0.037	0.308
Occupation	0.018	6	0.003	0.038	0.018	0.003	0.038
Number of bedrooms	0.145	4	0.083	0.211	0.145	0.087	0.211
Income	0.061	4	0.019	0.086	0.060	0.019	0.086
Tenure	0.081	4	0.046	0.129	0.076	0.040	0.141
Language	0.298	2	0.282	0.283	0.086	0.035	0.081
Landlord	0.194	3	0.062	0.256	0.061	0.013	0.101
Proficiency in English	0.083	5	0.024	0.282	0.024	0.003	0.045
Labour force status	0.026	3	0.007	0.037	0.024	0.005	0.036
Number in household	0.033	6	0.010	0.048	0.018	0.004	0.052

k = number of categories

ing, education, occupation/income and ethnicity. Some were significant predictors of the outcome variables, whilst others showed relatively high within-area homogeneity. The variables chosen for use in the CO software were: age by sex; country of birth (main English speaking, other); tenure type (own, rent, other) by ownership (house, townhouse, flat, other); highest year of school (year 11-12, year 10, year 9, ≤year 8); and occupation (manager, professional, driver, labourer, community worker, other); number of bedrooms (1, 2, 3, 4+); and four categories of income deciles (1, 2-3, 4-8, 9-10). The characteristics of these categories for the CD's in NSW are summarised in Table 5 in the Appendix.

4 SMM Validation and Comparison

Using the constraint categories and the simulation parameters defined in Appendix 6.1, a population of 6,378,163 individuals was simulated for 11,879 populated CDs in NSW. Whilst the CO software accepts several different measures of goodness-of-fit (Voas and Williamson, 2000), in this project, the overall relative sum of Z-scores (ORSZ) (Voas and Williamson, 2001) was used to assess the fit for each area, with a target value of 1.0. This metric creates a Z-score for the difference between the observed and expected counts for each level of each constraints, and it has been used by several authors (Ryan et al., 2009, Williamson, 2013).

Goodness-of-fit measures for the CO output are summarised in Table 3. For 90% of areas, ORSZ < 0.51, and the 95th percentile of the ORSZ statistic was 1.16. However there were some areas for which a good fit could not be obtained. Most areas also fit well according to the overall total absolute error per household (OTAE/HH) GOF criterion. This is a commonly used statistic for assessing the fit of SMM which is equal to the sum of the absolute difference between the observed

Table 3: Fit statistics for the simulated data from the CO Program

	OTAE	OTAE/HH	ORSZ	Number of Duplicates	Number of Households
Mean	77.7	0.16	0.61	18.3	537
Variance	6280	0.02	34.51	445	66641
Min.	3	0.02	0.00	0.0	3
Median	61	0.13	0.06	10.5	510
80th percentile	107	0.22	0.22	22.2	738
90th percentile	139	0.27	0.51	46.4	871
95th percentile	174	0.33	1.16	76.7	990
99th percentile	278	0.70	9.35	93.5	1288
Max	1936	2.73	332.06	99.8	2755

and benchmark counts for all of the constraint categories per household in each area. The 80th percentile for the statistic is 0.22, which is close to the recommended value when it is being used as a constraint (Smith et al., 2009).

Lower values of the fit statistics mean that the simulated population more closely resembles the chosen constraint categories. However, this does not necessarily correspond to a good representation of the true population. How well the simulated population reflects the true population depends on both the fit of the model, and how well the constraints represent the actual population. Maps of the ratio of observed to expected counts for Angina and Diabetes are shown for NSW CD's in Figure 1. The ratio ranges from 0 to 12.5 for angina (mean=1.01, median 0.94) and from 0 to 4.4 for diabetes (mean=0.99, median=0.98).

A well documented drawback of microsimulation models is that direct validation of the model is not possible (Edwards et al., 2011). Instead, a combination of internal and external validation is used. Validating the SMM using external data is an important step in creating a SMM and Edwards and Tanton (2013) includes a recent description of options for external validation. These techniques have not been applied to the present study, due to the nature of the research. However, the importance of external validation when model estimates are used for substantive applications and inference cannot be overstated. Several internal validation techniques are also widely used (Rahman et al., 2013). These include aggregating to a level for which known values for the variables are available (Morrissey and O'Donoghue, 2011); the use of total absolute error measures; regression analyses; plots of simulated versus actual error; and tests of statistical significance (Hynes et al., 2009, Rahman et al., 2010). The *d*-statistic can be added to these techniques as another useful tool for model validation. It provides a way to compare the within-area homogeneity of the actual and simulated data. In this paper, simulated and observed counts are compared using several established techniques as well as the *d*-statistic.

