More for Less?

Using statistical modelling to combine existing data sources to produce sounder, more detailed, and less expensive Official Statistics

Prof Stephen Haslett\(^1\), Dr Geoff Jones\(^1\), Dr Alasdair Noble\(^1\) and Dr Dimitris Ballas\(^2\)

\(^1\)Institute of Fundamental Sciences – Statistics, Massey University, New Zealand

\(^2\)Department of Geography, University of Sheffield, England

Corresponding author: Prof Stephen Haslett s.j.haslett@massey.ac.nz

This report was commissioned by Official Statistics Research, through Statistics New Zealand. The opinions, findings, recommendations and conclusions expressed in this report are those of the author(s), do not necessarily represent Statistics New Zealand and should not be reported as those of Statistics New Zealand. The department takes no responsibility for any omissions or errors in the information contained here.
Abstract

The principal aim of this research has been to establish when and how Official Statistics data sources, particularly surveys and censuses or surveys and administrative sources, can be combined using statistical models based on mass imputation, spatial microsimulation, and small area and domain estimation, to produce cost-effective, accurate, fine-level statistics. The research considers, summarises and further develops a range of statistical methods, and considers their application in principle to set of case studies in sociology, economics, and business from a variety of New Zealand government departments and ministries. Guidelines, based on this research, for use of these three core techniques are also provided.

Keywords

Mass imputation, small area estimation, spatial microsimulation.

Acknowledgements

The authors would like to thank all those who have provided input into this project, particularly the Advisory Committee (Helen Stott - Ministry of Social Development, Robert Templeton - Ministry of Heath, Dr David Turner - Ministry of Justice, Mike Doherty - Statistics New Zealand), people in government departments involved in existing projects (Frances Krsinich - Statistics New Zealand, Dr Chungui Qiao - Ministry of Social Development, Dr David Turner and Sally Faisandier - Ministry of Justice), A/Prof David O’Sullivan - Geography Department, Auckland University, Dr Walter Davis - Statistics New Zealand, Kim Cullen, Sophia Caldwell and Rebecca Stoks - Official Statistics Research, and the reviewer. While any errors and omissions remain the responsibility of the authors, without their inputs this research would not have been possible.
CONTENTS

1. Introduction
   1.1 Background
   1.2 What are spatial microsimulation, mass imputation and small area estimation?
   1.3 New Zealand - government projects: outline of case studies
      1.3.1 IBULDD imputation
      1.3.2 Small area estimation in outcome evaluation of Youth Transition Services
      1.3.3 Imputing Victim Forms from the New Zealand Crime and Safety Survey

2. Methodology
   2.1 Small area estimation – ELL / World Bank method
      2.1.1 Small area estimation
      2.1.2 Clustering
      2.1.3 The ELL method
      2.1.4 Bootstrapping
      2.1.5 Interpretation of standard errors
   2.2 Spatial microsimulation: Synthetic reconstruction, spatial microsimulation and combinatorial optimisation methods
      2.2.1 Aspatial microsimulation
      2.2.2 Spatial microsimulation
      2.2.3 Synthetic probabilistic reconstruction
      2.2.4 Probabilistic reweighting approaches to spatial microsimulation modelling
      2.2.5 Deterministic microsimulation approaches to spatial microsimulation modelling
      2.2.6 Microsimulation approaches to small area population projection
   2.3 Mass imputation
   2.4 Associated techniques

3. A unifying theoretical framework for ELL-type small-area estimation, spatial microsimulation, and mass imputation

4. Simulation study
   4.1 Investigation of small area estimation – ELL method
   4.2 Comparing spatial microsimulation and ELL-type small area estimation
   4.3 Effect of Model Uncertainty
   4.4 Updating via projections
   4.5 Mass Imputation

5. Conclusions
   5.1 Simulations
   5.2 New Zealand - government projects: general comments on case studies
      5.2.1 IBULDD imputation
      5.2.2 Small area estimation in outcome evaluation of Youth Transition Services
5.2.3 Imputing Victim Forms from the New Zealand Crime and Safety Survey:

5.3 General conclusions: links and comparison of small area estimation (ELL), spatial microsimulation, and mass imputation

References

Appendix 1: Guidelines for using mass imputation, spatial microsimulation, and ELL-type small area estimation

Appendix 2: Guidelines for designing surveys for mass imputation, spatial microsimulation, and ELL-type small area estimation

Tables

Table 1: Microsimulation procedure for the allocation of economic activity status (after the similar example of tenure allocation procedure given by Clarke, 1996: 3)

Table 2: The BHPS microdata format

Table 3: The BHPS microdata variable descriptions

Table 4: An example of small area information

Table 5: A hypothetical microdata set (original weights: table w)

Table 6: Hypothetical small area data tabulation (table s)

Table 7: The hypothetical microdata set, cross-tabulated by age and sex. (table m)

Table 8: Reweighting the hypothetical microdata set in order to fit. Table 6

Table 9: A simple example of the microsimulation procedure for mortality (after Ballas, Clarke and Wiemers, 2005).

Table 10a: Coverage of nominal 95% confidence intervals

Table 10b: Average length of nominal 95% confidence intervals

Table 10c: Bias = Average of (Estimate – True Value)

Table 10d: Standard error (Standard Deviation of Estimate)

Table 11a: Design of census and sampling

Table 11b: Average mean square error (100 simulations)

Table 11c: Performance of nominal 95% confidence intervals (1000 simulations)

Table 12a: Distribution of chosen model in 100 simulations (from design B1)
More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics

Table 12b: Comparison of performance of estimators using true and chosen models (1000 simulations)

Table 13: Effect on estimation of accuracy of additional noise added to the variable of interest, using census/sample design B2

Table 14a: Proportions of households with a given number of changes in the X variables resulting from added noise

Table 14b: Performance of mass imputation estimator for fixed census (designs as in Table 11a).

Figures

Figure 1: Coverage versus Bias (difference between true value and estimate) for the two censuses.

Figure 2: Graph showing variability in the census values (Y.c) the sample (Y.s) and the two estimates by iterative proportional fitting (Y.ipf) and small area estimation using a linear mixed effects model (Y.lme). These results are for one realization of design A1.

Figure 3: Increase in MSE resulting from errors in updating or projection for different values of the Noise Variance Factor

Figure 4: Graphs showing the Area-level census means (Y.c), and the means (Ybar.i) and standard deviations (Isd) of the corresponding multiple mass imputation estimates (from design A1).
1. Introduction

1.1 Background

Across Official Statistics internationally, there is wide discussion and increased usage of techniques which combine survey with census or administrative data to reduce overall costs and/or provide more detailed, finer-level statistics. There is also increasing awareness that emphasis in Official Statistics needs to continue to move toward supplementing data collection with better data utilisation. However the underlying statistical theory to do this well requires further research, especially for modelling methods that combine data as anonymously as possible to limit confidentiality concerns.

Mass imputation, spatial microsimulation, and small area estimation (including small domain estimation) are three statistical methods for extending usage of survey and census data. Although their research literature is essentially separate, the three techniques have underlying similarities. The nature and extent of those similarities, along with the important differences, and guidelines for their proper use, are linked in this paper to a set of case studies based on current projects within New Zealand government departments and ministries. The central aim of the research has been to establish when and how Official Statistics data sources, particularly surveys and censuses or surveys and administrative information, can be combined using statistical models to produce cost-effective, accurate, fine-level statistics and to enhance social and/or economic data collection and outputs.

The principal aim of this research project then is to explore how best to combine survey and census data, or survey and administrative data, using sound statistical models to produce finer-level statistics for variables collected only by sample survey, without formally linking individual records and raising such confidentiality issues. Essentially, the methods work by imputing or predicting variables of interest that are not collected or are partially missing in the larger dataset. Prediction is usually via models that use variables common to both datasets. The fitted model can then provide multiple predictions for all missing census observations, which when combined can give not only area or subpopulation estimates but also with estimates of accuracy conditional on the model being correct. The statistical techniques considered include mass imputation, spatial microsimulation, and certain types of small area and small domain methods which, while similar in intent, can differ in the form of the underlying statistical model used to connect data sources.

