National Centre for Research Methods
Methodological Review paper

Evaluations and improvements in small area estimation methodologies

Adam Whitworth (edt), University of Sheffield

NCRM National Centre for Research Methods

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Evaluations and improvements in small area estimation methodologies

Edited by Adam Whitworth, University of Sheffield

with

Grant Aitken, University of Southampton
Ben Anderson, University of Southampton
Dimitris Ballas, University of Sheffield
Chris Dibben, University of St Andrews
Alison Heppenstall, University of Leeds
Dimitris Kavroudakis, University of the Aegean
David McLennan, University of Oxford
Nick Malleson, University of Leeds
Graham Moon, University of Southampton
Tomoki Nakaya, Ritsumeikan University
Robert Tanton, NATSEM, University of Canberra
Joanna Taylor, University of Southampton
Nikos Tzavidis, University of Southampton
Paul Williamson, University of Liverpool

Table of Contents

Abstract

Small area estimation (SAE) of survey data down to small area level has become an increasingly widespread activity as scholars and policy-makers have sought to gain ever more detailed spatial information to better target interventions or resources and to evaluate local policy impacts. The availability of small area data has improved dramatically since the late 1990s yet many spatial variables of interest – income, fear of crime, health-related behaviours, and so the list goes on – remain impossible to access at small area geographies (i.e. beneath local authority level in the UK context). Various alternative methodologies have emerged to carry out SAE and these can be grouped broadly into statistical approaches and spatial microsimulation approaches, each with multiple differing approaches within them. A recent network, funded by the ESRC National Centre for Research Methods, brought together experts from across these methodological approaches and relevant external partners in order to enhance the state of the art in SAE through stimulating detailed comparative methodological discussion so as to better understand the strengths, weaknesses, similarities and differences between these methodologies. This methodological review paper emerges from the network discussions and aims to: summarise the main methodological approaches to SAE and their linkages; discuss the role of the small area covariate data and the opportunities and challenges around such data; identify the main methodological priorities around SAE in need of collective research attention; and, propose the need for a collective multi-methods SAE project in order more fully explore the conceptual and technical linkages between the statistical and spatial microsimulation methodologies

Introduction

Small area estimation (SAE) of survey data down to small area level has become an increasingly widespread activity as scholars and policy-makers have sought to gain ever more detailed spatial information to better target interventions or resources and to evaluate local policy impacts. The availability of small area data has improved dramatically since the late 1990s yet many spatial variables of interest – income, fear of crime, health-related behaviours, and so the list goes on – remain impossible to access at small area geographies in many national contexts. Within this context SAE methodologies have become increasingly demanded, increasingly used and increasingly refined. Yet the methodological landscape around SAE remains in need of attention in at least three key ways.

Firstly, various alternative SAE methodologies have emerged and it is often unclear to the 'new' researcher what these alternative approaches are, how they relate to each other and how they compare in terms of their estimation performance. These methodological approaches, discussed in greater detail below, can be classified broadly either as spatial microsimulation (which tend to be used by geographers predominantly) or statistical approaches (the use of which is dominated by statisticians). Secondly, despite recent advances in SAE methodologies there remain key methodological challenges are uncertainties to explore (e.g. how exactly can each method best be implemented in relation to weights, constraints, seeding, etc) as well as innovative methodological advances to bring together and extend (e.g. any role for agent based modelling, estimating distributional functions or spatially varying interactions). Thirdly, the different methodological approaches to SAE in large part operate in parallel to one another without a clear understanding of the conceptual and methodological linkages between them. This is particularly true between the statistical and spatial microsimulation approaches and greater understanding of the linkages between methodologies within these two differing approaches could support important contributions to the effectiveness of current SAE best practice. These issues form the focus of this review paper.

This paper emerges from the meetings of a one year network funded by the ESRC National Centre for Research Methods which brought together a wide range of methodological expertise from both the UK and internationally across alternative SAE methodologies in order to explore these issues and identify the key challenges, opportunities and priorities for SAE. Further details of the network's membership and events can be found in Appendix A. The network events, and the activity amongst network members between those events, have generated a range of outputs and resources (notes, podcast, R-library, PowerPoint slides, materials from a one-day training event) which will be of value to those researching or teaching SAE. These are available via the network webpage (http://www.shef.ac.uk/geography/staff/whitworth_adam/ncrm) and are openly available for others to use.

Overview of the main methodological approaches to SAE

This section seeks to provide this methodological orientation by providing an overview of the two overarching SAE methodological frameworks – spatial microsimulation and statistical estimation – as well as of the various distinctive approaches *within* these two overarching frameworks (Rahman 2008; Marshall 2012). The reviews are necessarily brief, seeking as they do to clarify the similarities and differences between the methodologies as well as the options open to the SAE researcher, though are designed to offer enough detailed that the informed non-specialist can gain a good sense of how the alternative methodologies operate in practice. References are provided throughout for those who wish to read further about a particular methodological approach.

Spatial microsimulation approaches

Spatial microsimulation approaches represent the first group of approaches to be discussed and three main approaches – iterative proportional fitting (IPF), combinatorial optimisation (CO), and generalised regression (GREGWT) – dominate the literature and these have been applied to diverse small area research projects in a wide range of national contexts (Voas and Williamson 2000; Ballas 2004; Ballas et al 2006; Nakaya et al 2007; Anderson 2007; Lymer et al 2008; Tanton and Vidyattama 2010; Zaidi et al. 2009; Rahman et al. 2010; Tanton et al. 2011; Birkin and Clarke 2011; Kavroudakis et al. 2013). In all three cases the task is essentially as follows:

1. take a survey in which there is the variable of interest that the researcher would like to take to small area level (the target variable, e.g. income) as well as a set of variables that relate to and help predict that target variable (the constraint variables). For those more familiar with multivariate regression it may be helpful to think of the target variables as the outcome variable of the regression model and the constraint variables as akin to explanatory variables of the model;
2. for each small area that the researcher wishes to create small area estimates of the target variable it is necessary to create tables of aggregated counts for each of the constraint variables, whether as one-way tables (e.g. number of men living in the area), as two-way tables (e.g. the number of men aged 25-34 in the area) or, if the data are available, as *n*-way tables. These data are usually, though not necessarily, obtained from national census data;
3. the task next is to adjust the weights attached to the survey cases (i.e. each person or household within the survey containing the target variable) such that the weighted aggregation of the survey cases sums as closely as possible to the actual small area counts of those constraint variables as calculated in Step 2 above from census (or other) data. The notion in doing so is that the weighted survey cases now collectively reflect a synthetic population micro-dataset of each small area, with the weights given to each survey case varying across the small area dependent upon the characteristics of those small areas;
4. the final step for each small area is to use the adjusted weights in the survey to calculate a weighted value (typically a mean, sum or percentage, though not necessarily as discussed below) of the target variable of interest (e.g. income).

The key differences between the three main spatial microsimulation approaches – IPF, GREGWT and CO – are seen in Step 3. IPF and GREGWT are sometimes described as deterministic approaches in that they do not involve any use of random numbers and give the same results each time. They differ however in that IPF reweights sequentially over the constraint variables one by one whilst GREGWT seeks to optimise the weights in one step. Hence, in both approaches it is the adjustment of the weights to precise non-integer values that achieves the fit o the small area constraint totals rather the selection of survey cases. In contrast, CO operates by selecting the

required number of individuals or households from the survey data according to the small area aggregates (i.e. if 200 households are in the small area then only this many are drawn from the survey) and then swapping those households with households not yet selected in an attempt to optimise the 'fit' between the final selection of households and the characteristics of the small area. Hence, CO is sometimes described as a probabilistic method in that a degree of randomness is involved, meaning that one would not expect to reach exactly the same results on each run.

The following sections offer further details to these three main spatial microsimulation approaches.

Iterative proportional fitting (IPF)
Dimitris Ballas and Ben Anderson

IPF relies on the relatively simple process of adjusting cell totals for small area tables given known margin (row/column) totals of the constraint variables derived from census or other small area data sources. The IPF algorithm itself is well known with a history of use in adjusting statistical tables to fit known margins of constraint variables. Its behaviour is well understood (Deming and Stephan 1940; Fienberg 1970; Wong 1992), it has been widely used (Ballas et al. 2005a; Ballas et al. 2007; Anderson 2007), and its statistical similarity with other techniques is discussed at length in Johnston and Pattie (1993).

The simplest situation in which IPF can be used is to adjust a simple two dimensional table derived from a sample survey so that the row margins and column margins fit other known values. For example, we may want to adjust the cell counts of a table of car ownership by tenure so that the margins fit a similar table from a larger survey, or from a census. Thus if we denote S_{ijt} as the cell count of row i and column j for the survey data at iteration t and C_i and C_j be the row and column margin totals from the known data sources (such as the census) then the cell counts at iteration t + 1 (adjusted to row margins) are given by:

Equation 1 $$S_{ij(t+1)} = (S_{ijt} / \Sigma_i S_{ijt}) * C_i$$

And at iteration t + 2 (adjusted to column margins) by:

Equation 2 $$S_{ij(t+2)} = (S_{ij(t+1)} / \Sigma_i S_{ij(t+1)}) * C_j$$

The process then iterates, repeatedly adjusting the cell counts until convergence is achieved. In survey re-weighting literature this is generally referred to as raking (Deville et al. 1993) or entropy maximization (Johnston and Pattie 1993) amongst others (Simpson and Tranmer 2005; Leyk et al. 2013). Agresti (2002) points out that this approach is essentially a loglinear iterative model fitting process and, as Wong notes, in theory the process will reach complete convergence at iteration n when:

Equation 3 $$\Sigma_i S_{ijn} = C_i \text{ and } \Sigma_j S_{ijn} = C_j$$

Under perfect conditions the process therefore satisfies the minimum information principle as it minimizes the information gain function (Macgill 1977). However, complete convergence, and thus minimal information gain, relies on margin totals being consistent and if multiple fitting tables are used this is not necessarily the case, especially in the small area context where record swapping or other disclosure protection-related perturbations may have occurred. In addition, any survey table cells with zero counts (where a small survey sample is being used for example) will also cause difficulties (see also Wong 1992).

If this can be done at the population level then clearly it can also be done at the small area level where instead of using national level known margin totals we can use margins obtained from a census or other covariate source (e.g. administrative or possibly commercial data source) for a given small area (also typically described as 'small area constraints'). In this context, instead of simply using the margin totals to adjust the cell counts of a table, the process is used to calculate a weight for each survey sample case for each small area and it is this reweighting process which is core to IPF. For example, if a survey file is to be re-weighted to fit each of 100 small areas then each survey case will be recalculated to take 100 (probably different) weights, one for each small area, once the whole process has finished, with each weight reflecting the extent to which that survey case 'fits' each small area across the set of chosen constraint variables. In this situation we calculate the weight for all cases in cell i,j at iteration t+1 (adjusting for row margins) as:

Equation 4 $$w_{ij(t+1)} = (S_{ijt} / \Sigma_i S_{ijt})$$

And at iteration t + 2 (adjusting for column margins) as:

Equation 5 $$w_{ij(t+2)} = (S_{ij(t+1)} / \Sigma_i S_{ij(t+1)})$$

If we chose only to use one table (e.g. car ownership by tenure) then we would simply iterate either until convergence (see Equation 3) or until a pre-defined stopping value of difference was achieved. In the more complex – and more typical – situation where a number of such small area constraint tables are used then we would iterate over each in turn before cycling back to the start to re-start the process, continuing this process until convergence. One benefit of this, and other spatial microsimulation approaches to SAE, is that for each small area a synthetic, reweighted population micro-dataset is created during the process which can be used for further analyses if desired.

The suitability of the variables and the number of small area constraints to be used in this process depends on the ultimate purpose of the estimation. For instance, if the aim is to estimate small area information on household income distributions, then it would make sense to select small area tables that include variables which are likely to be correlated with income (e.g. age, socio-economic group, number of cars, etc). Standard regression techniques can be used initially to model the relationship between the micro level constraints and what can be described as a 'target variable' or synthetic estimator (in the above example that would be 'household income') to be calculated in order to guide the selection of constraint variables (Anderson 2012; Ballas et al. 2007; Chin and Harding 2006).