Table 4 summarises the prevalence of key variables in the simulated data and provides a comparison of the simulated and actual data. The simulated totals for NSW are compared with the actual totals for NSW, Australia and the equivalent unweighted total for NSW. The totals were adjusted to the population of NSW to account for the different demographic structure in NSW compared with Australia

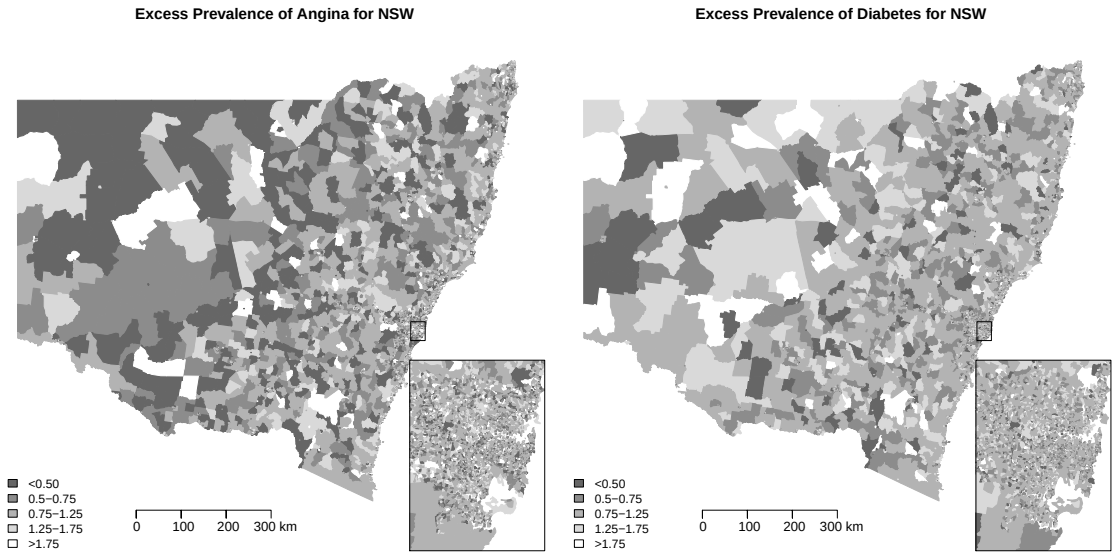

Figure 1: Map of the ratio of simulated observed to expected counts of Angina, for CD's in NSW. The inset shows the CD's in the Sydney region in more detail.

as a whole. Table 4 shows that the counts are reasonably similar for some variables, whilst large differences can be observed for others. For example, angina and dietary fat were within 5% of the actual counts adjusted to the Australian population whilst diabetes and overweight were almost 20% lower.

Table 4: Comparison of the total simulated and actual counts for the health outcomes and risk factors obtained using the simulated data and weighted population estimates from the NHS sample data.

	Sim '000	NSW '000	% Diff	Aus.Adj. '000	% Diff	Unwtd '000	% Diff
Diabetes	199	247	-24	237	-19	244	-23
Angina	121	83	31	116	4	133	-10
smoker	1049	1084	-3	1096	-5	1120	-7
sedentary	1955	2142	-10	2133	-9	2117	-8
Overweight	1135	1418	-25	1356	-20	1325	-17
Obese	763	905	-19	908	-19	910	-19
Dietary Fat	2924	3412	-17	3068	-5	2934	0

Correlation coefficients between the simulated and actual data were greater than 0.97 for the constraint variables. However, for the categories of other variables there was a larger range in correlation coefficients. For example, main language spoken at home (0.88 – 0.95); landlord type (0.18 – 0.87), proficiency in English (0.68 – 0.92); labour force status (0.76 – 0.999) and number of persons in the house (0.60 – 0.93). The categories with small counts and those with high within-area homogeneity were the most highly variable.

As correlation coefficients only provide a summary statistic for each variable, a plot of simulated versus actual counts is frequently used to asses how well the