The research also links to enhanced geospatial capability across the Official Statistics System (OSS) since it develops and discusses methodologies useful for unit-level, area-based information that can be amalgamated by area to produce means or totals that can be mapped. For confidentiality and accuracy reasons, the very fine-level data generated from statistical models by combining survey and census (or administrative) data is not suitable for publication itself but, in principle, after amalgamation, it is publishable in tables or suitable for mapping, and can include standard error estimates. Simple amalgamations of finer-level data to different area-boundaries, once well-defined, can minimise the impact of administrative boundary changes. Extension to official statistics that involve time series is possible, where the techniques are applied under suitable conditions at each point in time, or data is projected into the future and suitable models fitted to the projections. Such extensions allow changes to be monitored, using maps for example, to provide understanding of social and economic change, including dynamics and transitions.
The research has allowed development of guidelines on when, in principle, methods that combine data from different sources can and cannot assist state sector agencies that collect Official Statistics both to reduce respondent load and cost, and to improve accuracy of existing surveys by use of supplementary data sources. At a more detailed level, statistical modelling such as small area estimation has the potential to lower respondent load in particular subpopulations, such as Māori. (Haslett, Jones and Enright, 2008). Some general comment is made on how surveys should best be designed in future to integrate with census and administrative data, a topic with consequences both for survey data collection and respondent load.

Details of research methods are given below in Sections 3 and 4. The aim of the research is to compare and contrast the available statistical techniques for producing census-level datasets by modelling, to assess when and why such methods work well, and when they cannot be recommended. Conclusions are linked to existing Official Statistics projects in a range of government departments and ministries. These projects provide case studies that connect the theoretical research with particular situations encountered in the New Zealand Official Statistics System.

The current government projects used as case studies are:

- **IBULDD imputation** – Improved Business Understanding via Longitudinal Database Development: Statistics New Zealand - Frances Krsinich
- **Small Area Estimation in the Outcome Evaluation of YTS (Youth Transition Services)**: Ministry of Social Development - Dr Chungui Qiao
- **Imputation of Victim Forms from the New Zealand Crime and Safety Survey (NZCASS)**: Ministry of Justice - Dr David Turner.

Other government projects were also considered, but not included in this report in further detail for a variety of reasons (usually related to their being at too-preliminary a stage for methodology to have been fully specified). These were: tax benefit simulation for childcare to assess total cost for MSD; A Simulated System for Estimating Taxation (ASSET) which is a spatial microsimulation method without standard errors used initially by Statistics New Zealand in the 1980s at the time Goods and Service Tax (GST) was introduced; a study to follow complex transitions of religious affiliation in census data at Statistics New Zealand; and a spatial microsimulation approach using multilevel models, demographic data, and interpolation by geographical smoothing for relative risk of smoking by the Ministry of Health.

### 1.2 What are spatial microsimulation, mass imputation and small area estimation?

**Spatial microsimulation** is a relatively new technique, based on an older one, for generating detailed synthetic datasets that describe household characteristics at a local level, that has been developed primarily by geographers. The technique combines aggregate census data and sample household survey datasets to get area based information. The aim is to get the best out of the combination of data sources, since the census data contains more households than the samples, but fewer variables of interest. The method produces fine-level information on neighbourhoods and household composition at local level. The analysis of spatial variation in such simulated data, it is claimed, can be used to study the impact of social policy initiatives, for example in taxation, justice, education, and health. In essence then, the method includes statistical techniques that may add census averages to household data where those averages are taken over neighbourhoods, some types
of sample reweighting techniques, and prediction based on what is usually an implicit model.

Mass imputation is a statistical technique for producing detailed synthetic datasets that has been studied, used, and essentially abandoned for survey data (in favour of sample reweighting techniques) by major central government statistical agencies internationally, particularly in Canada and the Netherlands. In mass imputation, data is imputed at individual, or household (or in business studies, at establishment or enterprise) level. The data is then aggregated to the required level for publication. However, many of the historical imputation methods were specified primarily for ease of computation and suitability for highly aggregated statistics. The open questions are whether better statistical models are possible now given improvements in computing power, and if so whether such mass imputation models would improve estimates at finer levels of aggregation.

Small area estimation, although often applied to spatial statistics, in its widest sense also includes techniques for small domain estimation for population subgroups. There are a very broad range of statistical methods used to produce finer level estimates than is possible using traditional survey estimation alone. The techniques have developed from simple models that reweight sample survey data to more sophisticated linear and non-linear mixed models, some of which include census as well as survey data. Some small area estimation techniques have strong similarities to spatial microsimulation and mass imputation. Among these, one of the most often used is the technique developed, used and supported by the World Bank for small area estimation of poverty (see for example, Zhao, 2006). In the World Bank technique, which has now been used in over 50 countries, household-level data for expenditure (which is not generally collected in the census) are simulated (or multiply imputed) for the entire census under a regression model fitted to the survey data. As for spatial microsimulation, census means at neighbourhood or cluster level are also included in the model. Although important technical questions remain, the World Bank method, which was initially developed to estimate income and expenditure, can in principle be equally well applied to a much wider range of variables (e.g. Jones, Haslett and Parajuli, 2006). Indeed, there are obvious parallels between the ELL method and some of the early papers linked to spatial microsimulation (e.g. Birkin and Clarke, 1989). Some of these papers, e.g. Habib (1986), Krupp (1986), Paas (1986) and other related papers in Orcutt, Mertz and Quinke (1986), even have a finance or economic focus. Curiously however, we have not been able to find any common peer-reviewed papers referenced in both the ELL-related and the spatial microsimulation literature. Other small area techniques that combine survey and census data exist (e.g. Haslett, Noble and Zabala, 2008) but these are outside the scope of the current study.

The three statistical techniques are clearly linked, but the extent of overlap, and when exactly caution is required in their use remain open questions. This problem is of particular importance and possible concern to the New Zealand OSS because such techniques are now computationally feasible even on personal computers, given suitable software, and a number of such projects (such as those detailed in the case studies below) are already underway in New Zealand government departments and ministries.

The three methods are connected because all of them “complete” databases consisting of many individual records (of people or businesses) and many variables, for which some data is missing, by substituting an estimate or prediction for the missing values, i.e. by filling in the ‘holes’. The methods differ to some extent in how they do this substitution, which turns a dataset with holes into the type of complete dataset that is found (usually on a smaller scale) in spreadsheets with no missing cells. The original database may be a census or survey. The gaps or missing data may be missing by intention (as when a variable such as expenditure is not collected at all for a census) or at random (as for the non-selected part of the frame in a
survey, or in certain situations when responses to particular items for some respondents are missing), or in some possibly non-random fashion (as for example, where there are non-response complications). Generally (in contrast to the more usual imputation methods for surveys), all three methods are applied to generate a considerable proportion of the data, at least for some variables. The aim is always the same: to fill up all the holes or gaps in the data. The central questions are how good the datasets created this way are, and for what purposes they can or cannot be used.

All three methods work by imputing or predicting variables of interest that are not collected or are partially missing in census or administrative data by using, explicitly or implicitly, statistical models based on a subset of observations that do not have any missing values (which, even for census or administrative data, often come from an associated survey). Although similar imputation models can be used for surveys that have some missing data, the current research focuses more on the larger scale problem of ‘completing’ censuses and administrative records, and on surveys where the missingness is extensive. In all cases, surveys (rather than some records from the census or administrative data) are usually used to estimate the underlying statistical model, which is then used to generate single or multiple predictions of a variable of interest for all the missing census or survey observations or administrative records. When these predictions for a particular variable are summed, along with the values that were collected from the other respondents, they give area or subpopulation estimates, for example of proportions, means and totals, possibly with estimates of their accuracy. Mass imputation, spatial microsimulation, and certain types of small area / domain methods are similar in intent, but may use different types of statistical model. For spatial microsimulation and mass imputation, but not for small area estimation, the model is usually implicit rather than fully specified.

A central part of our research is a series of simulation studies, which are based on known models and use data structures reflecting those found in the set of case studies within government departments and ministries. The aim of the simulations has been to construct a range of populations based on the case studies, delete some or even a large proportion of that data for some variables, and see how well the original data can be reconstructed using each of the three methods. Of course, the reconstructed data is not perfect, so even if a method is ‘good’ the other question is exactly what the data are useful for, e.g. estimation of simple averages and totals, or more complicated statistics such as standard errors, correlation coefficients, regression relationships, or variance components. The simulations consider all three methods, a range of percentages of missing data, different patterns of missingness, and several populations some of which are comparatively easy and others more difficult to predict. These simulation runs involve a considerable amount of design, model fitting, compute time and testing.