Overall then, in contrast to IPF's traditional use in creating or adjusting survey table totals to match population margins (i.e. 'weighting up'), its use in the small area context is to 'weight down' since we are attempting to re-weight survey data to fit within given small area constraint (margin) totals. The following worked example shows how survey data can be reweighted using IPF to fit to small area constraint tables. More worked examples (as well as open source code) can also be found in Lovelace and Ballas (2013).

As noted above, the IPF algorithm itself requires two sets of tables for each constraint for each small area: the small area tables (typically, as below, drawn from census data, though not necessarily) for the constraints (Table 1) and the analogous small area tables constructed from the survey data by taking the weighted aggregation of the constraint variables using the adjusted weights (here with survey cases only for the relevant region retained for the reweighting) (Table 2). The constraints need to have identical definitions in the census and survey data and in this case

this meant adjusting some census totals: in this example, for example, it was necessary to assume that all household reference persons aged over 74 were retired from paid work since employment status was not asked of this age group in the census. In addition, the algorithm requires non-zero cell counts and so any zero counts in the census data were replaced with a fractional count (0.001) to ensure minimal disruption to the re-weighting calculations.

Table 1: Small area table for number of earners derived from the census for small area 1

Region	Number of households	Number of earners = 0	Number of earners = 1	Number of earners = 2	Number of earners = 3+
North East	784	397	221	142	24

Table 2: Small area table for number of earners derived from the survey for the relevant region for small area 1

Region	Number of households	Number of earners = 0	Number of earners = 1	Number of earners = 2	Number of earners = 3+
North East	1231	544	326	309	52

All survey household weights (w_i) are initially set to 1 whilst the weights of households that did not belong to the same region as the area in question were set to 0 rather than w_i to implement the regional weighting scheme. Then, for each constraint in turn, the weights were adjusted using the formula:

$$Nw_h = w_{ih} * c_{hj}/s_{hj}$$

where Nw_h was the new household weight for household h, w_{ih} was the initial weight for household h, c_{hj} was element hj of the census data table (Table 1) and s_{hj} was element hj of the survey table (Table 2).

Table 3: First four survey households with weights after fitting to constraint 1

Case	Region	Number of earners	W_1
26115	North East	1	= 1 * (221/326) = 0.6779
26116	North East	0	= 1 * (397/544) = 0.7298
26117	North East	2	= 1 * (142/309) = 0.4595
26118	North East	1	= 1 * (221/326) = 0.6779
..	..	..	..

As an example, using the number of earners constraint Table 3 shows the calculations of the first weights for the first four households so that the survey sample fits the census distributions on this one dimension. As can be seen the weights in this case are less than 1 (the number of census

households is less than the number in the survey region) because the final weights will need to reduce the weighted household counts (we have too many survey cases for the small area aggregates). Some cases are 'downweighted' more than others depending on how well, in a proportionate sense, the case matches the aggregated census count for this particular constraint.

Having passed over all constraint variables, repeating this process with the weight sequentially adjusted each iteration, the process then loops back to constraint one and repeats. Ballas et al (2005a) found that iterating the procedure between 5 and 10 times produced weights that reduced the error in fitting households to areas to a point where it no longer declined materially. Anderson (2007) suggested that 20 iterations were sufficient to achieve a stable indicator value. Others have suggested more formalised convergence thresholds (Tanton et al. 2011) but there remains no agreement as to the level at which this should be set. However many iterations are required, once this point is reached the simulation moves on to the next small area and repeats the process. The end result is a set of weights linking all selected survey households to all small areas in the sense that the weights represent the 'fractional existence' of the each household in each small area. Conceptually the results can be thought of as a matrix of small areas (rows) and households (columns) where each cell contains the weight for a given household in a given small area. Having completed the re-weighting process, calculating the 'target variable' or synthetic estimator (e.g. household income, number of households living in households that below a particular income threshold, etc.) is a straightforward matter of calculating weighted distributional statistics (e.g. means, percentiles, etc) as appropriate for each small area using the final calculated weights for each household and the values of the desired outcome variables for those cases.

Generalised Regression Reweighting
Robert Tanton

The Spatial MSM model is a spatial microsimulation model developed by the National Centre for Social and Economic Modelling at the University of Canberra (Tanton et al. 2009; Vidyattama and Tanton 2010; Tanton and Vidyattama 2010; Tanton et al 2011; Rahman et al. 2013). It uses a generalised regression reweighting methodology developed by Singh and Mohl (1996) and implemented by the Australian Bureau of Statistics in a SAS macro called GREGWT (Bell 2000). This procedure is normally used to reweight the observations in a national survey dataset to national benchmarks in order to ensure that the results from the survey match those national benchmarks. The SpatialMSM procedure uses this same technique to reweight the observations from the survey to a number of small area benchmarks, in line with previous work by Creedy (2003) and, in particular, Deville and Sarndal (1992). Specifically, GREGWT does so using the truncated Chi-square distance function to optimally match the margins within a single step (unlike IPF's sequential and iterative approach looping across constraints) and, in so doing, to optimally reweight the cases given those margins and that selected distance function. As with IPF, the end result is a dataset of adjusted weights for each survey household (columns) relating to each separate small area (rows).

Notable features of this approach are:

- It provides a survey file with a set of weights for each area, allowing cross-tabulations to be created;
- The weights are calculated using a number of generalised benchmarks, so one set of weights can be used to calculate a number of indicators;
- It uses a deterministic method so given the same input datasets and constraints the weights calculated will be exactly the same each time the procedure is run;
- It is highly parallelisable – one processor can be used per area estimated;

- As with all of the spatial microsmulation methodologies, it is an iterative approach and so is computationally intensive.

Constraint Selection
An important part of this method is the selection of constraints. As with other methodologies discussed, the constraints should be:

- Available on both the survey and small area data, defined in the same way, and with the same categories. Categories can be aggregated on either dataset to match them;
- Correlated with the final variable being taken from the survey;
- Correlated with any variables that the user wants to cross-tabulate this final variable with (age, sex, etc).

For best results, the benchmark tables should be cross-tabulations of the reliable small area data so that the cross-tabulations extracted from the final spatial microsimulation dataset will be as accurate as possible.

Refinements to the basic approach
A number of refinements have been outlined in Tanton and Vidyattama (2010) and include:

- Adding a number of constraint tables;
- Using univariate benchmarks rather than bivariate cross-tabulations;
- Limiting the source of households for the microsimulation to the broad area of estimation (i.e. using Sydney observations only to estimate areas in Sydney);
- Adding constraint tables adds specificity and subtlety to the process but also makes it more difficult for some small areas – particularly atypical small areas – to reach convergence. One option in these situations is to run the reweighting with all constraints initially and then for those small areas that do not reach convergence to gradually reduce the number of benchmarks constrained against one by one to deliver acceptable, if less subtle, results to the remaining small area each time.

Extensions: Projections and linking to tax/transfer models
One of the advantages of the SpatialMSM reweighting method is that projections can be easily created. The easiest way to achieve this is to inflate the weights in each area by the population growth by age and sex. The more complex method is to use population projections and labour force projections to develop projections for each benchmark table (see Vidyattama and Tanton 2010; Harding et al. 2011).

Because SpatialMSM provides new small area weights for a survey, if a Tax/Transfer microsimulation model uses the same base survey the small area weights can be merged onto the Tax/Transfer microsimulation model and the small area effects of a Tax/Transfer policy change on small areas can be calculated (see Tanton et al 2009 and Harding et al 2009). This 'what if' testing of the estimated impact of alternative policy scenarios on small areas is a benefit of SAE more broadly – and a key point of interest for spatially sensitive policy makers – as has been the subject of considerable work (Ballas et al. 2005b).

Combinatorial Optimisation
Paul Williamson

Combinatorial optimisation (CO) involves the selection of a combination of households from an existing survey micro dataset that best fit published small-area census tabulations (see Figure 1) (Williamson et al. 1998; Voas and Williamson 2000; Williamson 2002). In effect this is an integer reweighting approach in which most households are assigned weights of zero (i.e. not present) whilst selected households are assigned weights of one (i.e. present). Consequently, the 'correct' number of households is drawn from the survey as needed in the particular small area. Critically, selected households are then randomly swapped with remaining non-selected survey households and retained or returned to the survey according to one of several optimisation algorithms (e.g. hill climbing, simulated annealing, genetic algorithms) such that the pool of selected households comes to optimally reflect the small area margins across the range of constraint variables selected. The technique was originally devised to furnish inputs to dynamic microsimulation models that simulate the lifepaths of individuals and which cannot operate using fractionally weighted persons and households.

Although IPF, GREGWT and CO are therefore all variants of spatial microsimulation approaches to SAE it is worth clarifying two key differences between them. Firstly, CO is typically considered a probabilistic method in that a degree of randomness is involved in the household selection such that results will not be expected to be the same on each run. This contrasts with the deterministic approaches of IPF and GREGWT which involve no randomness and will therefore return the same results on each run. Secondly, IPF and GREGWT typically create fractional weights for all selected households. Usually all households in the survey are selected for reweighting although sometimes sub-samples of households can be selected according to, for example, the local region or geodemographic type[1] of the area in which the survey case lives (Smith et al. 2009; Birkin and Clarke 2012), to try to achieve more specific fitting). CO, in contrast, seeks to select the optimal combination of the 'correct' number of households needed for each small area with each of those households taking a weight of one and all other remaining survey households taking a weight of zero.

Notable features of this approach are that it:

- simultaneously satisfies a set of household and individual weights;
- resolves conflicts between constraints (e.g. as a result of differing sub-group small area totals across different small area tables) by producing weights that are the average of the conflicting constraints;
- has a high computational demand (the estimation process is highly iterative);
- is highly parallelisable (one processor per area being estimated).

Constraint selection

In general, the more constraints, and the more constraints that involve interactions between variables, the better. For example, Williamson (2002) reports the use of 817 constraints, drawn from 14 census cross-tabulations involving a total of 17 census variables.

Quality of estimates

Voas and Williamson (2000) report that combinatorial optimisation produces:

- good estimates of 'wholly constrained interactions' (i.e. of interactions included in the estimation constraints), even when many constraints are used;

[1] geodemographic classifications categorise areas into 'types' based upon their characteristics across selected variables and often using a method such as cluster analysis to group similar cases together.

- reasonable estimates of 'margin constrained interactions' (i.e. of interactions that involve table margins constrained as part of the estimation process);
- poor estimates of 'unconstrained interactions' (i.e. of interactions involving one or more table margins not constrained as part of the estimation process).

Refinements of the basic algorithm

Search strategy

Figure 1 below outlines a basic 'hill climbing' algorithm in which swaps of selected households for non-selected households are made only if they lead to an improvement in fit. Williamson et al. (1998) evaluated alternative approaches and concluded that a 'simulated annealing' approach, which allows 'one step back to take two steps forward', performs more effectively.

Stratified household selection

To improve the fit of the estimate it has been proposed that the households considered for selection from the survey should be limited to those that come from the same region, or the same geodemographic type, as the area being estimated. Williamson (2013) shows that both types of stratification can lead to poorer estimates for areas containing many persons atypical for the survey being sampled from (e.g. all student households) and that a better strategy is normally to maximize the survey records available for swapping.

Evaluation criterion for household swapping

Validation is a central issue in SAE. In Figure 2 Total Absolute Error is used as the criterion for judging whether or not a household swap should be made. This takes no account of the relative size of the error. Williamson (2013) shows that an alternative measure, Relative Sum of Squared modified Z-scores, provides better estimates.

In-sample sampling

For the very hardest to estimate areas (i.e. those most atypical of the survey average) it has been found beneficial, after a long phase of initial whole survey sampling, to switch to a final phase in which replacement households are sought only from those already within the selected sample.