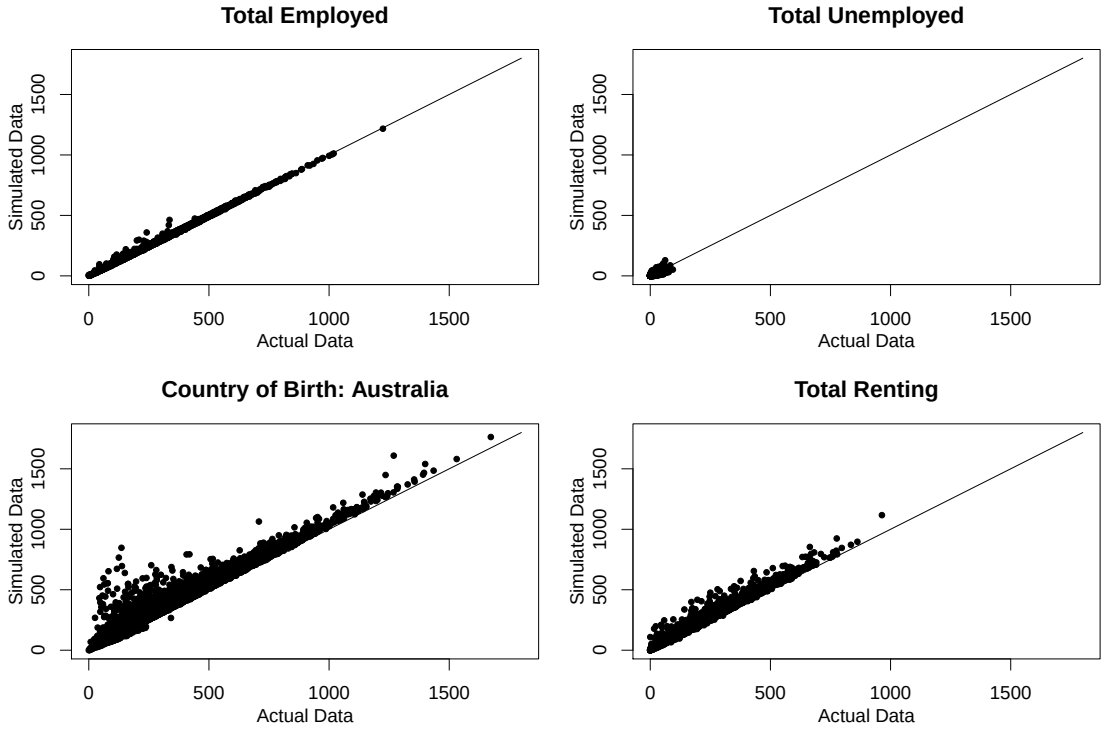

Figure 2: Comparison of simulated and actual counts for CD's in NSW using two employment categories; country of birth: Australia and total persons renting.

simulated counts fit the actual distribution of counts (Ballas et al., 2007a, Tanton et al., 2011). When the data are well simulated, the values lie close to the 45 degree line, and hence dispersion about this line provides a graphical summary of the fit of the simulated data. Variables that were available in both the Census and simulated datasets were compared using plots of simulated and actual counts for each area (e.g. Figure 2). Some variables, such as the categories of employment variables, were well simulated, with most values lying close to the 45 degree line. However, for other variables, such as country of birth or total renting, counts in the simulated data exceed those in the actual data. The results suggest that the spatial distribution of the constraints does not match the distribution of these variables, and records with these characteristics are over-represented in the simulated data.

The empirical cumulative probability density function (ECDF) of a variable for a set of areas gives a useful indication of the spatial concentration of the variable (Rahman et al., 2010, 2013). For each value along the x-axis, it shows the probability that the category counts in a randomly selected area will be less than the chosen value. It can also be used to compare the distribution of counts in the simulated and actual data for any category. For example, the ECDF's for the categories of landlord and main language spoken at home, are shown in Figure 3. They show that the simulated and actual counts for each area have a similar distribution, but that systematic over or underestimation does occur. The ECDF for the difference in the simulated and actual data for both categories of the language variable confirms that there is systematic under-representation (or bias) of the other language category in the simulated data.

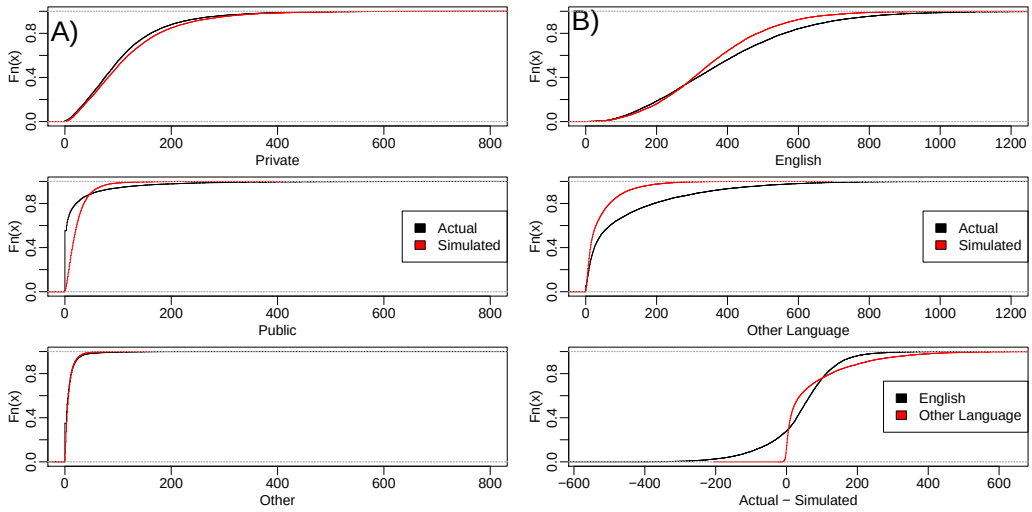

Figure 3: Empirical cumulative distribution function for A) the three categories of landlord and B) two categories of language for simulated and actual counts in each area, truncated to counts of 800 and 1200 respectively. The difference between actual and simulated counts of A) and B) are compared in the lower right panel.