The initial part of the case studies in New Zealand government departments and ministries has considered how these techniques are being or have been applied, and the type and characteristics of survey and census (or survey and administrative) data for which they are or have been being used. Time and other resource constraints mean that it has not been feasible to be directly involved in the implementation of such existing applied projects within government agencies as an intrinsic part of the proposed research; multiple applications to existing data are beyond the limitations imposed by a one-year OSResearch project. For this reason, we instead collected and used detailed information from a range of government projects as the basis of the simulation studies.

One benefit of this methodology is that the “real” answer is known in a simulation, so core questions can be answered about the effects of data distribution, the nature of the missing data, sample size of the survey data set, model fit, variance components where relevant, and the relative size of survey and census. General conclusions have also been drawn, and a set of recommendations for government
produced. Although there has not been involvement in the day-to-day aspects of these government projects, the results of our research are referenced back as individual case study summaries. Taken as a collection, the case studies are intended to connect the results of the theoretical research with the particular situations encountered in the New Zealand Official Statistics System.

Mass imputation, spatial micro-simulation and small area / domain estimation are also connected to record linkage methods used to merge distinct data sources into a single dataset in the absence of unique identifiers, generalized regression (GREG) estimation methods, weighting and reweighting for sample surveys, fractional imputation, and inverse sampling techniques. While these additional techniques are discussed briefly in Section 2.4, the focus remains on the three core methods.

In summary, the aim of the research is to compare and contrast three statistical techniques, (mass imputation, spatial micro-simulation, and small area / domain estimation using survey and census or administrative data) for producing census-level datasets by modelling, to assess when and why such methods work well, and when they cannot be recommended. The central question that our proposed research addresses is how these three techniques are linked, and in what circumstances they do and do not produce reliable statistics at local level.

1.3 New Zealand - government projects: outline of case studies

Official Statistics has long recognised the potential for leveraging administrative data to achieve more accurate statistics, because the use of existing Official Statistics data collections and sound statistical modelling has the potential to save money, provide more accurate statistics at finer levels, and reduce future data collection costs and respondent burden. More recently, the focus has expanded to include the possibilities of combining administrative and/or census data with survey data via statistical modelling, e.g. by using mass imputation, spatial microsimulation and certain types of small area and small domain estimation. These methodological developments have varying levels of theoretical backing, but are nevertheless the first and last are already being used within New Zealand government departments and ministries on what are now extensive and wide-ranging electronic databases containing social, economic, health, epidemiological and other information. A course on the second, spatial microsimulation, was run in Auckland and Wellington in February 2008 by Dr Dimitris Ballas from University of Sheffield through the Social Policy and Evaluation Research Committee (SPEAR – www.spear.govt.nz). Spatial microsimulation is being used extensively in the UK and may soon be used in New Zealand, as it provides a link to approaches geographers take to smoothing spatial data in their Geographical Information Systems (GIS) and outputs.

1.3.1 IBULDD imputation

IBULDD is the Improved Business Understanding via Longitudinal Database Development project at Statistics New Zealand. The database that results from this project is the prototype Longitudinal Business Database (LBD). The data imputed to date is for the period 2000-2005, and for each business includes variables such as region, industrial classification, number of employees, total salaries and wages, sales and services, and purchases. A number of these variables are ‘complete’, i.e. available for all businesses. All economically significant New Zealand businesses are included, approximately 800,000 in all over this period including births (plus rebirths) and deaths. Approximately 60% have zero sales. Data sources include Inland Revenue Department data, the Annual Enterprise Survey, the Business Operations Survey, and the Linked Employer-Employee Data (LEED), which as Statistics New
Zealand indicates “provides information on New Zealanders’ interaction with the labour market and their sources of income”.

Imputation for missing values on some variables would be extensive if the survey data were also imputed, in order to create a rectangular data set with no missing values on any variable for any case. In practice, imputation has to date been limited to two variables in the Inland Revenue Department (IRD) data, one of which is missing in 30% of cases and the other in 8%. Missingsness in the survey data is much more extensive. The aim of the imputation is to use imputation classes, which are based on cross-tabulations of categorical variables such as region and industry class, and number of employees, which are available from IRD, to impute based on nearest neighbour matching for continuous variables using Banff (Kozak, 2005). Phase 1 for IBULDD used hot-deck imputation for those businesses paying zero Goods and Services Tax (GST). Size of business was not used as part of the imputation scheme although, for the business surveys used, industry class and size class were important design variables. Many of the variables of interest are highly skewed. The intention was to provide a single ‘complete’ dataset for analytical purposes, for the two key IRD variables, not simply area or industry based averages.

1.3.2 Small area estimation in outcome evaluation of Youth Transition Services

The Youth Transitions Services (YTS) initiative was implemented by the Ministry of Social Development (MSD) to achieve the Government’s goal of having all 15-19 year-old youth in work, education, training or other activities that contribute to their long-term economic independence and wellbeing. YTS was designed as an early intervention strategy to support young people at risk of disengagement in their transition from school to economic independence in their employment pathway.

Eligibility for unemployment benefit generally requires recipients to be 18 years of age. If YTS is effective, those young people who received the YTS support at the age of 15, 16, and 17 will become economically independent so that they do not have to take up unemployment benefit when they turn 18. Hence, MSD used as outcome measure the proportion of youth in receipt of working-age benefits at the age of 18 in the outcomes evaluation of YTS, for comparison of the intervention areas with those similar non-intervention areas.

Results show that the proportion of youth aged 18 on working age benefits declined for both intervention and non-intervention areas over the past seven years. Since the beginning of 2006, when YTS began to engage more at-risk youth as service delivery capacity increased, the proportion of youth receiving a benefit continued to decline in intervention areas, whereas the proportion appeared to flatten in non-intervention areas. The difference between intervention and non-intervention areas appears to become slightly smaller in the period towards the end of 2006, when intervention areas became fully operational, and more youth were receiving YTS assistance and support.

Due to time constraints it was not feasible to determine whether the gap between intervention and non-intervention areas will finally close beyond 2007. The failure to confirm any YTS impact using this approach could also be due to the small number of at-risk youth who participated in YTS, relatively representing a small proportion of the total youth population. The data is not missing in the usual statistical sense, since it is essentially complete at a recent date, but possible future scenarios are of interest and consequently statistical modelling and prediction or projection are required. The key question to answer is whether the proportion of youth aged 18 on working age benefits differ significantly when the time series is sufficiently long and the number of at risk youth who receive YTS support has become sufficiently large.

Another outcome measure would be the official unemployment rate for youth aged 18 and 19, to demonstrate the impact of YTS in the intervention areas over non-intervention areas. Such an approach would require unemployment data for the
18 and 19 year olds at territorial local authority level, to be pooled for estimates for both intervention areas over non-intervention areas. Since Statistics NZ does publish unemployment data down to such fine levels, small area estimation will have to be used for further in-depth evaluation of the programme.

1.3.3 Imputing Victim Forms from the New Zealand Crime and Safety Survey:

In the New Zealand Crime and Safety Survey 2006 (NZCASS) respondents were asked to provide answers to core questions and to provide information on any victimisations. The design means that maximum number of incidents for each respondent for which victim forms are to be completed is three. However, some respondents are heavily victimised, so that 67% of incidents reported in NZCASS have no victim forms.

Most of the data collected in victim forms was analysed using sample reweighting techniques, so in itself this missingness does not require imputation. However, direct calculation of incidence and prevalence rates needs some information about all incidents (i.e., those experienced by all survey participants), particularly three items on each victim form: whether the incident occurred during the year 2005; whether the incident was an offence within the scope of the survey (termed “being relevant”); and the detailed offence code under which the incident falls. Because this data is missing for the 67% of incidents that do not have victim forms, imputation is required to replace this missing data. Imputation affects the sampling error of the results, and the NZCASS technical report (Reilly and Sullivan, 2008, p. 32) recognises that for some imputation methods “it is hard to figure out how much”. Multiple imputation (Rubin, 1987) with ten imputations has been used for the 2006 NZCASS to quantify this effect, via Lumley’s (2006) ‘mitools’ package.

If the number of incidents at any screener question was missing, it was assumed that the respondent reported being a victim of just one incident. Reilly and Sullivan (2008) suggest this may underestimate the true level of victimisation, but point out that other common imputation methods (e.g. random hot-deck) would not perform any better. They note that “primary reasons for imputing a value of 1 were that this approach was used in the 2001 survey, and that no clearly superior method was identified”.