Figure 1: A simplified Combinatorial Optimisation process

Step 1: Obtain sample survey microdata and small area constraints

Survey microdata (dataset 1)

Household	Characteristics		
	size	adults	children
(a)	2	2	0
(b)	2	1	1
(c)	4	2	2
(d)	1	1	0
(e)	3	2	1

Known small area constraints (dataset 2) [e.g. Census tabulations]

1. Household size (persons per household)

Household size	Frequency
1	1
2	0
3	0
4	1
5+	0
Total	**2**

2. Age of occupants

Type of person	Frequency
adult	3
child	2

Step 2: Randomly select *two* households from survey sample [e.g. Households A & E] to act as an initial small-area microdata estimate. (Two households because small area constraint 1 specifies that the area contains two households.)

Step 3: Tabulate selected households (*estimate*) and calculate absolute difference from *observed* small-area constraints

Household size	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
1	0	1	1
2	1	0	1
3	1	0	1
4	0	1	1
5+	0	0	0
		Sub-total:	*4*

Age	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
Adult	4	3	1
Child	1	2	1
		Sub-total:	*2*

Total absolute difference = *4 + 2* = **6**

Step 4: Randomly replace one of selected households with another household selected at random from the survey sample (e.g. replace Household A with Household D). Tabulate the new selection and calculate absolute difference from known constraints. If the new selection has a lower total absolute difference, then retain; otherwise revert to the previous selection.

Household size	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
1	1	1	0
2	0	0	0
3	1	0	1
4	0	1	1
5+	0	0	0
		Sub-total:	*2*

Age	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
Adult	3	3	0
Child	1	2	1
		Sub-total:	*1*

Total absolute difference = *2 + 1* = **3**

Step 5: Repeat step 4 until no further reduction in total absolute difference is possible.

Result: Final selected households: (c) & (d)

Household size	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
1	1	1	0
2	0	0	0
3	0	0	0
4	1	1	0
5+	0	0	0
		Sub-total:	*0*

Age	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
Adult	3	3	0
Child	2	2	0
		Sub-total:	*0*

Statistical approaches to Small Area Estimation

Joanna Taylor and Grant Aitken with Graham Moon and Nikos Tzavidis

This review and outline focuses on statistical approaches to small area (synthetic) estimation (SASE) and draws upon recent excellent summaries of the field (Rao 2003; Bajekal et al. 2004; Pickering et al. 2004; Marshall 2012). Its emphasis is on approaches that build from the generally well known standard, single-level regression model. SASE procedures which use either solely individual *or* area level covariates as well as those which incorporate *both* are all reviewed alongside a discussion as to why it is so important to simultaneously take into account both the individual and the area, and an indication of areas where the statistical approach to SASE is being extended.

Basics

The statistical approach to SASE is based on the regression model. Regression models enable the relationship between a characteristic of interest and explanatory variable(s) to be formally assessed. They estimate a target outcome (Y) using one or more predictors (X_1, X_2 … X_n) that aim to provide a good fit and minimise error variance. The target outcome is the variable for which we need small area estimates and predictors will typically include factors like age, sex and, socio-economic status. An additional error term within the model captures the extent to which there are other missing measurable or unmeasurable effects that impact on the outcome beyond the chosen predictors.

Translating the familiar regression model to a SASE scenario is superficially simple. SASE seeks to provide small area data on subject matter that is not otherwise available. For example in the UK there is no consistent, robust small area data on smoking behaviour; we do not have the data to 'map' smoking prevalence directly at small area scales. There are however other data sources that cover smoking behaviour, most obviously administrative records and surveys. These other sources provide candidate input data for the development of a regression model that 'predicts' smoking behaviour in that administrative/survey data given a chosen set of predictors. Provided the chosen predictor variables are themselves available at the required small area level we can take the coefficients derived from the model, apply them to the same covariates at the small area level and so rework the regression equation to generate small area estimates:

- Build model using survey or other data. $Y=\beta_0+ \beta_1X_1+ \beta_2X_2+ \beta_nX_n$
- Create SASEs: insert small area data on X_1, X_2 … X_n, for example from a census and combine with modelled coefficients.

The statistical approach to SASE is a form of indirect estimation. Direct estimation uses *only* the data collected in the survey – for example, an estimate of the prevalence of victimisation in a certain locality based solely on respondents who reside in that area. It does not enable the production of estimates for small areas where survey data are not available. This is typically a problem in circumstances in which many small areas tend not to be sampled within such surveys. Moreover, even where small areas do appear within surveys direct estimates suffer from large variances in the case of small area-specific sample sizes. Instead, indirect estimation takes findings (survey data) at a national level and using additional data extrapolates these results to the type of people living in the locality.

Indirect estimates such as those based on the statistical approach to SASE are design biased since what they estimate is the underlying *expected* value for any area given the socio-demographic independent variables included in the model and not the *real* value for the small area in question. For this reason literature reporting statistical SASE measures is usually at pains to stress that it is

reporting risk, chance, probability, estimates – all with a degree of (admittedly typically quantified) uncertainty – and *not* a direct measure of the construct of concern.

Individual level SASE models

Data to calibrate SASE regression models are usually drawn from large government surveys. The data in these surveys are for individuals but the objective of SASE is the provision of area estimates. So we need to get from individuals to areas. Individual level models work in two stages using regression modelling. Firstly, the survey data is used to predict the probability of the characteristic of interest based on the attributes of the individuals in the survey (such as gender, age and marital status). The aggregate levels of a cross tabulation of these individual characteristics for each local area are obtained, usually from the census, and the coefficients from the regression model are applied to those small area covariate values so as to calculate the expected value of the target outcome variable conditional on the area's characteristics. The steps are relatively straightforward:

- Ensure that the predictor variables are available in both survey data and for small areas
- Fit a regression model to survey data to predict the probability of chosen outcome
- Use Wald tests to consider dropping non-significant variables
- Extract parameter estimates and apply to small area data
- Sum to SASE

The choice of explanatory variables is based on three criteria. First, they must be measured in both the survey and the census. Second, the measure has to be consistent in the census and survey across the whole age range. Finally, the predictors should be assessed to see whether they are significantly associated with the outcome and it is common for non-significant predictors and interaction terms to be removed from the model. The modelled outcomes for each combination of chosen explanatory variables are then multiplied by the area counts of people in these population group combinations using census cross-tabulations.

There are several limitations to the individual level synthetic estimation methodology. The first is that the individual level data has to be measured in the same way in both the survey and the census and this is often not the case. Second, census data, of course, dates progressively after the census year, often leaving analysts with data from several years earlier but for which there is no obvious more recent alternative. Third, all the predictor variables must exist as a single census cross-tabulation. To preserve confidentiality, limits are in place in most countries for the number of covariates that can be included in any census cross tabulation and this disclosure control over the small area covariate data can place severe constraints on the number of individual predictors that can be used in a model. There *are* possible ways to increase the number for explanatory variables, but these are not unproblematic themselves: the use of census microdata is highly sensitive and if typically difficult to access; samples of census microdata which may be released lack comprehensive spatial coverage; commissioning a customised crosstabulation from the national statistical organization may be possible but often must be paid for; and estimating a multi-way cross-tabulation is possible but, of course, just an estimate (on which the research wishes to base another estimate). A further limitation of individual level studies is commonly referred to as the 'individualistic fallacy' whereby one assumes that individual-level outcomes can be explained exclusively in terms of individual-level characteristics. Indeed, even if one captures effects within those individual effects it is possible of course that these variables (and effects) may be (at least in part) drive by correlations with external variables at this or other scales.

Area level SASE models

Area level SASE models only use aggregate data, in other words average values or proportions relating to all individuals or households in an area. In the UK many social surveys are representative to the region level so one tactic is to develop an ecological regional model and then simply apply the model coefficients to smaller areas using census data by:

- fitting a model predicting the area variability of the dependent variable based on the aggregate values of the covariates;
- applying the model coefficients to the known local area values in order to predict the average local value for the dependent variable.

The disadvantage of area based models is that, by their very nature, they focus on area-based associations emphasising the context within which the behaviour or perception takes place. They ignore the many individual factors or indeed potential interactions between individuals and their surroundings. Furthermore, using only aggregate data crucially leaves results vulnerable to two classic conundrums of quantitative spatial statistics: the ecological fallacy and the modifiable areal unit problem (MAUP). By the ecological fallacy one refers to the danger of inappropriately using results at one geographical scale to say something about processes at a different scale. For example, a common example is that if one analyses area voting patterns then it is plausible to think that support for political parties opposed to immigration may increase as the percentage of the area population who are immigrants increases; after all, such parties tend not to get much support where there are no immigrants. However, to use relationships at this area level to infer something about processes at the individual level is not appropriate: it is not the case that individuals who are immigrants are more likely to vote for political parties that oppose the presence of immigrants. In the context of area level SASE the risk is that one is relying on areas based models to seek to model processes which very often take place at another scale, usually that of individuals. By MAUP one refer to the reality that the scale and location of boundaries may well affect the results from quantitative spatial analysis. In terms of their scale the questions then becomes whether it necessarily reasonable to apply coefficients from one scale to another scale and how do we know that we have selected to 'correct' scales for our SASE models?). Area based models are also relatively crude in the sense that all of the covariates, as well as the predicted outcome variable, are smoothed average values and do not offer distributional insights. Area based models do, however, have the advantage that, unlike individual level models, they are relatively undemanding in terms of their data requirements. A further benefit is that if the relationship between the independent and dependent variables is strong then estimates of good quality can usually be produced at a relatively low cost.

Individual and area level SASE models

Both individual only and area only models have been argued to lack reality. No account is taken of the connections between individuals and the localities where they live their lives: the concepts of composition and context. In contrast, SASE models that combine individual and area characteristics recognise that individual behaviours or outcomes are dependent on *both* place *and* personal characteristics, that is to say factors associated with both context and composition. Moreover, we have every reason to assume that smaller areas within regions will also vary and it would be helpful to be able to capture local variation more effectively. A multilevel modelling framework achieves this goal. Most of the national surveys that we can use to model target outcomes have a hierarchical design in which individuals are sampled within primary sampling units (PSUs – often postcode sectors) and then within regions. Effective modelling should include consideration of this structure, giving added impetus to the case for a multilevel approach to SASE.

The most straightforward multilevel model with individuals (level 1) nested within areas (level 2) is detailed below where *i* represents an individual in area *j,* x_j is a single area based explanatory and u_{0j} and e_{0ij} are the error or random terms at the area and individual level respectively.

$$y_{ij} = \beta_0 + \beta_1 x_j + u_{0j} + e_{0ij}$$

Incorporating just area level terms is referred to as the 'ONS method' (Heady et al. 2003) in the major UK Department of Health's review of SASE (Bajekal 2004) after the work of the UK nation statistical agency the Office for National Statistics. The probability of an individual being, say, a smoker, is modelled in terms of predictors at the area level; there are no predictors at the individual level. The resulting equation is then reworked using census data on all areas to generate SASEs, possibly adjusted using the area residuals. Area residual adjustment is not possible if not all areas feature in the survey dataset. The approach can be extended to include nested higher areas, for example wards in regions. An approach sometimes termed the Twigg method (Twigg and Moon 2002) extends the ONS approach to include a small set of individual predictors as well as area predictors within the multilevel framework. The choice of individual predictors is driven by the availability of multiway cross tabulations in census data where such cross tabulations effectively identify the numbers of individuals in small area with key characteristics. Models including these characteristics can therefore be used to generate SASEs for groups of individuals within a small area, taking into account the area level characteristics to which they are exposed. Age and sex are the obvious first choice for individual predictors. The Twigg method has the advantage of recognising that area constraints on individual outcomes are likely to be affected by individual characteristics. Multilevel models which incorporate both individual and area based measures in one model begin to address the ecological and individualistic fallacies. Further, by working at more than one level, not only is it possible to start separating out compositional from contextual differences but cross-level interactions can be incorporated to recognise the reality that environments affect individuals.