Comparing the spatial distribution of variables in a set of areas using within-area homogeneity can also be used to validate SMM. Table 2 shows that the range of values for the d-statistic in the simulated variables is similar to the actual data, with age by sex in particular having very low homogeneity and country of birth and dwelling structure having high within-area homogeneity. For some variables, particularly those that were not utilised as constraints (i.e., main language spoken at home), the within-area homogeneity is substantially lower for the simulated data than the actual data. This result is verified by the map of population percentage whose main language at home is English (Figure 4), which shows that the range of values for the simulated areas is narrower than the actual range and that there appears to be much less similarity between neighbouring areas than in the actual data.

5 Discussion

The results in the previous section highlight two important advantages of using within-area homogeneity for SMM. First, using within-area homogeneity to identify the variables that exhibit spatial variation provides useful information for the selection of constraints for the SMM. The use of constraints with high within-area homogeneity retains spatial variation in the data for these and other correlated variables. However, the variables with high within-area homogeneity may not be the same as those which are useful for the prediction of a specific response variable. The appropriate strategy for selecting constraints in such a case depends on the purpose of the microsimulation. For a targeted microsimulation with a narrow focus, the use of highly predictive variables may be appropriate. In other cases (e.g. Birkin and Clarke, 2011), the use of a combination of constraints that are either predictive of

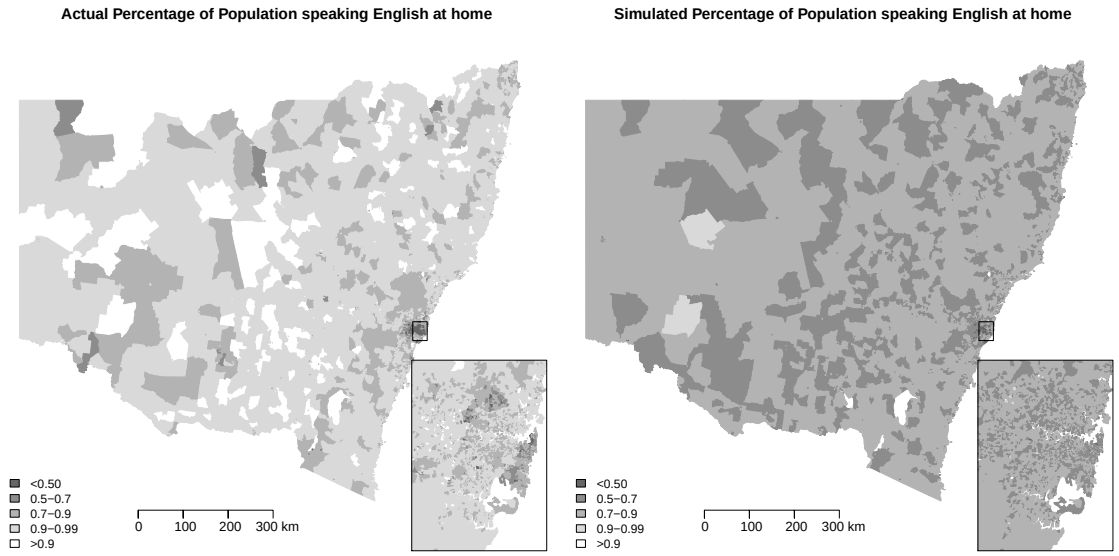

Figure 4: Maps of actual and simulated percentage of people whose main language is English, for CD's in NSW

the main variables of interest or representative in the diversity of the population is more appropriate and within-area homogeneity provides a valuable tool for selecting constraints in these cases.

The second advantage is that within-area homogeneity, in conjunction with other techniques, can be used to interrogate and validate the spatial structure of the simulated data. The results in previous sections highlight that whilst the simulated and actual totals are similar for some variables, other variables show evidence of substantial bias in the counts per area. The bias may arise due to confounding, whereby the distribution of the variable in an area, given the constraints, does not reflect the true distribution of the variable, given the constraints, due to the presence of other unmeasured variables. The use of two different data sources to obtain the sample records and geographic summaries may also compound the errors. This was recently considered by Vidyattama et al. (2013), who concluded that detailed validation was required to identify differences in the variable distributions from each database. Similarly, if the distribution of the selected constraints in a specific area is not representative of the overall distribution of the constraints, spatial disparity and sampling bias may occur (Harland et al., 2012). In all of these cases, the d-statistic provides a useful tool that can assist in validating the model and identifying variation between the simulated and actual data.