For date imputation where required, in general, for each of the ten imputations per incident, the date at which the incident occurred was imputed randomly assuming each day between 1 January 2005 and date of interview was equally likely. Reilly and Sullivan (2008) note that assuming uniform spread does not account for recall bias.

Different types of offence have different percentages of completed victim forms for which the victimisation was in scope and recorded as in scope. This percentage is called the relevance rate and it varies markedly from 38% to 90% depending on the type of offence. There are also different proportions of missing data for different offences ranging from 34% to 88%. An imputation model was chosen by stepwise selection, leading to a model with the screener question, age, ethnicity, household size, tenure/landlord, New Zealand Socio-Economic Index (NZSEI), and the New Zealand Index of Deprivation score as predictors. Details of the model are given in Reilly and Sullivan (2008, Table 9.2). This model was used to multiply impute relevance status from the main screener questions for incidents without victim forms, with a slight variation for incidents (usually those of a more sensitive nature) for which the questionnaire required self-completion by the respondent.

Offence codes were imputed using “hot deck” imputation (the approximate Bayesian bootstrap), with imputation classes defined by the source screener question. As Reilly and Sullivan (2008) note, this reproduces the distribution of offence codes from each screener, on average. Mode imputation, used in the 1996 and 2001 surveys and which simply selects the mode for each variable, would have
depressed estimated rates for offences without dedicated screener questions and overstated rates for offences with their own screener question.

There was also imputation to adjust for duplicated incidents for imputed victim forms, based on the number of such duplications in the incidents for which there were victim forms. After imputation, a cut-off of thirty valid offences was applied to ‘improve the reliability’ of the estimated rates.

The effect of imputation on the variance estimates for victimisation rates was carried out assuming the imputation and analysis models were congenial (Meng, 1994). As Reilly and Sullivan (2008, p. 39) note, “Model misspecification can cause multiple imputation to produce biased variance estimates”. Putting aside this issue, they also note that the imputation effect alone usually scaled up variances by a factor between 1.2 and 1.8 (the lower and upper quartiles), and that sexual offences (which were self-completed) have large imputation effects ranging in the range two to ten.
2. Methodology

2.1 Small area estimation – ELL / World Bank method

In this section, we present a brief overview of small area estimation, and the ELL (Elbers, Lanjouw and Lanjouw, 2003) method which is supported by the World Bank and has been used primarily for small area poverty estimation in developing countries. ELL has strong links to the economic rather than the statistical literature (e.g. Bramley, 1991a, 1991b, 1992; Bramley and Smart, 1995, 1996; Bramley and Lancaster, 1998) and was followed by various related publications (e.g. Elbers, Lanjouw and Lanjouw, 2000; Elbers, Lanjouw, Lanjouw and Leite, 2001; Elbers, Lanjouw and Lanjouw, 2002) and has been followed by others (e.g. Elbers, Lanjouw and Leite, 2008).

2.1.1 Small area estimation

Small area estimation refers to a collection of statistical techniques designed for improving sample survey estimates through the use of auxiliary information (Ghosh and Rao, 1994; Rao, 1999, 2003). We begin with a target variable, denoted \( Y \), for which we require estimates over a range of small subpopulations, usually corresponding to small geographical areas. Direct estimates of \( Y \) for each subpopulation are available from sample survey data, in which \( Y \) is measured directly on the sampled units (households or eligible children). Because the sample sizes within the subpopulations will typically be very small, these direct estimates will have large standard errors and hence not be reliable. Some subpopulations may not be sampled at all in the survey. Auxiliary information, denoted \( X \), can be used under some circumstances to improve the estimates, giving lower standard errors for sampled areas and estimates even for unsampled areas.

In the situations examined in this report, \( X \) represents additional variables that have been measured for the whole population, either by a census or via a GIS database. A relationship between \( Y \) and \( X \) of the form

\[
Y = X\beta + u
\]

can be estimated using the survey data, for which both the target variable and the auxiliary variables are available. Here \( \beta \) represents the estimated regression coefficients giving the effect of the \( X \) variables on \( Y \), and \( u \) is a random error term representing that part of \( Y \) that cannot be explained using the auxiliary information. If we assume that this relationship holds in the population as a whole, we can use it to predict \( Y \) for those units for which we have measured \( X \) but not \( Y \). Small area estimates based on these predicted \( Y \) values will often have smaller standard errors than the direct estimates, even allowing for the uncertainty in the predicted values, because they are based on much larger samples. Thus the idea is to “borrow strength” from the much more detailed coverage of the census data to supplement the direct measurements of the survey.

2.1.2 Clustering

The units on which measurements have been made are often not independent, but are grouped naturally into clusters of similar units. Households tend to cluster...
together into mesh blocks or other small geographic or administrative units, which are themselves relatively homogenous. Put simply, households that are close together tend to be more similar than households far apart. When such structure exists in the population, the regression model above can be more explicitly written as

\[ Y_{ij} = X_{ij}\beta + c_i + e_{ij} \]  \hspace{1cm} (3.1)

where \( Y_{ij} \) represents the measurement on the \( j \)th unit in the \( i \)th cluster, \( c \) the error term held in common by the \( i \)th cluster, and \( e_{ij} \) the household-level error within the cluster. The relative importance of the two sources of error can be measured by their respective variances \( \sigma_c^2 \) and \( \sigma_e^2 \). Ghosh and Rao (1994) give an overview of how to obtain small area estimates, together with standard errors, for this model. Where individual-level rather than household-level data is used, an additional error term at within household is added. In the general explanation given below we focus on equation (3.1) in order to establish general principles useful for distinguishing the characteristics of variation at 'higher' and 'lower' levels. When there are three error terms rather than two, the three form a sequence in which the cluster remains the highest level of aggregation, household takes an intermediate status, and individual level variation is at the finest level. It is also possible, and indeed advisable to add an error term at small area level, to check whether the variables included in the model are sufficient to rule out the requirement for small area level random effects.

We note that the auxiliary variables \( X_{ij} \) may be useful primarily in explaining the cluster-level variation, or the household-level variation. The more variation that is explained at a particular level, the smaller the respective error variance, \( \sigma_c^2 \) or \( \sigma_e^2 \). The estimate for a particular small area will typically be the average of the predicted \( Y \)s in that area. Because the standard error of a mean gets smaller as the sample size gets bigger, the contribution to the overall standard error of the variation at each level, household and cluster, depends on the sample size at that level. The number of households in a small area will typically be much larger than the number of clusters, so to get small standard errors it is of particular importance that, at the higher level, the unexplained cluster-level variance \( \sigma_c^2 \) should be small. (A parallel comment applies to any small-area level variance in comparison with the cluster level variance.) Two important diagnostics of the model-fitting stage, in which the relationship between \( Y \) and \( X \) is estimated for the survey data, are the \( R^2 \) measuring how much of the variability in \( Y \) is explained by \( X \), and the ratio \( \sigma_c^2 / (\sigma_c^2 + \sigma_e^2) \) measuring how much of the unexplained variation is at the cluster level. Note that although \( \sigma_c^2 \) and \( \sigma_e^2 \) are parameters they are different for different models with different regressors. GIS data and cluster-level means can be particularly useful in lowering this ratio.

Another important aspect of clustering is its effect on the estimation of the model. The survey data used for this estimation cannot usually be regarded as a simple random sample, because they have been obtained from a complex survey design which although it is random, nevertheless involves stratification and cluster sampling. To account properly for the complexity of the survey design requires the use of specialized statistical routines (Skinner et al., 1989; Chambers and Skinner, 2003; Lehtonen and Pakhinen, 2004; Longford, 2005) in order to get a consistent estimate for the regression coefficient vector (i.e. \( \hat{\beta} \)) and its variance (\( V_{\hat{\beta}} \)).

### 2.1.3 The ELL method
The ELL methodology was initially designed specifically for the small area estimation of poverty measures based on per capita household expenditure. Here the target variable \( Y \) is log-transformed expenditure, the logarithm being used to make more symmetrical the highly right-skewed distribution of untransformed expenditure. It is assumed that measurements on \( Y \) are available from a survey.