The multilevel SASE methodology has limitations. Firstly, where it utilises individual level data it suffers from the same data restrictions on the availability of census cross tabulations described earlier. Further, the process for estimating the standard errors and therefore confidence intervals is complex and, somewhat counter-intuitively, there is little evidence that the models which include both individual and area level indicators statistically outperform the more parsimonious, but still multilevel, models that only incorporate area level data.

Developments

The industry standard approach described above uses what are known as random effects models. For those with an interest in more recent, advanced or specific SAE approaches, other areas of research activity around SASE modelling during recent decades include (Ghosh and Rao 1994; Rao 1999; Pfefferman 2002; Rao 2003; Rahman 2008; Pfefferman 2013):

- Outlier robust estimation with the nested error regression model
- Spatial and temporal models (Pratesi and Salvati 2008)
- Non-parametric models (Opsomer et al. 2008)
- Empirical likelihood methodologies
- Measurement error models.

Outcome modelling estimation has also sought to move beyond averages and totals and towards:

- Estimation of percentiles of the distribution of survey variables using techniques such as M-quantile regression (Chambers and Tzavidis 2006; Tzavidis et al. 2010; Marchetti et al. 2012)
- Estimation of complex indicators (e.g. poverty indicators)

- Estimation with multinomial and count outcomes
- Estimation that incorporates survey weights for achieving design consistency
- Benchmarked estimates
- Ordered small area estimation
- Bayesian estimation (Ghosh and Rao 1994; Gomez-Rubio et al. 2010; Molina and Rao 2010)
- Development of SASE quality diagnostics

Software

The above approaches can readily be calibrated using standard multilevel software such as MLwiN, SAS, Stata, and so on. R routines have been written for implementing a range of small area methodologies. In a European context one collection of SAE routines has been developed under the SAMPLE (Small Area Methodologies for Poverty and Living Condition Estimates) Framework 7 project funded by the European Commission[2].

[2] http://www.sample-project.eu/en/the-project/deliverables-docs.html

Areas for development in current SAE practice: challenges and possible extensions

The methodological approaches outlined above represent the dominant ways in which SAE is currently conducted. In this section the focus shifts to consider three specific issues facing the SAE community that offer potential extensions and connections that may help to advance current methodological practice in SAE: linkages with agent based modeling; challenges with cross-national SAE research; and the lack of a commonly agreed software platform amongst SAE researchers.

Agent-Based Modelling: a possible complement to SAE
Nick Malleson and Alison Heppenstall

Sitting apart from the statistical and spatial microsimulation approaches to SAE, agent-based modelling has in the past decade or so grown as a flexible and powerful method for dynamic, spatially detailed analysis of social phenomena. Agent-based modelling (ABM) is an increasingly popular technique, having been described as a "breakthrough in computational modelling in the social sciences" (Gilbert and Terna 2000:60) and "one of the most exciting practical developments in modelling since the invention of the relational database" (Macal and North 2005:2). An agent-based model is comprised of virtual 'agents' who are able to behave autonomously, who exist in a navigable virtual environment (which is often spatial) and who are able to make decisions about what they would like to do in a given situation.

One of the areas of ABM that shows great promise is the linkage to SAE techniques, most obviously the spatial microsimulation approaches to SAE which create synthetic population microdata for small areas during their analysis as an intermediate step towards estimating the target variable of interest but which in doing so offer rich and realistic synthetic representations of small area populations. By linking this approach to ABM, researchers can exploit the ABM paradigm and explore impacts of policies on individuals within a dynamic and realistically modelled spatial framework. There is, however, little evidence of this in the literature to date and creating a population of agents from a synthetic population is not a trivial task.

This section focuses on outlining the main characteristics of ABM and suggests some areas where it could be used in conjunction with SAE to extend the potential of both methodological approaches that have to date been largely separate.

Agents

There are many definitions of the term 'agent' but the following are regularly applied:

- Autonomy: an agent should be free to control its own state, interact with other agents and its environment and make decisions without direct control from some central source. This is an ideal mechanism for modelling people, particularly when heterogeneity and individual behaviour is important;
- Heterogeneity: agents need not be identical. This allows for the incorporation of different qualitatively obtained data and theories;
- Reactivity: agents should be able to respond in a proactive way to changes in their environment;
- Bounded rationality: particularly with modelling in the social sciences it is important that agents do not always act perfectly rationally. Agents can be programmed with bounded rationality by limiting their knowledge of the world so that choices are not always perfectly optimal but, rather, a more accurate reflection of human behaviour (Castle and Crooks

2006).

One of the most substantial advantages of ABM is the 'natural description' (Bonabeau 2002) of a system which it provides. Social systems, whose behaviour is characterised by the behaviour/interactions of its individual components, cannot usually be described effectively by mathematical equations. One risk therefore is that often simplified assumptions are required which are often implausible or reduce the realism of the model (Evans 2011).

The Relationship between ABM and spatial microsimulation

There are a number of criticisms that can be levelled at spatial microsimulation (MSM) and the main ones can be summarised as follows (Birkin and Wu 2012):

1. MSM requires high quality and large datasets;
2. MSM models are computationally intensive;
3. MSM models examine the impact of policy but not the impact of individuals on policy;
4. MSM is typically weak in behavioural modelling.

High quality data is now much more widely available and large scale process intensive simulations are better supported by the computational abilities of contemporary hardware and software than in the past. However the robustness of the behavioural basis to MSM can still be questioned and even dynamic MSM cannot match ABM's capability to simulate the behaviours of each individual in terms of individual preferences, decisions, plans, and so on. Importantly, the interactions between individuals are not modelled in the MSM simulation. Thus ABM seems to offer considerable potential for simulation where a large number of heterogeneous individuals need to be simulated.

As Birkin and Wu (2012) highlight, it is perhaps more constructive to "view the relationship of ABM to MSM as one of complementarity rather than supremacy". ABM is a relatively new paradigm and can benefit from a relationship with a well defined and established methodology such as MSM. This rationale stresses that computational modelling is not just an applied tool but, in addition, a means for the production, testing and refinement of social theories. This view also allows for the development of more refined theories about social agents, for example moving away from static and unsophisticated views of individual actors which overemphasise either rationality or simple social learning as a basis of behaviour. We therefore advocate that the fusion of microsimulation and agent perspectives offers a powerful approach to the study of both social structures and social behaviours and can offer a great deal to one another through the potential to develop models with are rich in terms of both population attributes and behaviour.

Cross-national comparative SAE analysis: possibilities and challenges

Dimitris Ballas and Tomoki Nakaya

There are relatively few examples of cross-national research on small area estimation. However, small area estimation can be an important tool to overcome limitations that we often face when we compare detailed geographical aspects of different societies. One of these examples is the collaborative work between researchers in Britain and Japan (including ourselves) using IPF-based methodologies and simulated annealing techniques to estimate small area income distributions in the historic cities of Edinburgh and Kyoto (Ballas et al. 2012). In doing so our work revisits previous social geography research of class-based residential segregation comparing the two historic cities by Fielding (2004), drawing on and further developing recent and on-going small area microsimulation work in the two countries (Campbell and Ballas 2013; Hanaoka 2011; Hanaoka et al. 2013). Although both of the censuses in the two countries have small-area based socio-

economic variables such as occupational composition and unemployment, the concept of social classes and categories of related variables are different so that household-size adjusted household income measure, equivalised household income, was considered as a more direct measure of socio economic position in each country. It should be noted that since the both country's censuses do not have income information, we needed to estimate the income measure at a small areal level.

This research demonstrates the potential and power of spatial microsimulation to enhance comparative research of the social geography of cities internationally (in this case inequality), but also some of the challenges in doing so with respect to both data and methodological issues. A particular difficulty was the different definitions of concepts in the input data (e.g. household income, social class) and harmonised definitions and collection of social survey and census data between countries (in the same way that this is done for national level variables, for example, for OECD countries) would enhance the capability for comparative SAE (Ballas et al. 2012; 2013).

Another issue that needs to be considered when conducting comparative spatial microsimulation between different cities, regions and countries is the choice of spatial unit and potential issues pertaining to the Modifiable Areal Unit Problem (MAUP), an issue whereby the choice of spatial units themselves (either in their scale or their location) affect findings. For example, unlike output areas in the UK, Japanese small areas called 'cho-cho' used for detailed census tabulation have large variations in their areal and population sizes which makes fair comparison difficult (indeed, whether between themselves or with the UK's more population standardised geographies). In this context one interesting difference between the two countries is that the population census of Japan provides national 500m and 1km gridded data; Martin et al. (2011) argue around the viability and reliability of constructing similar gridded population from UK output area data. Indeed, small area estimation with data transformed form one geography of areal units into another geography – such as transforming irregular census units to grids – is a further research topic in its own right. Any further cross-national research needs to consider these issues around comparability of data (both survey and small area) and spatial units, with population gridded potentially offering one way to bypass concerns around MAUP and, in so doing, facilitate comparative SAE analyses.

Lost in translation: Do we need to speak the same language?
Dr Dimitris Kavroudakis, University of the Aegean, Lesvos, Greece

R is a powerful, flexible and freely available open source software package that is becoming increasingly popular amongst both academic and non-academic communities. R is the name of the programming language itself and RStudio is a convenient Graphical User Interface for executing R commands or R scripts. In common with most open source software platforms, a useful feature of R is that it is user-driven such that user-written programmes can be created oneself and downloaded from others. There is a lack of R libraries relating to spatial microsimulation and this library seeks to contribute to this gap. It is hoped that further developments will be added by other R users and that this will help both to move towards a shared programming language for SAE researchers as well as to facilitate new users to learn or use SAE techniques.

The current spatial microsimulation (sms) library uses a combinatorial optimization approach (CO) to optimize the random selection of a micro-dataset with characteristics that mach a macro, small area, description. The sms library contains functions for preparing micro-data from census and longitudinal datasets and these functions connect various data-sources and produce a small area population dataset. The functions use multi-core approaches of modern personal desktop

computers in order to simulate relatively large areas in reduced computational time and does so via the parallel interface of the R platform to divide the main simulation process into smaller simulations which then run in parallel. The library also offers a structural skeleton of key tabular and visual output for examining the success of the results and the intermediate states of the data fitting process. As noted earlier, this spatial R-library is freely available via the project webpage as described further in Appendix A.

The role of the covariate data: the foundations for SAE

Underpinning all of the SAE methodologies outlined above is the need for good quality, spatially detailed covariate data relating to the small area scale at or below that which the target variable of interest is desired. Internationally it is census data which tends (though not always) to form the bedrock for such data although in several countries, there is growing interest in opening up administrative and commercial data for SAE research purposes. Timely, accurate data across a range of key constraint variables and at sufficiently detailed spatial scales is key to the success of all of the SAE methods outlined above. Hence, this section focuses on the centrality of the two key sources of small area covariate data for SAE in most national contexts – the census and administrative records – and uses the current shifting and uncertain UK context in relation to these sources of covariate data to speak to the broader debate about the key role of covariate data and its implications for SAE.

The UK context is currently in a state of uncertainty and flux with respect to both census and administrative sources of small area covariate data. In terms of the census data, the national government statistical agency, Office for National Statistics, is currently conducting the Beyond 2011 project which has been tasked with returning recommendations to Parliament in 2014 around the viability of alternatives to the census for 2021 and thereafter in England and Wales, potentially heralding the end of the decennial census as the key source of small area data in these countries. At the same time, government, academia and funding councils alike are all aware of the significant potential that large, timely and spatially detailed administrative data can play in a wide range of academic research, including as a source of covariate data for SAE, and considerable emphasis, energies and resources are being expended to seek to 'open up' such data to research use. In this section we offer an overview of this changing landscape around the covariate data and consider their implications for SAE in terms both of the challenges and risks which such changes to the covariate data create but also in terms of a potentially enhanced role for SAE which they may also encourage.