Biased counts may also arise because the constraints do not include enough variables that exhibit spatial correlation, possibly when appropriate data are not available, in which case alternative modelling strategies or the inclusion of additional data sources may be required. For example, where accurate population totals are required, the modelling procedure can be combined with calibration techniques (e.g. Morrissey and O'Donoghue, 2011) to adjust the constraints to a given value, rate or mean at defined levels of spatial aggregation. The use of the d-statistic can identify and assist in rectifying this situation, although some bias may be unavoidable.

An alternative source of bias may have arisen in this study because the NHS

sample weights were ignored when simulating the population. This changes the probability of selection of each sample record, increasing (decreasing) the selection probabilities of records with low (high) weights as all records were equally weighted in the simulation. Moreover, selecting records from a sample flattens out extreme demographics, to be closer to the sample mean (Harland et al., 2012). When the probability of disease varies throughout the population, these effects can result in biased estimates in the simulated population. If the sample mean is different from the population mean, additional bias may result. For this study, the differences between the weighted and unweighted populations are relatively small (see Table 4). The prevalence of most variables is similar to those obtained using the sample weights, so the NHS sample weights are not considered to be a substantial cause of bias. The main factor appears to be the choice of constraints and their association with the response variables and covariates, which was limited by the available data.

A final consideration is that the constraint variables in areas with small populations at the CD level have been perturbed, introducing noise into the fitting algorithm. Harland et al. (2012) account for this by adjusting the constraints to sum to the population totals. In this study, several of the constraints did not apply to every member of the population for each area. For example, the variable "year left school" only applies to individuals over 15. In this case the variability was taken into account as part of the "other" category. This approach potentially biases the results when the population totals do not match for each area. However, as the error is added to the data at random it should not contribute to the systematic bias observed in the simulated results.

In conclusion, the d-statistic is a useful measure for selecting appropriate SMM constraints. The statistic applies to categorical variables, and provides information which can substantially improve the model fit and hence spatially disaggregated analysis of the resulting microdata. Its usefulness extends to model validation, where it can provide valuable information to compare the within-area homogeneity of the simulated data to that of the observed data.

6 Appendices

6.1 CO Program Parameters

The following simulated parameters were used to simulate the NSW population in private dwellings using data from the 2006 Australian Census and NHS data.

- Initial 'temperature' = default(no. of constraint cells) = 51
- Evaluations before temperature change = default(10 x temp0) = 510
- Rate of decrease of 'temperature' = default(0.95) = 0.95
- step size = Default(200000) = 200000
- Minimum no. evaluations per estimation area = Default(200000) = 200000
- max. no of evaluations during normal sampling = Default(4000000) = 4000000
- max. evaluations per estimation area = Default(4000000) = 5000000
- minimum target for ORSZ (Overall relative sum of Z2) = 1.96
- minimum target value for OTAE (overall total absolute error) = 250

Table 5: Statistics for the constraint categories used to create simulated data. Basic statistics are for the proportion of the population from each area in each category