The first step is to identify a set of auxiliary variables \( X \) that are in the survey and are also available for the whole population. It is important that these should be defined and measured in a consistent way in both data sources. The model (3.1) is then estimated for the survey data, by incorporating aspects of the survey design for example through use of the “expansion factors” or inverse sampling probabilities. The residuals \( \hat{u}_y \) from this analysis are used to define cluster-level residuals \( \hat{c}_i = \hat{u}_y - \hat{c}_i \), the dot denoting averaging over \( j \), and household-level residuals \( \hat{e}_{ij} = \hat{u}_{ij} - \hat{e}_{ij} \).

It is assumed that the cluster-level effects \( c_i \) all come from the same distribution, but that the household-level effects \( e_{ij} \) may be heteroscedastic. This is modelled by allowing the variance \( \sigma_{e}^2 \) to depend on a subset \( Z \) of the auxiliary variables:

\[
g(\sigma_{e}^2) = Z\alpha + r
\]

where \( g(.) \) is an appropriately chosen link function, \( \alpha \) represents the effect of \( Z \) on the variance and \( r \) is a random error term. Fujii (2006) uses a version of the more general model of ELL involving a logistic-type link function, fitted using the squared household-level residuals. Fujii’s model is:

\[
\ln\left( \frac{\hat{e}_{ij}^2}{A - \hat{e}_{ij}^2} \right) = Z\alpha + r_{ij}
\]

From this model the fitted variances \( \hat{\sigma}_{e,ij}^2 \) can be calculated and used to produce standardized household-level residuals \( \hat{e}_{ij}^* = \hat{e}_{ij} / \hat{\sigma}_{e,ij} \). These can then be mean-corrected or mean-centred to sum to zero, either across the whole survey data set or separately within each cluster. In many practical applications of ELL to economic poverty measures the heteroscedasticity is small, and using equation (3.2) is unnecessary.

In standard applications of small area estimation, the estimated model (3.1) is applied to the known \( X \) values in the population to produce predicted \( Y \) values, which are then averaged over each small area to produce a point estimate, the standard error of which is inferred from appropriate asymptotic theory. In the case of poverty mapping, interest is not always directly in \( Y \) but in several non-linear functions of \( Y \), such as poverty incidence, gap and severity. This is one of the advantages of the ELL method: because it predicts at household or individual level, non-linear functions can be calculated at this level and aggregated. Such non-linear transformations are not possible for models fitted to aggregate data. The ELL method obtains unbiased estimates and standard errors for linear and non-linear functions by using a bootstrap procedure as described below.

2.1.4 Bootstrapping

Bootstrapping is the name given to a set of statistical procedures that use computer-generated random numbers to simulate the distribution of an estimator (Efron and
Tibshirani, 1993). In the case of poverty mapping, we construct not just one predicted value

$$\hat{Y}_{ij} = X_{ij} \hat{\beta}$$

(where $\hat{\beta}$ represents the estimated coefficients from fitting the model) but a large number of alternative predicted values

$$Y_{ij}^b = X_{ij}\beta^b + c_i^b + e_{ij}^b , \quad b = 1, \ldots, B$$

in such a way as to take account of their variability. The statistical analysis of the chosen model for $Y$ yields information on how to appropriately insert variability into the calculation of the predicted values. We know for example that $\hat{\beta}$ is an unbiased estimator of $\beta$ with variance $V_{\beta}$ so we draw each $\beta^b$ independently from a multivariate normal distribution with mean $\hat{\beta}$ and variance matrix $V_{\beta}$. The cluster-level effects $h_i^b$ are taken from the empirical distribution of $h_i$, i.e. drawn randomly with replacement from the set of cluster-level residuals $\hat{h}_i$, since the appropriate cluster level residual is known only for the clusters in the sample not all the clusters in the census. To take account of any unequal variances (heteroscedasticity) in the household-level residuals, we can first draw $\beta^b$ from a multivariate normal distribution with mean $\hat{\beta}$ and variance matrix $V_{\beta}$ combine it with $Z_{ij}$ to give a predicted variance and use this to adjust the household-level effect

$$e_{ij}^b = e_{ij}^* \times \sigma_{e_{ij}}^b$$

where $e_{ij}^*$ represents a random draw from the empirical distribution of $e_{ij}$, either for the whole data set or just within the cluster chosen for $h_i$ (consistent with the mean-centring of Section 2.1.3).

Each complete set of bootstrap values $Y_{ij}^b$, for a fixed value of $b$, will yield a set of small area estimates. In the case of poverty estimates we exponentiate each $Y_{ij}$ to give predicted expenditure $E_{ij} = \exp(Y_{ij})$, then apply an appropriate summation. This is not equivalent to totalling the $Y_{ij}$ in each small area and exponentiating, which is one reason that fitting the model at household (or individual level in the case of a three level model), rather than to aggregated data at area-level, is the better alternative. The mean and standard deviation of a particular small area estimate, across all $b$ values, then yields a point estimate and its standard error for that area.

2.1.5 Interpretation of standard errors

The standard error of a particular small area estimate is intended to reflect the uncertainty in that estimate. A rough rule of thumb is to take two standard errors on each side of the point estimate as representing the range of values within which we expect the true value to lie. When two or more small area estimates are being compared, for example when deciding on priority areas for receiving development assistance, the standard errors provide a guide for how accurate each individual estimate is and whether the observed differences in the estimates are indicative of real differences between the areas. They serve to reminder users of the maps derived from plotting small area estimates, that the information in them represents estimates which may not always be very precise.
The size of the standard error depends on a number of factors. The poorer
the fit of the model (3.1), in terms of small $R^2$, large $\sigma^2$ or $\sigma^2$, or a large
$\sigma^2/\sigma^2 + \sigma^2$ ratio, the more variation in the target variable will be unexplained and
the greater will be the standard errors of the small area estimates. The population
size, in terms of both the number of households and the number of clusters in the
area, is also an important factor. Generally speaking, standard errors decrease
proportionally as the square root of the population size. Standard errors will be
acceptably small at higher geographic levels but not at lower levels. If we decide to
create a map at a level for which the standard errors are generally acceptable, there
can nevertheless still be some of the smaller areas for which the standard errors are
larger than we would like.

The sample size used in fitting the model is also important. The bootstrapping
methodology incorporates the variability in the estimated regression coefficients
$\hat{\alpha}$, $\hat{\beta}$. If the sample size is small these estimates will be very uncertain and the standard
errors of the small area estimates will be large. This problem is also affected by the
number of explanatory variables included in the auxiliary information, $X$ and $Z$. A
large number of explanatory variables relative to the sample size increases the
uncertainty in the regression coefficients. We can always increase the apparent
explanatory power of the model (i.e. increase the $R^2$ from the survey data) by
increasing the number of $X$ variables, or by dividing the population into distinct
subpopulations and fitting separate models in each, but the increased uncertainty in
the estimated coefficients may result in an overall loss of precision when the model is
used to predict values for the census data. We must take care not to “over-fit” the
model.

Although usually relatively small, there will be some small uncertainty in the
estimates, and indeed the standard errors, due to the bootstrapping methodology,
which uses a finite sample of bootstrap estimates to approximate the distribution of
the estimator. This could be decreased, at the expense of computing time, by
increasing the number of bootstrap simulations $B$.

Finally, the integrity of the estimates and standard errors depends on the
fitted model being correct, in that it applies to the census population in the same way
that it applied to the sample. This relies on good matching of survey and census to
provide valid auxiliary information. We must also take care to avoid, as much as
possible, spurious relationships or artefacts which appear, statistically, to be true in
the sample but do not hold in the population. This can be caused not only by fitting
too many variables, but also by choosing variables indiscriminately from a very large
set of possibilities. Such a situation could lead to estimates with apparently small
standard errors, but the standard errors would be spurious.

The requirement for variables to match in this way between survey and
census is one reason that special care must be taken if survey and census are not
from the same period. The changes between periods can be structural changes, i.e.
the interpretation of a particular variables has changed, or simply a change in level.
Both types of change have the potential to add to standard errors of estimates, and in
some cases to produce bias.

See also Section 5.3 for a discussion of the relative size of estimated
standard errors from ELL in comparison with other small area estimation methods.

2.2 Spatial microsimulation: Synthetic reconstruction, spatial
microsimulation and combinatorial optimisation methods

2.2.1 Aspatial microsimulation
There are microsimulation models that do not take geography into account. It can be argued these are “aspatial” or non-geographical. There is an extensive literature on this topic, beginning with the work of Orcutt (1957), and Orcutt, Greenberger, Korbel and Rivlin (1961). An extensive review is given in O’Donoghue (2001).