The Beyond 2011 Programme at the Office for National Statistics and the future of socio-economic statistics[3]

Adam Whitworth, University of Sheffield

The UK Office for National Statistics is currently taking a fresh look at options for the production of population and small area socio-demographic statistics for England and Wales in the context that a future census (the next one would ordinarily occur in 2021) may not take place. The Beyond 2011 Programme has been established to carry out research on the options and to recommend the best way forward to meet future user needs.

The principle of Beyond 2011 is simple – the programme is investigating the best way of producing the population and small area socio-demographic statistics needed to support national and local decision making and the effective administration of the country. Beyond 2011 has been looking for options that might provide more flexible, frequent and richer data. A wide range of options have been considered and the overall process has been one of assessing the main alternatives, conducting public consultations and quality assuring the process. Clearly the viability of different options will vary across different national contexts and the Beyond 2011 project has assessed all of the possible options and approaches against a predefined and published set of criteria including

[3] Readers interested in further details about the Beyond 2011 project and the regular reports from the project are directed to the Beyond 2011 website at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/index.html

statistical quality, risk, cost, technical and legal feasibility, public acceptability and burden. The project has provided regular updates and gradually the focus is tightening as options are discarded. At the time that this review paper was published Beyond 2011 proposed two leading options – (i) an online census or (ii) the use of administrative data supplemented by a 4% rolling survey – and there are clear pros and cons to these two approaches in terms of quality, frequency and the nature of outputs and they each bring different risks. The final recommendations, which will be made in 2014, will balance user needs, cost, benefit, statistical quality, and the public acceptability of all of the options.

Clearly the census is a product which affects a wide range of users and stakeholders and any changes to the census would have implications in a wide range of ways given that the results will have implications for all population-based statistics in England and Wales and, potentially, for the statistical system as a whole. For anyone working with an interest in social and/or spatial statistics the outcome is clearly of central interest and our interest as SAE researchers in the census data is in one sense specific (we need accurate, timely, rich small area covariate data for our methodologies) but also reflect broader interest from other stakeholders – local government for example – who also require such information. For SAE researchers the interest is perhaps particularly acute because of the level of spatial detail that such methodologies require in their covariate data and the reliance to date on census data for those small area covariate data.

Although SAE (and other types of detailed geographical research) require small are statistics the financial case for ONS is not baseline funded to carry out the census. A business case has to be made to carry out each Census and a case is certainly needed to underpin the recommendations of Beyond 2011 – whether the eventual outcome is a modernised census or an alternative solution. In simple terms the minimum requirement for the census is to produce the statistics legally required and these are age-sex population data at local authority government level (equivalent to municipal governments in many countries). Given ongoing budgetary constraints in the UK public sector – as in many countries at present – anything beyond this will need to be strongly justified. The case for rich, broad covariate 'attribute' data for small areas is much more difficult to make at present in the UK context. ONS have already spoken to many users of local data in an attempt to identify cases where small area data are adding real, quantifiable, economic or social benefit. Without proper evidence it will be very difficult to justify the continued production of anything more than a very limited set of data below local authority level, whatever the recommended approach. Identifying the concrete, quantifiable benefits of small area data continues to be a key focus for the SAE community during this period.

Administrative data and SAE: options, possibilities and challenges
Chris Dibben and David McLennan

As outlined above, it is no surprise perhaps that one of the serious contenders as a potential new core data source for small area statistics within the ONS Beyond 2011 project is administrative datasets that are routinely collected by government departments during the course of their normal operations rather than data that are collected for the purposes of research. The census is clearly an extremely useful product for small area covariate data, especially for deriving population estimates and denominators from which rates or risks can be calculated, but it is limited by its range of variables and timeliness (only every 10 years). Spatially detailed administrative data is understandably a potentially extremely powerful alternative small area data source to be used within SAE, all of whose methodologies require accurate, rich covariate data at the small area level. Hence, it is understandable that the UK has witnessed something of a revolution during the last 15 years in the availability of small area covariate data derived from administrative data systems for use in small area synthetic data and simulation production.

In the UK, government departments are the main (although not exclusive) purveyors of large administrative databases, including welfare, tax, health and educational record systems. These datasets have for many years been used to produce official statistics to inform policy-making. The potential for this data to be accessed for the purposes of social science research is increasingly recognised, although as yet has not been fully exploited. Two areas of research – education and health – have seen fairly extensive use of administrative data, but most other administrative datasets have not been widely used for research purposes.

Administrative datasets are typically very large, covering samples of individuals and time periods not normally financially or logistically achievable through survey or even census methodologies. Alongside cost savings, the scope of administrative data is often cited as its main advantage for research purposes, though coverage is recognised to be imperfect. The lack of control the researcher has during the data collection stage and how this affects its quality, and therefore what can be done with the data, are the main problems for administrative data. More general concern has also been voiced about the lack of well-established theory and methodologies to guide the use of administrative data in social science research.

Opening up administrative data at the small area level: developments and lessons from the UK

Administrative data holds great research potential for SAE (and other) research in all national contexts although the research availability and use of such data varies significantly between countries. Linked to potential changes to the census in the UK, there are a variety of initiatives and recommendations that are driving forward the increased use of public (administrative) data for learning and research in the UK context and these shifts have relevance for all countries also reflecting on potential sources of small area covariate data for SAE.

The early work in this area in the late 1990s led to the setting up of 'neighbourhood statistics services' in the different parts of the UK[4]. The focus has now turned to processes that might allow researcher access to more potentially disclosive data. In the UK context the Administrative Data Liaison Service was established as a web-based portal in the late 2000s to facilitate researchers and policy makers to better understand the potential of administrative data for research purposes and to promote expertise around access, safety and ethics when researching using administrative data. As part of its service the site thematically details and describes many of the key administrative datasets across government, access conditions and processes, the spatial scale to which data are available, as well as research examples using the data. The Administrative Data Taskforce, formed in late 2011 to advise how best to make administrative data safely available for research use, has provided a range of recommendations for improving the research use of government administrative data (Administrative Data Taskforce, 2012). This included the development of Administrative Data Research Centres in each of the four countries in the UK in order to drive forwards the availability and use of administrative data in a safe and ethical manner.

The Shakespeare Review (2013) provides a set of recommendations to ensure that public sector information is used for the public good, whilst recognising the balance for individual privacy and citizen benefit and the current Coalition Government continues to push the agenda of increased access to 'public data' so as to ensure that we are, in its words, 'Unleashing the potential' (Cabinet Office, 2012) of such data. There is a government website for providing anonymised government data – www.data.gov.uk – including small area administrative data and the site is improving all the time with datasets being added and new ways to explore, visualise and filter data to enhance its usability. There is clearly forwards momentum around opening up administrative data more easily

[4] e.g. http://www.neighbourhood.statistics.gov.uk/dissemination (England), http://www.sns.gov.uk/ (Scotland), http://www.ninis2.nisra.gov.uk/public/home.aspx (Northern Ireland)

to researchers and this momentum may well gather ever more force if the Beyond 2011 project recommends that administrative data become the alternative data source to the traditional census for small area statistics in the future. Yet whilst administrative data certainly hold considerable potential for SAE (and other) research interested in geographically detailed analysis, reliance on administrative data presents some considerable challenges.

Disclosure risk at the small area level

Common concern around the use of detailed administrative data at the small area level are risks around confidentiality, anonymity and disclosure and this may lead to data controllers refusing to release the data or making it available within very controlled environments. An important consideration therefore for the release or publication of administrative data at individual or aggregate small area level is that the identity of individuals is protected. Assessment of disclosure risk is a complex process. Generally the more detail the data has and the higher the proportion of the population of interest that is captured in the data then the higher the risk. Intruders seeking the identity of an individual would either use some large scale information source to match against anonymised files or identify an unusual record in the dataset and attempting to identify that person in the population.

A recent disclosure risk report prepared by the Administrative Data Liaison Service (ADLS) of complete records of fire data between two stated dates held by the UK Department of Communities and Local Government provided examples of disclosure risk problems. Any third party who had response knowledge that there had been a fire at a certain location (through visibility, media or online reports for example) would be able to use that information with other information held within the dataset to try and identify an individual or address. Geocoding of fire data was also problematic in that it could be unique (particularly in rural areas) and this could then be used for re-identification purposes from response knowledge or other data available. It is also important to pay attention to unusual (e.g. unusual ethnicities) or zero values within data as these could also be used for identification purposes, particularly if released with detailed geographic information.

Administrative data linkage

There are various ways in which extracts of administrative data can be linked with other data sources to create more comprehensive and powerful datasets for analysis (both in terms of cases and variables). The most obvious is the linkage of different years of data within a data source. This is frequently done with datasets like the National Pupil Database (a database containing information on pupils at maintained schools including their examination results) where the same individuals can be identified in annual cuts of a data source for many years and a longitudinal record for individuals can therefore be created. However, administrative data can also be successfully linked with a variety of other data sources – to census records, to other administrative datasets, or to survey datasets – via a unique identifier or fuzzy matching methodologies (matching personal details like names, date of birth, address etc). However, administrative databases tend to be largely department or function specific and there has been little linkage between different datasets. Linkage often exacerbates risks around data disclosure and potential breaches of anonymity and this needs to be carefully managed by the appropriate use of secure data centres and training of research users.

Changes to the covariate data and implications for SAE

As outlined above, the UK context presents a changing landscape in terms of census and administrative covariate data sources which together offer the central foundations for virtually all SAE methodologies. The two leading options within the Beyond 2011 project – a shift to an online census or to reliance on administrative data supported by survey calibration – each present cost

risks to the quality and completeness of the type of small of small area statistics that have historically been available from the traditional census. Hence, such changes may open up greater opportunities for SAE methodologies to contribute small area estimates at the same time as weakening the covariate data foundations on which such estimates will have to be built. Administrative datasets, whether in isolation or whether linked to other (admin, census or survey) data, clearly offer a potential valuable addition to our traditional reliance on census data for small area covariate statistics of the sort required by all of the statistical and spatial microsimulation techniques discussed in this report. This may become more true if it is decided that England and Wales will no longer continue to run a census in the future, with administrative data one of the potential alternatives under consideration. Within this context the various efforts outlined above to increase access to such routine administrative data for research purposes are critical, as is greater awareness of issues and processes to manage key legal, safety and ethical issues around administrative data sharing. At the same time, and even if perfectly shared and perfectly linked, it is known that administrative data do not offer complete coverage – either in terms of population or variables of interest – and more widespread use, and quite possibly dependence upon, administrative data for research purposes will inevitably continue efforts to tackle these issues where they exist. More broadly, routine administrative data at present remain under the remit, and therefore under the control of, the individual departments themselves rather than, for example, ONS, raising important, and difficult, legal and policy issues around what the function and ownership of such data are in the modern era's desire to open up researcher access to such data. SAE researchers will have to remain flexible to these potential changes in the covariate data in the years to come.

Identifying gaps and setting out future priorities

A key aim of this review paper is to identify existing gaps in current SAE practice and to highlight methodological priorities for the SAE to work together to overcome in the medium term. These gaps and priorities can be differentiated into two separate types:

- to identify gaps and priority areas *within* each methodological approach, or relating to groups of methodologies, that are in need of attention and development;
- to identify gaps and priority areas *between* different SAE approaches in order to explore methodological connections, similarities and differences between the approaches which a view to using that knowledge to improve existing approaches. In particular, there is an interest in considering possible connections between the statistical and spatial microsimulation approaches and reflecting on whether greater awareness of any such connections can advance existing methodological practice.

Identifying priorities within the main SAE methodologies

In this section we focus on the first of these two areas, namely the identification of gaps and priorities *within* the various main methodological approaches SAE

Distributional estimates as well as means
It is typically the case that SAE analyses focus on estimating point values – usually mean or medians – rather than seeking to unpick distributions around those point estimates. Yet such distributional information is useful in and of itself to give a clearer sense of the estimated profile of a variable's values within a small area. When using SAE within 'what if' policy scenario testing, moreover, we may well want to explore distributional impacts of those policy scenarios within small areas as well as any impacts of point values. Certainly some SAE research does estimate distributional values (e.g. various percentile estimates) but this is relatively rare and should in our view be encouraged. In spatial microsimulation approaches doing so are relatively straightforward given that once the reweighting has been completed the researcher has flexibility to use the reweighted synthetic population micro-data to pick up whatever distributional values of the target variable are desired. Within a statistical framework approaches such as M-quantile regression can also be used to estimate distributional values within SAE (Chambers and Tzavidis 2006; Tzavidis et al. 2010; Marchetti et al. 2012). This issue is discussed further in the next section.