	Mean	Std.Dev	Median	Min	Max	d_k
0-9yrs×M	0.07	0.02	0.07	0.00	0.38	0.01
10-19yrs×M	0.05	0.02	0.06	0.00	0.60	0.01
20-29yrs×M	0.08	0.04	0.07	0.00	1.00	0.02
30-39yrs×M	0.07	0.03	0.06	0.00	0.75	0.01
40-49yrs×M	0.07	0.02	0.07	0.00	1.00	0.00
50-59yrs×M	0.07	0.02	0.06	0.00	0.64	0.00
60-69yrs×M	0.05	0.02	0.04	0.00	0.36	0.01
70-79yrs×M	0.03	0.02	0.03	0.00	0.27	0.01
80+yrs×M	0.01	0.01	0.01	0.00	0.30	0.01
0-9yrs×F	0.06	0.02	0.06	0.00	0.27	0.01
10-19yrs×F	0.05	0.02	0.05	0.00	0.27	0.01
20-29yrs×F	0.08	0.04	0.07	0.00	0.41	0.02
30-39yrs×F	0.07	0.03	0.07	0.00	0.38	0.01
40-49yrs×F	0.08	0.02	0.07	0.00	1.00	0.00
50-59yrs×F	0.07	0.02	0.07	0.00	0.25	0.00
60-69yrs×F	0.05	0.02	0.04	0.00	0.67	0.01
70-79yrs×F	0.03	0.02	0.03	0.00	0.28	0.01
80+yrs×F	0.02	0.02	0.02	0.00	0.44	0.02
COB: Main Eng Spk	0.07	0.04	0.06	0.00	1.00	0.02
COB: Other	0.15	0.15	0.08	0.00	0.82	0.18
Yr12	0.33	0.14	0.31	0.00	1.80	0.08
Yr11	0.05	0.02	0.05	0.00	0.21	0.00
Yr10	0.21	0.08	0.22	0.00	0.52	0.03
Yr9	0.07	0.03	0.06	0.00	0.38	0.02
Yr8 or lower	0.06	0.04	0.06	0.00	0.28	0.02
House×Own	0.26	0.14	0.28	0.00	1.00	0.09
House×Mortgage	0.30	0.17	0.31	0.00	0.86	0.14
House×Rent	0.14	0.11	0.12	0.00	1.00	0.09
House×Other	0.02	0.04	0.01	0.00	1.00	0.04
T/House×Own	0.02	0.04	0.00	0.00	0.59	0.08
T/House×Mortgage	0.02	0.05	0.00	0.00	0.57	0.10
T/House×Rent	0.03	0.07	0.01	0.00	0.88	0.13
T/House×Other	0.00	0.01	0.00	0.00	0.44	0.06
Flat×Own	0.02	0.04	0.00	0.00	0.67	0.09
Flat×Mortgage	0.02	0.05	0.00	0.00	0.48	0.13
Flat×Rent	0.08	0.15	0.01	0.00	0.92	0.30
Flat×Other	0.00	0.01	0.00	0.00	0.71	0.05
Manager	0.07	0.06	0.05	0.00	0.72	0.03
Professional	0.09	0.06	0.08	0.00	0.75	0.04
Driver	0.03	0.02	0.03	0.00	0.50	0.01
Labourer	0.04	0.02	0.04	0.00	0.75	0.01
Community Worker	0.04	0.02	0.04	0.00	0.36	0.00
Other	0.17	0.05	0.17	0.00	1.33	0.01
1 bedroom	0.03	0.05	0.01	0.00	0.70	0.08
2 bedroom	0.17	0.17	0.12	0.00	0.91	0.21
3 bedroom	0.40	0.15	0.41	0.00	1.00	0.09
4+ bedroom	0.34	0.20	0.32	0.00	1.00	0.19
Income: 1 dec	0.03	0.03	0.02	0.00	0.42	0.02
Income: 2-3 dec	0.21	0.12	0.20	0.00	1.00	0.07
Income: 4-8 dec	0.52	0.12	0.54	0.00	1.00	0.05
Income: 9-10 dec	0.06	0.08	0.03	0.00	0.60	0.09

References

Abello, A., Lymer, S., Brown, L., Harding, A., and Phillips, B. (2008). Enhancing the australian national health survey data for use in a microsimulation model of pharmaceutical drug usage and cost. *Journal of Artificial Societies and Social Simulation*, **11** : 2.

AIHW (2009). Prevention of cardiovascular disease, diabetes and chronic kidney disease: targeting risk factors. Cat. no. PHE 118., AIHW, Canberra.

Australian Bureau of Statistics (2006a). *Census Dictionary*. cat. no. 2901.0 (reissue) ABS, Canberra.

Australian Bureau of Statistics (2006b). *Socio-Economic Indexes for Areas (SEIFA) - Technical Paper*. cat. no. 2039.0.55.001. ABS, Canberra.

Australian Bureau of Statistics (2008). *National Health Survey 2007-08*. Basic CURF, CD-ROM. Findings based on use of ABS CURF data.

Australian Bureau of Statistics (2009). *National Health Survey: Users' guide - Electronic Publication, 2007-08*. Cat. No. 4363.0.55.001. Viewed 21 June 2012 http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/CC0FB5A08570984ECA25762E0017CF2B/$File/4363055001_2007-08.pdf.

Ballas, D., Clarke, G., and Dewhurst, J. (2006). Modelling the socio-economic impacts of major job loss or gain at the local level: a spatial microsimulation framework. *Spatial economic analysis*, **1** : 127–146.

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

Ballas, D., Clarke, G., Dorling, D., and Rossiter, D. (2007a). Using simbritain to model the geographical impact of national government policies. *Geographical Analysis*, **39** : 44 – 77.

Ballas, D., Clarke, G., Hynes, S., Lennon, J., Morrissey, K., and O'Donoghue, C. (2013). A review of microsimulation for policy analysis. In *Spatial microsimulation for rural policy analysis*, Advances in spatial science, pages 35 –54. Springer, Heidelberg and New York.

Ballas, D., Kingston, R., Stillwell, J., and Jin, J. (2007b). Building a spatial microsimulation-based planning support system for local policy making. *Environment and Planning A*, **39** : 2482–2499.

Birkin, M. and Clarke, G. (2012). The enhancement of spatial microsimulation models using geodemographics. *The Annals of Regional Science*, **49** : 515–532.

Birkin, M. and 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. and Clarke, M. (1989). The generation of individual and household incomes at the small area level using SYNTHSIS. *Regional Studies*, **23** : 535–548.

Birkin, M. and Clarke, M. (2011). Spatial microsimulation models: a review and a glimpse into the future. In Stillwell, J. and Clarke, M., editors, *Population Dynamics and Projection Methods*, Understanding Population Trends and Processes, pages 193–208. Springer.