Aspatial microsimulation is a technique developed, particularly by economists, that has been widely used for over 50 years. The results of microsimulation models are also widely quoted in the UK and USA media when covering possible impact of government budget changes upon different types of households. The models have aimed to build large-scale data sets on the attributes of individuals or households (and/or on attributes of individual firms or organisations) and to analyse policy impacts based on these micro-units (e.g. Orcutt et al., 1986; Birkin and Clarke, 1995; Clarke, 1996). By analysis at the level of the individual, family or household, they claim to provide the means of assessing variations in the distributional effects of different policies (Hancock and Sutherland, 1992; Mertz, 1991; Harding, 1996; Mitton, Sutherland and Weeks, 2000; Redmond, Sutherland and Wilson, 1998). Microsimulation modelling frameworks have also been used to define the goals of economic and social policy, and to study the effect on people of structural changes due to socio-economic policy measures (Krupp, 1986), and have become accepted tools in the evaluation of economic and social policy, as well as analysis of tax-benefit options and other areas of public policy.

Among the first applied microsimulation models was TAX, developed at the US Treasury department in the 1960s. From the end of the 1960s onwards, microsimulation became the dominant quantitative method for forecasting the impacts of policy changes in social welfare in the USA. Nelissen (1993) gives an overview of these models.

One perceived advantage of microsimulation modelling is the ability to simulate different policy scenarios with the same model. For instance, the OTA model (Office for Tax Analysis) (Nelissen, 1993), developed for personal income tax analysis, was used to simulate thousands of proposals for tax changes.

Statistics Canada (http://www.statcan.ca/english/spsd/) has produced several microsimulation models which assist in policy relevant research development and analysis. One of these models is the Social Policy Simulation Database and Model (SPSD/M) which is a micro computer-based product which was designed to analyse the financial interactions of governments and individuals in Canada and the cost implications and income redistributive effects of changes in the personal taxation and cash transfer system.

Generally, microsimulation models have wider application when they become dynamic, by updating once a microsimulation database is built. Among the first usable dynamic microsimulation models is DYNASIM (DYNAmic Simulation of Income Model; see Orcutt, Greenberger, Korbel and Rivlin, 1961; Wertheimer, Zedlewski, Anderson and Moore, 1986), which was the base for more sophisticated developments such as CORSIM (Cornell Microsimulation Model – Caldwell, Clarke and Keister, 1998) and DYNACAN (Dynamic Microsimulation Model for Canada). One of the descendants of DYNASIM was DYNASIM2, developed and maintained at the Urban Institute in Washington D.C. (Wertheimer et al., 1986) which comprised two sub-models: a Family and Earnings History (FEH) model and a Jobs and Benefit History (JBH) model. The FEH model required an input sample of the US population, which included earnings history information for all the micro-units and it aimed at simulating the basic demographic and labour market activities of the members of the sample for a given number of years. Its JBH sub-model required as input a sample of earnings, employment and demographic histories produced by the FEH sub-model. The output of the JBH sub-model was an augmented set of histories, which it was claimed could be processed like any panel study of a real population (Wertheimer et al., 1986). Other relevant microsimulation models worldwide include DYNACAN in Canada, and DYNAMOD in Australia. DYNACAN is based on the CORSIM template...
and is used for fiscal and policy-oriented analysis of Canadian Social Security Schemes (Caldwell and Morrison, 2000). In Australia, DYNAMOD is a dynamic model of population designed to project population characteristics over a fifty-year period using a one percent population sample (King, Baekgaard and Robinson, 1999).

Other examples of dynamic microsimulation modelling for economic and social policy analysis is NEDYMAS (Netherlands Dynamic Micro-Analytic Simulation model; see Nelissen, 1993), and the LIFEMOD model (Falkingham and Lessof, 1992) which is a dynamic cohort microsimulation model, simulating the life histories of a cohort of 2,000 males and 2,000 females in which each individual experiences major life events such as schooling, marriage, childbirth, children leaving home, employment and retirement.

Atkinson and Sutherland (1988) built TAXMOD, which is a model for the analysis of tax and benefit in the UK. A key difference between TAXMOD and LIFEMOD is that the former imposed a set of rules on an existent microdata population, whereas the latter applied these rules upon a simulated micro-population. A further application of dynamic microsimulation is the work of Hancock, Mallender and Pudney (1992) who built PENSIM, which simulated pensioners’ incomes in the UK up to the year 2030.


As the previous comments and references indicate, microsimulation modelling has been used extensively. Its widest application has been to assess the future effect of policy changes. Such assessments are complicated not only by choice of underlying model, but also by what should be a requirement to assess accuracy of predictions and the long term nature of many of the predictions. Historically, due to limitations of computing power and the size of databases and the need to consider a range of scenarios, modelling accuracy has not been assessed. Models have tended to be run once only for each possible scenario. Where differences in consequences between scenarios has been marked, accuracy is perhaps not so important, but where differences are more subtle, accuracy measures become more so. Models used have often been decided a priori based on expert opinion, and this too has added to uncertainty in predictions for scenarios, because the models have been implicit rather than developed or formally tested statistically. This can be a major complication to the effectiveness of microsimulation, especially where long term predictions and assessments are involved or required.

Despite its variety of uses and applications, microsimulation modelling has been used in the main for conducting analysis at the national level and for undertaking international comparisons. Putting aside modelling and standard error issues for the moment, one possible extension would be to estimate and investigate income and wealth distributions for regions, cities and even smaller areas. Potentially such an extension would allow us to go beyond estimation of indices (such as mean income or percentage below a given threshold) by creating micro-datasets which we can analyse to assess the potential effects of policy scenarios and interventions on these indices. The following section discusses ways in which these models and similar approaches have been adopted in a geographical context.

2.2.2 Spatial microsimulation

Microsimulation models become geographical or spatial when area-based information about the simulated entities is available (or estimated). In particular,
More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics

geographical microsimulation can be defined here as a method to construct small area population microdata for one point in time and then to update these microdata. This definition of microsimulation is different from that used by economists involved in building statistical and mathematical microsimulation models, since these generally do not take geography into account. The focus in spatial microsimulation is to provide small area socio-economic information that can be used for the spatial analysis of policies, as well as the inter-household distributional effects (traditionally the main concern of economists and social policy practitioners).

Spatial microsimulation involves the creation of large-scale population microdata sets and the analysis of policy impacts at the micro-level. Population microdata contain information on individuals rather than aggregate data. Population microdata can be individual microdata that contain information on individuals; household microdata which may contain household information only and household microdata which may contain individual and household information.

This section discusses how various aspatial methods and techniques can be refined and applied in a geographical context in order to provide small area microdata.

Small area microdata can be built with the use of static spatial microsimulation methods. We can distinguish between the following types:

- Synthetic probabilistic reconstruction models, which involve the use of random sampling
- Reweighting probabilistic approaches, which typically reweight an existing national microdata set to fit a geographical area description on the basis of random sampling and optimisation techniques
- Reweighting deterministic approaches, which reweight a non geographical population microdata set to fit small area descriptions, but without the use of random sampling procedures

These approaches are now reviewed in turn.

2.2.3 Synthetic probabilistic reconstruction

Synthetic probabilistic reconstruction models calculate or estimate conditional probabilities of having particular attributes and then assign these attributes on the basis of random sampling procedures (Monte Carlo simulation). Table 1 depicts the steps that need to be followed in the procedure for allocating economic activity status.

Assuming that data on the age, sex, and marital status of each head of household (hh) is available in aggregated form from a published data source such as the Census, it is possible to estimate probabilities of economic activity status conditional upon age, sex, and marital status and location (at the enumeration district level). The first synthetic household in Table 1 has the following characteristics: male head of household, aged 16-29, Single or Widowed or Divorced (SWD), living at location DAFA01 (the code for the first enumeration district in Aireborough ward in Leeds). The estimated probability that a household of this type would be economically active is 70%. The next step in the procedure is to generate a random number to see if the synthetic household gets allocated to the economically active category. The random number in this example is 0.55 which is within the 0.001 - 0.700 range needed to qualify as economically active. The next step is to assign economic activity status. The estimated probability that the head of the first synthetic household is an employee is 60%. Likewise, the estimated probabilities that this particular head of household is self-employed, on a government scheme or
unemployed are 20%, 5% and 15% respectively. Again, a random number is generated and in this example the number is 0.4 and falls within the 0.001 - 0.600 range needed to qualify as an employee. If the random number were larger than 0.6 and smaller than the sum of 0.6 with the next probability rate (i.e. the 0.2 probability of being self-employed) then the head of household would be assigned the according economic activity status. If the random number was larger than this sum then it would be compared to the sum of the first two probabilities plus the next one and so on and so forth.