Estimating variance and 'confidence intervals' within spatial microsimulation approaches
Statistical methodologies place the estimation of variance and confidence intervals at their heart and, similarly, statistical approaches to SAE almost always provide such estimates alongside any point estimates so as to offer an indication of the plausible range within which such point estimates might realistically fall. In doing so such approaches highlight that one of the challenges for SAE, particularly as it relates to policy applications, is the width of these intervals and the reality that there is often considerable uncertainty in our estimates at small area level. In this context a common challenge is not just the width of intervals around point estimates but also the overlap of intervals between different small areas which makes it difficult to clearly differentiate small area values/ranks if one wishes to take into account the uncertainty of the estimates in addition to the values of the point estimates themselves.

Such variance estimates do reflect the cautious reality of most SAE claims yet the estimation of variance and confidence intervals – or, rather, what are more usually referred to as 'credible intervals' – are not commonplace within spatial microsimulation approaches where the estimation of a single point estimate without such information about uncertainty is widespread. Some spatial

microsimulation approaches do seek to build credible intervals and this is typically done through boot-strapping estimates for each small area. One of the challenges in this approach however is the computational intensity (and hence time) of building such credible intervals (meaning that for practical purposes most such intervals are based on fewer than 50 iterations for each small area) as well as the increased risks around non-convergence within a larger set of iterative loops. An alternative route, discussed further below, comes through the recognition of similarities between the statistical and spatial microsimulation approaches such that it may be possible to exploit these connections to more efficiently build credible intervals around estimates stemming from spatial microsimulation approaches based on improved understanding of their statistical equivalences. However such estimates of variance and confidence/credible intervals are constructed analysts should in our view be encouraged to provide them alongside any point estimates.

Clearer understanding of the impact of incorporating geodemographics into SAE

As noted in some of the earlier sections, one relatively common way to seek to add local specificity to small area estimates is to incorporate knowledge of geodemographic classifications into SAE (Smith et al. 2009; Birkin and Clarke 2012). Within a spatial microsimulation approach this is most commonly done at the point of survey case selection whereby one, for example, only retains cases within the same geodemographic types as the target small area. Within a statistical approach, most obviously a multi-level framework, one approach is to make random error terms specific to different geodemographic types so as to 'adjust' each small area's estimates in a manner reflective of the type of area which they are. Whilst intuitively plausible there has to date been little work exploring the empirical effects of such approaches. The issue seems less problematic within the statistical approach. Within the spatial microsimulation approaches, however, the added specificity of geodemographically specific sub-sample selection comes at the price of a reduce sample size. As noted above, Williamson (2013) finds that the more effective strategy is usually to maximise the number of records available rather than to seek to tailor the selection of survey cases (usually either via geodemographic type or through retaining only those cases in the small area's region) and we suggest further attention be paid to assessing the impact of attempts to incorporate greater 'local' specificity (whether through sub-sample selection within spatial microsimulation or through tailored random error terms within multi-level modelling frameworks).

Greater consideration of n-way outcomes

Estimating one-way target variables – by which we mean an outcome such as poverty (poor/non-poor) or ill-health (ill/well) – dominates current SAE practice. Estimating such target variables is a valuable task and may well be sufficient for many purposes. In addition, however, we feel there is scope for greater attention to be placed on two-way (and possibly *n*-way) outcomes (e.g. poor old people, non-poor old people, poor young people, non-poor young people) in order to add greater depth and specificity to the target variables being estimated. This may be particularly pertinent within 'what if' policy scenarios where one may well wish to know now just what the distributional impacts of a proposed reforms is expected to be (quite possibly using distributional as well as point estimates as discussed above) but also what the effects are expected to be across different groups within the small areas.

Clearer understanding of interactions between variables

Much SAE analysis is based on modelling with (statistical approaches) or constraining to (spatial microsimulation approaches) single variables. A further issue however is to seek to incorporate greater recognition of the role of interactions between variables (i.e. which variables vary spatially the most) and of the geographic scales at which these interactions are most powerful.

Software comparability

One challenge for collaborative or cross-methodological SAE research is the lack of a common software platform on which SAE is carried out, making it more difficult to collaborate and share coding as well as having uncertain (though probably not material) consequences for results (all software operates slightly differently). As outlined by Kavroudakis above, one option in this context is to consider the value of moving towards a situation of a shared software language. Various software tools are commonly used by SAE researchers – R, Stata, Java, Fortran, SAS, SPSS – and of these the momentum is perhaps shifting towards R as a potential platform for the future.

Small area covariate data

Typically it is sufficient to have aggregated counts at small area level (e.g. census tables) although for statistical approaches such as M-quantile regression which seek to model distributional outcomes micro-data at the small area level are required (usually anonymised individual census records). In some form, however, all SAE techniques require detailed small area data and, as discussed above, the availability of small area covariate data – census, administrative or potentially from other sources (e.g. commercial, social) – is changing in many contexts.

Related, it is important that researchers retain access to key secure data environments to access the census, geocoded survey data and, if necessary, administrative data. Given the push for increasingly open access and the desire to exploit the potential of such data it is important that access to secure environments and processes remains and, where needed, is put in place to enable researchers to access, in a secure and efficient manner, the detailed datasets which high quality research requires.

Starting weight

One issue specific to spatial microsimulation approaches which arose was the value at which the initial starting weight should be set before the reweighting process begins. There is a lack of agreement within the literature as to this issue with two different options – a starting weight either of one or set to the case's initial weight – both being used in different pieces of research. Yet the choice of starting weight is material, both in terms of the efficiency of the reweighting and to accurately capture interactions and compositional differences across the constraint variables. It is not currently clear precisely what the impact of differing starting weights is and further research to establish this issue is welcome. Such issues suggest however that a starting weight of one would not accurately reflect compositional differences across variables.

Validation criteria

Given that SAE is essentially concerned with the estimation of data to small area level where that data does not currently exist, a perennial, yet critical, issue is the process of validating whether estimates appear realistic. Despite being a central part of any SAE analysis, there remains a need to clarify more standardised and commonly accepted forms of validation both within the separate methodological approaches but also, more clearly, so that one could in principle compare the performance of statistical and spatial microsimulation approaches to the same task. Both internal and external validations are usually conducted and there are issues to develop in both areas (Edwards et al. 2011; Rahman et al. 2013; Williamson 2013).

Internal validation:

- For statistical approaches this largely involves model diagnostics (model fit, outliers and leverage, satisfaction of assumptions) to assess the acceptability of the model in and of itself, though not formal, justified thresholds of acceptability exist around these issues;
- For spatial microsimulation this largely involves an assessment of how closely the benchmark totals for the small area are hit both across the constraint variables and, more

challenging, across variables not included as constraints within the reweighting. As noted earlier, Total Absolute Error (TAE) is commonly used as the central criterion for internal validation. Williamson (2013) highlights however that this takes no account of the relative size of the error and instead suggests that an alternative measure, Relative Sum of Squared modified Z-scores, provides better estimates;

- Reaching consensus as to the most appropriate tests of internal validation may help to standardise the process of validation and it is also not clear how tests of internal validations could be compared between statistical and spatial microsimulation approaches, although this is perhaps less widely needed.

External validation:

- this can involve comparing estimates to already existing similar and/or highly correlated variables also at the small area level (e.g. estimated income against already measured deprivation), though it is often challenging to find appropriate external variables at that small area scale (that is the reason for the SAE typically);
- a common strategy is to calculate direct estimates from survey data at higher geographical scales (local authority or regional level in the UK context for example) and then to aggregate up the small area estimates to this higher scale and compare the correlation the two. Alternatively, alongside the SAE one can estimate the target variable to this higher scale simply for the purposes of validating the small area estimates;
- one might also ask local practitioners or experts to 'ground truth' the estimates based on their detailed local knowledge, perhaps focusing on particular case study areas;
- there are no clear guidelines as to what types of external validation should be conducted nor of objective criteria against which to assess them.

Constraint selection within spatial microsimulation

The selection and ordering of constraint variables within spatial microsimulation techniques is a common area of uncertainty yet many of the decisions in this area depend upon the specific technique being used and the nature of the particular estimation task in hand. In principle, one is looking for constraints which correlate with the target outcome variable such that the estimation process successfully discriminates between different small areas. This in many ways reflects the principles of multivariate regression and, as Anderson (2012) and others note, regression models can be useful in guiding the choice of constraint variables. The number of constraints depends in part upon the degree of correlation between different constraint variables (highly correlated constraints represent unnecessary duplication whilst weakly correlated constraints may require all to be retained) as well as on the ability of the particular method to deliver results with a full set of constraints as this tends to strengthen conflicts between optimisation across different constraints and hence can make convergence more challenging. As noted above, this is an issue for a method such as GREGWT where atypical small areas can fail to converge with a full set of constraints, leading to the gradual simplification of the set of constraints in order to achieve convergence in these areas. For a method such as CO, in contrast, no such issue arises and a large number of constraints can be used without similar difficulties. The ordering of constraints within the reweighting processes of spatial microsimulation techniques is an area that would benefit from further clarification and whilst constraint order does matter – and should therefore be thought through – this differs according to the technique being used.

Pushing the boundaries of SAE: Exploring connections between statistical and spatial microsimulation approaches

The second area of potentially fruitful methodological progression is to consider conceptual and technical similarities, differences and possible linkages between the statistical and spatial microsimulation methodologies more deeply. Such a focus is rare within the literature and this section presents an overview of the main linkages and related opportunities followed in the next section by a proposed plan to explore these issues more fully.

Identifying linkages: some of the methodologies, but not all

In terms of connections between the methodologies, and as others have noted previously (Simpson and Tranmer 2005), iterative proportional fitting (IPF) can equivalently be expressed as a log-linear model. Similarly, GREGWT comes from within the generalized regression framework though can also be understood as calibration against known counts (margins) for small areas within a fixed matrix of variables in order to seek to optimise weights for each case (typically individual or household)(Deville and Sarndall 1992). Some spatial microsimulation approaches, however, may be more similar and better suited to connections with statistical approaches than other spatial microsimulation methodologies. Combinatorial optimisation, for example, is about randomly drawing households and so is a different type of approach to a more 'deterministic' IPF approach, with implications for potential linkages to the spatial microsimulation approaches.

Different outcomes, equivalent methodologies?

One interesting issue is that whilst IPF/log-linear modelling and GREGWT/generalised regression might be considered as equivalents they differ in what it is that they each estimate. Within the statistical approaches the 'model' estimates the values of the target variable itself (e.g. small area income/health/crime/etc estimates). For the spatial microsimulation approaches, in contrast, the 'model' estimates weights – a sort of intermediate step – and these weights are *then* used to estimate the target variable of interest at small area level. It is not presently clear whether the approaches are therefore conceptually or methodologically equivalent in their specifics or whether this difference in model outcome (target variable versus weights) is material in terms of the point estimates or in terms of any distributional estimates of the target variables. Further work, suggested in the next steps below, will be needed to test this issue.

Possibilities around credible intervals for spatial microsimulation approaches

Nevertheless, the similarities – and possible equivalences – between the methodologies offer possible opportunities for cross-fertilisation of approaches in order to improve the power and efficiency of the methodologies. One way in which gains might be made is in terms of thinking about confidence intervals/credible intervals or, related, estimates of variance around the point estimates. As discussed above, an area for development within spatial microsimulation approaches is to support the calculation of credible intervals around point estimates in a more computationally efficient manner. In an IPF framework, for example, boot-strapped estimates are an obvious way to seek to establish such credible intervals but this is computationally demanding and time consuming; in a log-linear regression framework, in contrast, one would tend to think about deriving the intervals statistically within a single step. It may be, therefore, that improved understanding of the statistical links between IPF and log-linear modelling can help in this respect. Indeed, these improvements may both be in terms of computational time but also in terms of accuracy of those variance estimates (it is not clear without further testing whether credible intervals based on boot-strapped estimates around a point estimate from IPF would necessarily be the same as the statistically 'equivalent' intervals). The same could be said for GREGWT in terms of its foundations in a generalised regression framework.