Bourguignon, F. and Spadaro, A. (2006). Microsimulation as a tool for evaluating redistribution policies. *The Journal of Economic Inequality*, **4** : 77–106.

Burden, S. and Steel, D. (2013). Characteristics of empirical zoning distributions for small area health data. Working Paper 15-13, University of Wollongong.

Campbell, M. and Ballas, D. (2013). A spatial microsimulation approach to economic policy analysis in Scotland. *Regional Science Policy & Practice*, **5** : 263–288.

Cataife, G. (2014). Small area estimation of obesity prevalence and dietary patterns: a model applied to Rio de Janeiro city, Brazil. *Health & place*, **26** : 47–52.

Chin, S.-F. and Harding, A. (2006). Regional Dimensions: creating synthetic small-area microdata and spatial microsimulation models.

Chin, S.-F., Harding, A., Lloyd, R., and McNamara, J. (2005). Spatial microsimulation using synthetic small-area estimates of income, tax and social security Benefits. *Australasian Journal of Regional Studies*, **11** : 303–335.

Clark, S. D., Birkin, M., and Heppenstall, A. (2014). Sub regional estimates of morbidities in the English elderly population. *Health & Place*, **27** : 176–185.

Cullinan, J. (2011). A Spatial Microsimulation Approach to Estimating the Total Number and Economic Value of Site Visits in Travel Cost Modelling. *Environmental & Resource Economics*, **50** : 27–47.

Cullinan, J., Hynes, S., and O'Donoghue, C. (2011). Using spatial microsimulation to account for demographic and spatial factors in environmental benefit transfer. *Ecological Economics*, **70** : 813–824.

Diez Roux, A. V. and Mair, C. (2010). Neighbourhoods and health. *Annals of the New York Academy of Sciences*, **1186** : 125–145.

Edwards, K. and Clarke, G. P. (2009). The design and validation of a spatial microsimulation model of obesogenic environments for children in Leeds, UK: SimObesity. *Social Science & Medicine*, **69** : 1127–1134.

Edwards, K., Clarke, G. P., Thomas, J., and Forman, D. (2011). Internal and external validation of spatial microsimulation models: small area estimates of adult obesity. *Applied Spatial Analysis and Policy*, **4** : 281–300.

Edwards, K. and Tanton, R. (2013). Validation of spatial microsimulation models. In *Spatial microsimulation*, pages 249 –258. Springer, Dordrecht and New York.

Elliott, P., Wakefield, J., Best, N., and Briggs, D. (2000). 1. spatial epidemiology: Methods and applications. In Elliott, P., Wakefield, J., Best, N., and Briggs, D., editors, *Spatial Epidemiology*, pages 3–14. Oxford University Press, London.

Getis, A. (2008). A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective. *Geographical Analysis*, **40**(3) : 297–309.

Gong, H., McNamara, J., Vidyattama, Y., Miranti, R., Tanton, R., Harding, A., and Kendig, H. (2012). Developing spatial microsimulation estimates of small area advantage and disadvantage among older australians. *Population Space and Place*, **18** : 561 – 565.

Hanaoka, K. and Clarke, G. (2007). Spatial microsimulation modelling for retail market anaysis at the small-area level. *Computers, Environment and Urban Systems*, **31** : 162 – 187.

Harland, K. and Heppenstall, A. (2009). Modelling individual consumer behaviour: ESRC/BSPS seminar series microsimulation modelling in the UK.

Harland, K., Heppenstall, A., Smith, D., and Birkin, M. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, **15** : 1.

Hermes, K. and Poulsen, M. (2012a). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, **36** : 281–290.

Hermes, K. and Poulsen, M. (2012b). Small area estimates of smoking prevalence in London. Testing the effect of input data. *Health & place*, **18** : 630–638.

Holm, E. and Sanders, L. (2007). Spatial microsimulation models. In *Models in spatial analysis*, pages 159 – 195. ISTE, London and Newport Beach, CA.

Hynes, S., Morrissey, K., O'Donoghue, C., and Clarke, G. (2009). A spatial microsimulation analysis of methane emissions from irish agriculture. *Ecological Complexity*, **6** : 135–146.

Kavroudakis, D., Ballas, D., and Birkin, M. (2013). Using spatial microsimulation to model social and spatial inequalities in educational attainment. *Applied Spatial Analysis and Policy*, **6** : 1–23.

Lovelace, R., Ballas, D., and Watson, M. (2014). A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels. *Journal of Transport Geography*, **34** : 282–296.

Lymer, S., Brown, L., Yap, M., and Harding, A. (2008). 2001 regional disability estimates for new south wales, australia, using spatial microsimulation. *Applied Spatial Analysis*, **1** : 99 – 116.

Ma, J., Heppenstall, A., Harland, K., and Mitchell, G. (2014). Synthesising carbon emission for mega-cities: A static spatial microsimulation of transport CO2 from urban travel in Beijing. *Computers, Environment and Urban Systems*, **45** : 78–88.