Table 1: Microsimulation procedure for the allocation of economic activity status
(after the similar example of tenure allocation procedure given by Clarke, 1996: 3)

<table>
<thead>
<tr>
<th>Steps</th>
<th>1st</th>
<th>2nd</th>
<th>...</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, sex and marital status and location (ED level) (given)</td>
<td>Age: 16-29 Sex: Male Marital Status: SWD GeoCode: DAFA01</td>
<td>Age: 75-84 Sex: Female Marital Status: married GeoCode: DAFA02</td>
<td>...</td>
<td>Age: 30-44 Sex: Male Marital Status: married GeoCode: DAGK45</td>
</tr>
<tr>
<td>Probability of head of household (hh) of given age, sex and location (ED level) being economically active</td>
<td>0.7</td>
<td>0.4</td>
<td>...</td>
<td>0.7</td>
</tr>
<tr>
<td>Random number</td>
<td>0.55</td>
<td>0.5</td>
<td>...</td>
<td>0.45</td>
</tr>
<tr>
<td>Economic activity assigned to hh on the basis of random sampling</td>
<td>Economically active</td>
<td>Economically inactive</td>
<td>...</td>
<td>Economically active</td>
</tr>
<tr>
<td>Probability of economically active hh being an employee</td>
<td>0.6</td>
<td>...</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Probability of economically active hh given age, sex, marital status and location (ED level) being self employed</td>
<td>0.2</td>
<td>...</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Probability of economically active hh given age, sex, marital status and location (ED level) being on a government scheme</td>
<td>0.05</td>
<td>...</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Probability of economically active hh given age, sex, marital status and location (ED level) being unemployed</td>
<td>0.15</td>
<td>...</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Random number</td>
<td>0.4</td>
<td>...</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Economic activity category assigned on the basis of random sampling</td>
<td>Employee</td>
<td>...</td>
<td>Self-employed</td>
<td></td>
</tr>
</tbody>
</table>

The same procedure can be carried out sequentially for the assignment of economic activity to all the synthetic households that would be included in a spatial microsimulation database. The procedure involves the crucial, though not fully explicit modelling decision to determine which household attribute should be
estimated first. Birkin and Clarke (1988, p1652-1653) point out that some characteristics are very important in determining others:

It would have been equally possible to start off by assigning some other household attribute, such as tenure, or an individual attribute such as socioeconomic status. In a formal sense, we assume there is some sort of continuum of attribute dependence: some characteristics are very important in ‘determining’ others (at least in a statistical sense - the association need not be causative), and thus need to be assigned at an early stage; whereas others are more dependent, and need to be introduced after the ascription of those characteristics on which they are supposed to depend.

Consequently, one of the most difficult tasks related to the probabilistic synthetic reconstruction method is to specify which variables are dependent upon others and to determine the ordering of probabilities (Birkin and Clarke, 1995; Clarke, 1996; Birkin and Clarke, 1989).

Curiously there is little published discussion of how this issue is linked to statistical modelling. The variables used in synthetic probabilistic reconstruction are almost invariably categorical, and involve counts, so the structure of what is often an inherent model decided a priori is loglinear. This makes it clearer why the order of the variables used in synthetic probabilistic reconstruction is so critical – A followed by B and then C implies a nested log linear model containing A, B(A), and C(A,B) - where B(A) indicate B nested within A etc. - while fitting C followed by B and then A implies the model is C, B(C), A(B,C). Clearly these are very different underlying models unless A, B and C are independent.

Additional references to synthetic probabilistic reconstruction include Williamson (1992); Caldwell and Keister (1996); Caldwell, Clarke and Keister (1998); Hooimeijer (1996); and Wegener and Spiekermann (1996).

2.2.4 Probabilistic reweighting approaches to spatial microsimulation modelling

Since Birkin and Clarke (1988; 1989), there have been major advances in data availability, and in computer hardware and software. These developments, when combined with suitable statistical modelling, have led to better techniques for generating small area microdata. These developments have tended to focus on algorithms for generating “complete” datasets, rather than on the statistical models lying behind the algorithms.

Williamson, Birkin and Rees (1998) suggested a “reweighting” approach to microsimulation modelling, although from a statistical point of view the technique would be better described as a reselection procedure, where sample records, selected from the sample as a whole, and deemed to better match population tables on selected variables than the original sample are added to the database. Williamson, Birkin and Rees used an existing microdata set, available at coarse geographies, adding and replacing observations to build population microdata by getting a better match fine level census statistics for selected variables. In particular, they explored different methods of finding the combination of UK Census household Samples of Anonymised Records (SARs) which best fit known small area constraints. They used various techniques of combinatorial optimisation such as hill climbing algorithms, simulated annealing, and genetic algorithms, to reweight an existing microdata sample from SARs (which only has coarse levels of geography coded at the household level), so that it would fit finer level population-based statistics tables. There have been several later refinements and applications based on Williamson, Birkin and Rees (1998), including Voas and Williamson (2000), Williamson (2002), and Ballas, Clarke, and Wiemers (2005). A simplified description of the “reweighting” or reselection procedure following Williamson (2002) is:
- Use both the sample survey unit record data on a range of variables (e.g. household size, number of adults, number of children), and the known small area constraints (for example from a census) for some of these variables (e.g. totals only by small area for each of household size, for number of adults, and number of children).
- Randomly select two households from the survey sample (not necessarily within the particular small area).
- Tabulate both the selected sample and the census information, and calculate absolute differences (in terms of counts) of the sample from the census table.
- Take one of the sampled units and replace it by a selection from the sample only if it results in a smaller absolute difference.
- Repeat reselection of sample elements until there is no further reduction in absolute difference.

Note that the matching criterion is a simple one, basically counting the extent of mismatching by variable.

Below is a more extensive example to illustrate how the procedure would operate on a more realistic dataset. The example involves reselecting from the British Household Panel Survey (BHPS), described in Tables 2 and 3 to “populate” small areas within a microdata file.

Table 2: The BHPS microdata format

<table>
<thead>
<tr>
<th>PERSON</th>
<th>*HID</th>
<th>PID</th>
<th>*AGE12</th>
<th>SEX</th>
<th>*JBSTAT</th>
<th>...</th>
<th>*HLLT</th>
<th>*QFVOC</th>
<th>*TENURE</th>
<th>*JLSEG</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100208</td>
<td>1002151</td>
<td>90</td>
<td>2</td>
<td>4</td>
<td>...</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>9</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>100382</td>
<td>1004291</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>100382</td>
<td>1001521</td>
<td>25</td>
<td>1</td>
<td>3</td>
<td>...</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>100668</td>
<td>1003857</td>
<td>53</td>
<td>2</td>
<td>1</td>
<td>...</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>101223</td>
<td>1013578</td>
<td>52</td>
<td>2</td>
<td>1</td>
<td>...</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>6</td>
<td>101223</td>
<td>10034608</td>
<td>53</td>
<td>1</td>
<td>2</td>
<td>...</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>7</td>
<td>100668</td>
<td>1003857</td>
<td>53</td>
<td>2</td>
<td>1</td>
<td>...</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>8</td>
<td>100412</td>
<td>1005848</td>
<td>31</td>
<td>1</td>
<td>2</td>
<td>...</td>
<td>2</td>
<td>-7</td>
<td>3</td>
<td>-7</td>
<td>...</td>
</tr>
<tr>
<td>9</td>
<td>100412</td>
<td>10016875</td>
<td>15</td>
<td>2</td>
<td>-8</td>
<td>...</td>
<td>-8</td>
<td>-8</td>
<td>3</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>1001505</td>
<td>10017931</td>
<td>41</td>
<td>2</td>
<td>1</td>
<td>...</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>11</td>
<td>1001505</td>
<td>10017962</td>
<td>45</td>
<td>1</td>
<td>2</td>
<td>...</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>-8</td>
<td>...</td>
</tr>
<tr>
<td>12</td>
<td>1001505</td>
<td>10017991</td>
<td>11</td>
<td>2</td>
<td>-8</td>
<td>...</td>
<td>-8</td>
<td>-8</td>
<td>2</td>
<td>-8</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 3: The BHPS microdata variable descriptions