Possibilities around distributional estimates

A further potential link between statistical and spatial microsimulation approaches is in the desire for distributional estimates of target variables rather than simply mean or median point estimates. As noted above, this can be achieved within a statistical framework by approached such as M-quantile regression (given suitable statistical expertise) but can also be done relatively simply and intuitively – albeit very differently – within spatial microsimulation approaches. Whilst spatial microsimulation approaches pick up these estimates empirically from the reweighted microdata, in statistical approaches these distributional estimates are derived from model-based theoretical distributions of the target variable for the small area. It is not clear whether the two approaches are conceptually or empirically equivalent nor whether these approaches might differ in their results. This warrants further attention. It is also particularly pertinent given that a spatial microsimulation approach to the calculations of distributional estimates seems to offer to possible advantages over an M-quantile regression approach. Firstly, the spatial microsimulation approach is considerably simpler to implement in terms of the level of statistical expertise required and, despite the need to be able to programme, would be expected to therefore be more accessible to quantitatively skilled but not statistically expert researchers. Secondly, creating distributional estimates in a statistical framework requires individual-level covariate microdata for each small area (typically census microdata). This is an intensive data requirement and can be a restriction if not available, or if all of the desired variables are not available. In a spatial microsimulation approach, in contrast, only aggregate counts for the small areas are required, significantly reducing the covariate data demands and confidentiality (and hence data access) concerns. Hence, an important question is whether the less data demanding spatial microsimulation approach to distributional estimates results in the same (or similar) estimates as the statistical approach. There are doubts over the ability to reliably estimate distributional values without microdata though this requires further testing both in and of itself as well as within a broader process of seeking to understand the conceptual and technical similarities and differences between these approaches.

(Stalled) possibilities around the small area microdata

Related, a further possible connection between statistical and spatial microsimulation approaches regards the small area microdata itself. Census data are the obvious source yet a by-product of spatial microsimulation approaches is the creation of synthetic population microdata for each small area. It is not clear to what extent spatial microsimulation could play a role in providing detailed small area covariate data which might be used as source data for M-quantile regression (which require such data for distributional estimates). However, this is problematic in that the rolling down of the small area microdata to smaller scales within the spatial microsimulation step inevitably loses and smoothes out much of the differing (and, indeed, random) variation in relationships across space (given that it is based on a model) yet this is precisely what one wishes to pick out during the next statistical SAE phase.

Reaching the tails of the distribution

A final possible linkage between statistical and spatial microsimulation methodologies relates to the desire to include information from the tails of the distributions so as to avoid overly smoothed estimates. In a statistical framework one approach is to add ‘noise’ to our estimates in order to achieve this aim but this does not tend to happen in spatial microsimulation approaches. Yet if (some of) these methodologies can be expressed in a statistical framework then there seems the possibility of seeking to mimic the introduction of noise within, for example, IPF or CO in order to seek to incorporate the tails of the distributions more effectively into our estimates, potentially enabling clearer differentiation between small area estimates.

Next steps: A collaborative multi-method project

The above highlights that there is much potential around methodological learning and advancement by more fully and more explicitly understanding how the statistical and spatial microsimulation methodologies relate to one another in the details of their operation. The review paper has identified key issues, priorities and possibilities around comparative methodological learning between the two broad approaches to SAE and, in so doing, proposes an agenda for a larger, deeper, empirically-grounded comparative methodological research project. It is proposed that this agenda should begin with a collaborative multi-methods project harnessing differing methodological expertise available.

The intention in doing so would be to test many of the questions identified above by creating a collective project in which specially designed, methodologically neutral synthetic data are used by a number of differing SAE methodological approaches to work in parallel towards a shared estimation task. The aim would be to identify a single research question and estimation process at a set scale and to then run different methodological approaches in parallel and to compare results at both final and intermediate stages. The aim in doing so is not only to compare the comparability of estimates (whether point, distributional or variance estimates) but, importantly, to seek to identify the underlying conceptual and methodological linkages, similarities and differences between the different methodologies. This would be a highly innovative collaborative methodological project with rich opportunities for methodological learning due to its wide range of SAE methodologies spanning statistical and spatial microsimulation techniques and its explicitly comparable, standardised and rigorous approach. It is hoped this would begin to unpick many of the complex conceptual and empirical questions underpinning the comparative methodological evaluation of these key SAE techniques outlined above and that it would lead towards the realisation of advanced inter-disciplinary methodological learning.

References

Administrative Data Taskforce (2012) *The UK Administrative Data Network: Improving Access for Research and Policy*.

Agresti, A (2002) *Categorical Data Analysis*. London: John Wiley & Sons.

Anderson, B (2007) Creating *small area income estimates for England: spatial microsimulation modelling*, a report to the Department of Communities and Local Government. London: Department of Communities and Local Government.

Anderson, B (2012) 'Estimating Small Area Income Deprivation: An Iterative Proportional Fitting Approach' in Edwaerds, K and Tanton, R (eds) *Spatial Microsimulation: A Reference Guide for Users*. London: Springer.

Bajekal, M., Scholes, S, Pickering, K and Purdon, S (2004) *Synthetic estimation of healthy lifestyles indicators: Stage 1 report*. London: NatCen

Ballas, D., Dorling, D., Nakaya, T., Tunstall, H and Hanaoka, K (2013) 'Income inequalities in Britain and Japan: a comparative study of two island economies', *Social Policy and Society*, (in press, published online before print, 13 March 2013; doi:10.1017/S1474746413000043).

Ballas, D., Campbell, M., Clarke, G., Hanaoka, K., Nakaya, T and Waley, P (2012) 'A spatial microsimulation approach to small area income estimation in Britain and Japan', *Studies in Regional Science*, 42, pp163-187.

Ballas, D., Clarke, G., Dorling D and Rossiter, D (2007) 'Using SimBritain to Model the Geographical Impact of National Government Policies', *Geographical Analysis*, 39(1), pp44-77.

Ballas, D., Clarke, G. and Wiemers, E. (2006) 'Spatial microsimulation for rural policy analysis in Ireland: the implications of Cap reforms for the national spatial strategy', *Journal of Rural Studies*, pp367-378.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B and Rossiter, D (2005a) 'SimBritain: A Spatial Microsimulation Approach to Population Dynamics', *Population, Space and Place*, 11, pp13-34.

Ballas, D., Rossiter, D., Thomas, B., Clarke, G. and Dorling, D (2005b) *Geography matters: simulating the local impacts of national social policies*. York: Joseph Rowntree Foundation.

Ballas, D. (2004) 'Simulating trends in poverty and income inequality on the basis of the 1991and 2001 census data: a tale of two cities', *Area*, 36(2), pp146-163.

Bell, P (2000) *GREGWT and TABLE macros - Users guide*. Canberra: Australian Bureau of Statistics.

Birkin, M. and Clarke, G. (2012) 'The enhancement of spatial microsimulation models using geodemographics', *Annals of Regional Science*, 49, pp515-532.

Birkin, M and Wu, B (2012) 'A review of microsimulation and hybrid agent-based approaches' in Heppenstall, A., Crooks, A., See, L and Batty, M (eds) *Agent-based models of Geographical Systems*. Dordrecht: Springer.

Birkin, M. and Clarke, G. (2011) 'Spatial microsimulation models: a review and glimpse into the future', *Population Dynamic and Projection Methods: Understanding Population Trends and Processes Volume 4*. London: Springer.

Bonabeau, E (2002) 'Agent-based modeling: Methods and techniques for simulating human systems', *Proceedings of the National Academy of Sciences 99* (90003), pp7280–7287.

Cabinet Office (2012) *Open Data White Paper: Unleashing the Potential*. London: The Stationery Office, Cm 8353.

Campbell, M and Ballas, D (2013) 'A spatial microsimulation approach to economic policy analysis in Scotland', *Regional Science Policy and Practice* (in press, published online before print, 16 May 2013; doi.wiley.com/10.1111/rsp3.12009).

Castle, C and Crooks, A (2006) *Principles and Concepts of Agent-Based Modelling for Developing Geospatial Simulations*. London: Centre for Advanced Spatial Analysis, UCL Working Paper Series.

Chambers, R and Tzavidis, N (2006) 'M-quantile models for small area estimation', *Biometrika,* 93, pp255-268.

Chin, S.-F and Harding, A (2006) *Regional Dimensions: Creating Synthetic Small-area Microdata and Spatial Microsimulation Models*. University of Canberra: National Centre for Social and Economic Modelling, NATSEM Technical Paper no. 33.

Creedy, J (2003) *Survey reweighting for tax microsimulation modelling*. New Zealand Treasury, Working paper 03/17

Deming, W and Stephan, F (1940) 'On a least square adjustment of sampled frequency tables when the expected marginal totals are known', *Annals of Mathematical Statsictics*, 6,pp427-444.

Deville, J-C., Sarndal, C-E and Sautory, O (1993) 'Generalized Raking Procedures in Survey Sampling', *Journal of the American Statistical Association*, 88(423), pp1013-1020.

Deville, J-C and Sarndal, C-E (1992) 'Calibration estimators in survey sampling', *Journal of the American Statistical Association*, 87, 418, pp376-382.

Edwards, K., Clarke, G., Thomas, J. and Forman, D. (2011) 'Internal and external validation of spatial microsimulation models: small area estimates of adult obesity', *Applied Spatial Analysis*,4, pp281-300.

Evans, A. (2011) 'Uncertainty and Error' in Heppenstall, A., Crooks, A and Batty, M (eds) *Agent-Based Models for Geographical Systems*. London: Springer.

Fielding, A (2004) 'Class and space: social segregation in Japanese cities', *Transactions of the Institute of British Geographers*, 29, pp64-84.

Fienberg, S (1970) 'An Iterative Procedure For Estimation in Contingency Tables', *The Annals of Mathematical Statistics*, 41(3), pp907-917.

Gilbert, N and Terna, P (2000) 'How to build and use agent-based models in social science', *Mind &*

Society 1(1), pp57–72.

Ghosh, M. and Rao, J. (1994) 'Small area estimation: an appraisal', *Statisical Science*, 9(1), pp55-76.

Gomez- Rubio, V., Best, N., Richardson, S., Li, G and Clarke, P (2010) *Bayesian statistics small area estimation*, an NCRM Methodological Review Paper http://eprints.ncrm.ac.uk/1686/

Hanaoka, K (2011) 'Estimating spatial distributions of earnings at the small area level in Japan – A spatial microsimulation approach', *Journal of the City Planning Institute of Japan*, 46(2), pp142-148.

Hanaoka, K., Nakaya, T and Tabuchi, T (2013) 'Small-area estimates of socio-economic inequalities using a spatial microsimulation approach: A case study of Osaka City, Japan', *Annals of the Japan Association of Economic Geographers*, 59(1), pp73-87.

Harding, A., Vidyattama, Y and Tanton, R (2011) 'Demographic change and the needs-based planning of government services: projecting small area populations using spatial microsimulation', *Journal of Population Research*, *28*(2-3), pp203–224.

Harding, A., Vu, Q., Tanton, R and Vidyattama, Y (2009) 'Improving Work Incentives and Incomes for Parents: The National and Geographic Impact of Liberalising the Family Tax Benefit Income Test', *Economic Record*, *85*(s1), S48–S58.

Heady, P., Clarke, P., Ellis, K., Heasman, D., Hennell, S., Longhurst, J and Mitchell, B (2003) *Model-based small area estimation series no. 2: small area estimation project* report. London: Office for National Statistics.