Miranti, R., McNamara, J., Tanton, R., and Harding, A. (2011). Poverty at the local level: national and small area poverty estimates by family type for Australia in 2006. *Applied Spatial Analysis and Policy*, **4** : 145–171.

Morrissey, K., Hynes, S., Clarke, G., and O'Donoghue, C. (2010). Examining factors associated with depression at the small area level in ireland using spatial microsimulation techniques. *Irish Geography*, **43** : 1 – 22.

Morrissey, K. and O'Donoghue, C. (2011). The spatial distribution of labour force participation and market earnings at the sub-national level in Ireland. *Review of Economic Analysis*, **3** : 80–101.

Procter, K., Clarke, G. P., Ransley, J. K., and Cade, J. (2008). Micro-level analysis of childhood obesity, diet, physical activity, residential socioeconomic and social capital variables: where are the obesogenic environments in Leeds? *Area*, **40** : 323–340.

Rahman, A., Harding, A., Tanton, R., and Liu, S. (2010). Methodological issues in spatial microsimulation modelling for small area estimation. *International Journal of Microsimulation*, **3** : 3 – 22.

Rahman, A., Harding, A., Tanton, R., and Liu, S. (2013). Simulating the characteristics of populations at the small area level: New validation techniques for a spatial microsimulation model in Australia. *Computational Statistics & Data Analysis*, **57** : 149–165.

Riva, M. and Smith, D. M. (2012). Generating small-area prevalence of psychological distress and alcohol consumption: validation of a spatial microsimulation method. *Social Psychiatry and Psychiatric Epidemiology*, **47** : 745–755.

Rosenthal, T. (2012). Geographic variation in health care. *Annual review of medicine*, **63** : 493–509.

Ryan, J., Maoh, H., and Kanaroglou, P. (2009). Population synthesis: comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, **41** : 181 – 203.

Smith, D. M., Clarke, G. P., and Harland, K. (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A*, **41** : 1251–1268.

Smith, D. M., Pearce, J. R., and Harland, K. (2011). Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. *Health and Place*, **17** : 618–624.

Steel, D. and Holt, D. (1996a). Rules for random aggregation. *Environment and Planning A*, **28** : 957–978.

Steel, D. and Tranmer, M. (2011). Measuring and analysing homogeneity of geographical areas for a categorical variable. *Journal of Statistical Theory and Practice*, **5** : 649–658.

Steel, D. G. and Holt, D. (1996b). Analysing and adjusting aggregation effects: the ecological fallacy revisited. *International Statistical Review*, **64** : 39–60.

Tanton, R. (2011). Spatial microsimulation as a method for estimating different poverty rates in Australia. *Population, Space and Place*, **17** : 222–235.

Tanton, R. and Edwards, K. (2013). Introduction to spatial microsimulation: history, methods and applications. In *Spatial microsimulation*, pages 3 – 8. Springer, Dordrecht and New York.

Tanton, R., Vidyattama, Y., Nepal, B., and McNamara, J. (2011). Small area estimation using a reweighting algorithm. *Journal of the Royal Statistical Society, A*, **174** : 931 – 951.

Tanton, R., Williamson, P., and Harding, A. (2007). Comparing two methods of reweighting a survey file to small area data: generalised regression and combinatorial optimisation. *1st Gen. Conf. International Microsimulation Association, Vienna*, National Centre for Social and Economic Modelling.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46** : 234.

Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M., Martin, D., Steel, D., and Gardiner, C. (2005). The case for small area microdata. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **168** : 29–49.

Tranmer, M. and Steel, D. G. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A*, **30** : 817–831.

van Leeuwen, E. and Dekkers, J. (2013). Determinants of off-farm income and its local patterns: A spatial microsimulation of Dutch farmers. *Journal of Rural Studies*, **31** : 55–66.

Vidyattama, Y., McNamara, J., Miranti, R., Tanton, R., and Harding, A. (2013). The challenges of combining two databases in small-area estimation: an example using spatial microsimulation of child poverty. *Environment and Planning A*, **45** : 344–361.

Vidyattama, Y. and Tanton, R. (2010). Projecting small area statistics with australian spatial microsimulation model (spatialmsm). *Australasian Journal of Regional Studies*, **16** : 99 –126.

Voas, D. and 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. and Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical & Environmental Modelling*, **5** : 177–200.

Williamson, P. (2007). *Working Paper 2007/1 (v. 07.06.25): CO Instruction Manual*. Population Microdata Unit, Department of Geography, University of Liverpool.

Williamson, P. (2013). Chaper 3. An evaluation of two synthetic small-area microdata simulation methodologies: synthetic reconstruction and combinatorial optimisation methodologies. In *Spatial microsimulation*, pages 19 –47. Springer, Dordrecht and New York.

Williamson, P., Birkin, M., and Rees, P. H. (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.