Johnston, R and Pattie, C (1993) 'Entropy-maximizing and the iterative proportional fitting procedure', *Professional Geographer*, 45(3), pp317.

Kavroudakis, D., Ballas, D. and Birkin, M. (2013) 'Using spatial microsimulation to model social and spatial inequalities in educational attainment', *Applied Spatial Analysis*, 6, pp1-23.

Leyk, S., Buttenfield, B and Nagle, N (2013) 'Modeling Ambiguity in Census Microdata Allocations to Improve Demographic Small Area Estimates', *Transactions in GIS*, 17(3), pp406-425.

Lovelace, R and Ballas, D (2013) `Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation', *Computers, Environment and Urban Systems*, 41, pp1-11 (dx.doi.org/10.1016/j.compenvurbsys.2013.03.004).

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, pp99-116.

Macal, C and North, M (2005) 'Tutorial on Agent-Based Modelling and Simulation' in Kuhl, M., Steiger, N., Armstrong, F and Joines, J (eds) *Proceedings of the 2005 Winter Simulation Conference*, pp.–15. Piscataway, NJ: Institute of Electrical And Electronics Engineers.

Macgill, S (1977) 'Theoretical properties of biproportional matrix adjustments', *Environment and Planning A*, 9(6), pp687-701.

Marchetti, S., Tzavidis, N and Pratesi, M (2012) 'Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area average, quantiles and poverty indicators', *Computational Statistics and Data Analysis,* 56(10), pp2889-2902.

Martin, D., Lloyd, C and Shuttleworth, I (2011) 'An evaluation of gridded population models using 2001 Northern Ireland Census data', *Environment and Planning A*, 43(8), pp1965-1980.

Marshall, A (2010) *Small area estimation using ESDS government surveys – An introductory guide* Economic and Social Data Service.

Molina, I and Rao, J. (2010) 'Small area estimation of poverty indicators', *The Canadian Journal of Statistics*, 38(3), pp369-385.

Nakaya, T., Fotheringham, S., Hanaoka, K., Clarke, G., Ballas, D. and Yano, K. (2007) 'Combining microsimulation and spatial interaction models for retail location analysis', *Journal of Geographical Systems,* 9, pp345-369.

Opsomer, J., Claeskens, G., Ranali, M., Kauermann, G. and Breidt, F. (2008) 'Non-parametric small area estimation using penalized spline regression', *Journal of the Royal Statistical Society B*, 70(1), pp265-286.

Pfefferman, D (2002) 'Small Area Estimation - New Developments and Directions', *International Statistics Review*, 70, pp125-143.

Pfeffermann, D (2013) 'New important developments in small area estimation', *Statistical Science*, 28, pp40-68.

Pickering, K., Scholes, S and Bajekal, M (2004) *Synthetic estimation of healthy lifestyles indicators: Stage 2 report.* London: NatCen.

Pratesi, M. and Salvati, N. (2008) 'Small area estimation: the EBLUP estimator based on spatially correlated random area effects', *Statistical Methods and Applications,* 17, pp113-141.

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 and Data Analysis,* 57, pp149-165.

Rahman, A., Harding, A., Tanton, R. and Liu, S. (2010) 'Methodological ssues in spatial microsimulation modelling for small area estimation', *International Journal of Microsimulation*, 3(2), pp3-22.

Rahman, A (2008) *A review of small area estimation problems and methodological developments.* University of Canberra: NATSEM Discussion Paper Issue 66.

Rao, J (2003) *Small Area Estimation.* New York: Wiley.

Rao, J. (1999) 'Some recent advances in model-based small area estimation', *Survey Methodology*, 25(2), pp175-186.

Shakespeare, S. (2013) *Shakespeare review: an independent review of public sector information*. London: Department for Business, Innovation and Skills.

Simpson, L and Tranmer, M (2005) 'Combining sample and census data in small area estimates: iterative proportional fitting with standard software', *The Professional Geographer*, 57(2), pp222-234.

Singh, A and Mohl, C (1996) 'Understanding calibration estimators in survey sampling', *Survey Methodology*, 22, pp107-115.

Smith, D., Clarke, G. and Harland, K. (2009) 'Improving the synthetic data generation process in spatial microsimulation models', *Environment and Planning A,* 41, pp1251-1268.

Tanton, R and Vidyattama, Y (2010) 'Pushing it to the edge: Extending generalised regression as a spatial microsimulation method', *International Journal of Microsimulation*, 3(2), pp23–33.

Tanton, R., Vidyattama, Y., McNamara, J., Vu, Q and Harding, A (2009) 'Old, Single and Poor: Using Microsimulation and Microdata to Analyse Poverty and the Impact of Policy Change among Older Australians', *Economic Papers: A journal of applied economics and policy*, 28(2), pp102–120.

Tanton, R., Vidyattama, Y., Nepal, B and Mcnamara, J (2011) 'Small area estimation using a reweighting algorithm', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(4), pp931-951.

Twigg, L and Moon, G (2002) 'Predicting small area health-related behaviour: a comparison of multilevel synthetic estimation and local survey data', *Social Science & Medicine* 54(6), pp931-937.

Tzavidis, N., Marchetti, S and Chambers, R (2010) 'Robust estimation of small area means and quantiles', *Australian and New Zealand Journal of Statistics*, 52(2), pp167-186.

Vidyattama, Y and Tanton, R (2010) 'Projecting small area statistics with Australian spatial microsimulation model (SpatialMSM)', *Australian Journal of Regional Studies*, 16(1), pp99–126.

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, pp349-366.

Williamson, P (2002) 'The aggregation of small-area synthetic microdata to higher level geographies: an assessment of fit', Working Paper 2002/1, Population Microdata Unit, Dept. of Geography, University of Liverpool.

Williamson, P (2013) 'An evaluation of two synthetic small-area microdata simulation methodologies: synthetic reconstruction and Combinatorial Optimisation' in Tanton, R and Edwards, K (eds) *Spatial microsimulation: a reference guide for users*. London: Springer.

Williamson, P., Birkin, M and Rees, P (1998) 'The estimation of population microdata using data from small area statistics and samples of anonymised records', *Environment and Planning A,* 30, pp785-816.

Wong, D (1992) 'The Reliability of Using the Iterative Proportional Fitting Procedure', The *Professional Geographer*, 44(3), pp340-348.

Zaidi, A., Harding, A. and Williamson, P. (eds.)(2009) *New frontiers in microsimulation modelling: public policy and social welfare*. Farnham: Ashgate.

Appendix A: Overview of the network participants and events

The network ran from May 2012-April 2013 and was deliberately constructed to be methodologically broad so as to incorporate expertise across the broad range of SAE methodological approaches as well as across broader areas of relevance to SAE (e.g. small area covariate data, agent based modelling). Table A1 below lists the network contributors and summarises their main areas of methodological expertise.

Table A1: Network Participants

Overarching Methodological Area	Specific Methodological Area	Network Expertise
Spatial Microsimulation	Iterative Proportional Fitting (IPF)	Dr Adam Whitworth, Dept of Geography, University of Sheffield (PI)
		Dr Ben Anderson, Senior Research Fellow, Engineering and the Environment, University of Southampton
		Dr Dimitris Ballas, Dept of Geography, University of Sheffield
		Dr Kimberley Edwards, School of Clinical Sciences, Univ of Nottingham
		Prof Graham Clarke, Dept of Geography, University of Leeds
		Dr Tomoki Nakaya, Dept of Geography, Ritsumeikan University, Japan
	Combinatorial optimisation (CO)	Dr Dimitris Kavroudakis, University of the Aegean, Lesvos, Greece
		Dr Paul Williamson, School of Environmental Sciences, Univ of Liverpool
		Dr Cathal O'Donoghue, Head of Rural Economy Research Centre, Teagasc, Ireland
		Dr Karyn Morrissey, Dept of Geography, University of Liverpool
	GREGWT	Dr Robert Tanton, NATSEM, University of Canberra, Australia
Statistical estimation approaches	Multi-level models	Prof Graham Moon, Centre for Geographical Health Research, University of Southampton
		Dr Nikos Tzavidis, Social Statistics and Demography, Univ of Southampton
		Joanna Taylor, Senior Research Assistant, Geography and Environment, University of Southampton
		Grant Aitken, PhD student, Geography and Environment, University of Southampton
	M-quantile regression	Dr Nikos Tzavidis, Social Statistics and Demography, Univ of Southampton
	Robust regression	Dr Nikos Tzavidis, Social Statistics and Demography, Univ of Southampton
	Spatial downscaling	Prof Peter Atkinson, Centre for Geographical Health Research, Univ of Southampton
	Bayesian approaches	Prof Nicky Best, Faculty of Medicine, Imperial College London
Dynamic models	Agent based modelling	Dr Alison Heppenstall, School of Geography, University of Leeds
		Dr Nick Malleson, Dept of Geography, University of Leeds
Spatial interaction models	Retail geographies	Prof Graham Clarke, Dept of Geography, University of Leeds
SAE analysis	Comparative analysis	Dr Tomoki Nakaya, Dept of Geography, Ritsumeikan University, Japan
		Dr Dimitris Ballas, Dept of Geography, University of Sheffield
Small area covariate data	Census data	Martin Ralphs, Beyond 2011 project, Office for National Statistics
		Kieran Martin, Beyond 2011 project, Office for National Statistics
		Pamela Dent, Beyond 2011 project, Office for National Statistics
		Philip Clarke, Beyond 2011 project, Office for National Statistics
	Administrative data	Dr Chris Dibben, School of Geography and Geosciences, Univ of St Andrews
		Mr David McLennan, Deputy Director, Social Disadvantage Research Centre, University of Oxford
		Dr Adam Whitworth, Dept of Geography, Univ of Sheffield (PI)
Demographics	Migration and population estimation	Adam Dennett, Centre for Advanced Spatial Analysis, UCL
		Fiona Aitchison, Office for National Statistics Small Area Population Estimation group
Software	R-programming	Dr Dimitris Kavroudakis, University of the Aegean, Lesvos, Greece

Network activities

Over the course of the year the network's activities, all held at the University of Sheffield, focussed around three 'key issues' workshops for network members and a free, open invitation one-day training event to conclude:

Workshop 1 (23-25 May 2012): 'Mapping network expertise and overview of the methodologies'

Workshop 1 focused on introducing network members to each other, mapping the expertise across the network and, in particular, familiarising one another with the details of the range of alternative spatial microsimulation and statistical approaches across the network. This began the process of conducting a synthesis and gap analysis of the state of the art across these main methodological approaches to SAE.

Workshop 2 (16-17 October 2012): 'Challenges and opportunities around small area covariate data'

As noted above, a central issue to the network is the changing landscape around the covariate data required by all SAE methodologies, both in terms of the work of the ONS Beyond 2011 team and potential changes to the census as well as in terms of growing availability of administrative data for research purposes. Workshop 2 focussed on exploring this changing landscape around the covariate data in terms both of the challenges for SAE methodologies but also in terms of the potential roles and opportunities which these changes open up to such methodologies.

Workshop 3 (10 April 2013): 'Linking the methodologies and next steps'

Whilst methodological linkages inevitably flowed through discussion within earlier events the explicit focus of workshop three was on identifying methodological linkages between the spatial microsimulation and statistical approaches. Given that these two broad sets of approaches might best be described as operating largely in parallel at present, it is important to better understand, formalise and make explicit any methodological connections between them with a view to using that learning to improve the methodologies and their estimation.

Training event (26 April 2013): 'Principles and practices in small area estimation methodologies: a one day introductory training course', delivered by Dr Adam Whitworth (Univ of Sheffield) and Dr Kim Edwards (Univ of Nottingham)

The training day took place at the end of the network's duration and was a free, widely advertised and open access event that aimed both to provide an overview of principles and practices in differing approaches to SAE and to provide an opportunity for hands-on supported practical time with pre-prepared estimation examples using both a spatial microsimulation (IPF) and a statistical approach. The day took place within a dedicated IT teaching lab at the University of Sheffield.