<table>
<thead>
<tr>
<th>Person</th>
<th>*HID</th>
<th>PID</th>
<th>*AGE12</th>
<th>SEX</th>
<th>*JBSTAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>person number</td>
<td>household identifier (number of household to which the listed individual belongs) (in year *)</td>
<td>person identifier (a unique number to identify the individual) (in year *)</td>
<td>Age at 1 December of in year *</td>
<td>Sex</td>
<td>Current labour force status (e.g. self employed, in paid employment,</td>
</tr>
</tbody>
</table>
More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics

| *HLLT | Health status in year * |
| *QFVOC | Vocational qualifications in year * |
| *TENURE | Tenure status in year * |
| *JLSEG | Socio-economic group: last job (in year *) |

The BHPS is a sample of households and all their occupants. The method samples from all sample records aiming to find the set of household records that best matches the population in the census area statistics tables for each area. First, a set of area tables (e.g. from the census or other sources) that describe the area of interest must be selected, followed by sampling from the BHPS to find a suitable combination of households that would fit the data described by those tables, as in Table 4.

**Table 4**: An example of small area information

<table>
<thead>
<tr>
<th>Small area table A (household type)</th>
<th>Small area table B (economic activity of household head)</th>
<th>Small area table C (tenure status)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area 1</td>
<td>Area 1</td>
<td>Area 1</td>
</tr>
<tr>
<td>60 &quot;married couple households&quot;</td>
<td>80 employed/self-employed</td>
<td>60 owner occupier</td>
</tr>
<tr>
<td>20 &quot;Single-person households&quot;</td>
<td>18 unemployed</td>
<td>20 Local Authority or Housing association</td>
</tr>
<tr>
<td>20 &quot;Other&quot;</td>
<td>2 other</td>
<td>20 Rented privately</td>
</tr>
<tr>
<td>Area 2</td>
<td>Area 2</td>
<td>Area 2</td>
</tr>
<tr>
<td>40 &quot;married couple households&quot;</td>
<td>60 employed/self-employed</td>
<td>60 owner occupier</td>
</tr>
<tr>
<td>20 &quot;Single-person households&quot;</td>
<td>37 unemployed</td>
<td>20 Local Authority or Housing association</td>
</tr>
<tr>
<td>40 &quot;Other&quot;</td>
<td>3 other</td>
<td>20 Rented privately</td>
</tr>
</tbody>
</table>

The aim is to select records for the microdata set from SARs or the BHPS to best match these tables. However, even for a moderately sized census, there are a vast number of possible sets of households that can be drawn from an existing sample. This computational complexity complicates generating the dataset, which is usually done just once, using a wide range of techniques ranging from linear programming to complex combinatorial optimisation methods (Williamson, Birkin and Rees, 1998; Ballas and Clarke, 2000; Ballas, Clarke and Turton, 2003) and simulated annealing (Kirkpatrick, Gelatt and, Vecchi., 1983; Dowsland, 1993; Pham and Karaboga, 2000; van Laarhoven and Aarts, 1987; Openshaw and Schmidt, 1996; Ballas, Kingston, Stillwell and Jin, 2007).

Because probabilistic reweighting approaches to spatial microsimulation modelling are computationally intensive, especially when generating larger datasets, research and publication in this area have focused on methods of generating the dataset itself. There is usually only one such dataset created. The statistical model that is being used to generate the data is left implicit, and not only seldom stated but also seldom tested. Estimated standard errors even under the implicit model are seldom given or discussed, and are indeed impossible to estimate with only one microdata set.

One core assumption in the method is that, given the variables that are to be matched with census tables, no further relationship between key variables and area remains. Another is that all variables used in assessing the matching (through the total absolute difference) are taken to be equally important to the matching score,
although this is a minor matter since this latter measure could be easily adjusted. The central issue however is that (while the computational algorithm used for re-selection of respondents for the microdata set that acts as a pseudo-census is important for computation time and is where the technical discussion of the technique has focused in publication) the criteria for selection of the microdata set are seldom explicitly stated. The choice of the set of area tables (e.g. from the census or other sources) that are selected to describe the area of interest is critical, but treated as a side issue. For example, for the BHPS data above, the implicit model is loglinear with three effects, household type, economic activity of household head and tenure status, all assumed to operate independently of one another.

By focusing on algorithms rather than statistical model choice, the essential question of why a particular random selection would (or would not) be deemed suitable or best for the microdata set has not received the required level of research attention.

2.2.5 Deterministic reweighting approaches to spatial microsimulation modelling

Probabilistic reweighting approaches to spatial microsimulation modelling involve use of random sampling procedures. In contrast, Ballas, Clarke, Dorling, Eyre, Rossiter and Thomas (2005) presented an alternative deterministic approach to reweighting survey microdata so that they fit given small area statistics tables. A particular characteristic of this method is that it does not use random number generators at any stage (hence the term deterministic) and it therefore produces the same results with each run. An example of how this method readjusts the BHPS household records so that they fit small area constraints is described in Tables 5-8. In particular, Table 5 gives a hypothetical individual microdata set comprising five individuals, in two age categories. Table 6 depicts a small area statistics table for a hypothetical area, and Table 7 depicts a cross-tabulation of the hypothetical microdata set, so that it can be compared with Table 6.

**Table 5:** A hypothetical microdata set (original weights: table \( w \))

<table>
<thead>
<tr>
<th>Individual</th>
<th>Sex</th>
<th>Age-group</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(^{st})</td>
<td>Male</td>
<td>Over-50</td>
<td>1</td>
</tr>
<tr>
<td>2(^{nd})</td>
<td>Male</td>
<td>Over-50</td>
<td>1</td>
</tr>
<tr>
<td>3(^{rd})</td>
<td>Male</td>
<td>Under-50</td>
<td>1</td>
</tr>
<tr>
<td>4(^{th})</td>
<td>Female</td>
<td>Over-50</td>
<td>1</td>
</tr>
<tr>
<td>5(^{th})</td>
<td>Female</td>
<td>Under-50</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 6:** Hypothetical small area data tabulation (table \( s \))

<table>
<thead>
<tr>
<th>Age/sex</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under-50</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Over-50</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 7:** The hypothetical microdata set, cross-tabulated by age and sex. (table \( m \))

<table>
<thead>
<tr>
<th>Age/sex</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under-50</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Over-50</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Using these data it is possible to readjust the weights of the hypothetical individuals, so that their sum would add up to the totals given in Table 6. In particular, the weights can be readjusted by multiplying them by the value of the cell in Table 6, which denotes the category in which they belong over the respective cell in Table 7. This can be expressed as:

\[ n_i = w_i \frac{s_{ij}}{m_{ij}} \]

where \( n_i \) is the new household weight for household \( i \), \( w_i \) is the original weight for household \( i \), \( s_{ij} \) is element \( ij \) of table \( s \) (small area statistics table, which is the equivalent of Table 6) and \( m_{ij} \) is element \( ij \) of table \( m \) (reproduced table using the household microdata original weights, the equivalent of Table 7 in the example. Table 8 depicts how this simple formula is used to readjust the weights of the individuals in the example.

### Table 8: Reweighting the hypothetical microdata set in order to fit Table 6.

<table>
<thead>
<tr>
<th>Individual</th>
<th>Sex</th>
<th>age-group</th>
<th>Weight</th>
<th>New weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st}</td>
<td>Male</td>
<td>Over-50</td>
<td>1</td>
<td>( 1 \times 3/2 = 1.5 )</td>
</tr>
<tr>
<td>2\textsuperscript{nd}</td>
<td>Male</td>
<td>Over-50</td>
<td>1</td>
<td>( 1 \times 3/2 = 1.5 )</td>
</tr>
<tr>
<td>3\textsuperscript{rd}</td>
<td>Male</td>
<td>Under-50</td>
<td>1</td>
<td>( 1 \times 3/1 = 3 )</td>
</tr>
<tr>
<td>4\textsuperscript{th}</td>
<td>Female</td>
<td>Over-50</td>
<td>1</td>
<td>( 1 \times 1/1 = 1 )</td>
</tr>
<tr>
<td>5\textsuperscript{th}</td>
<td>Female</td>
<td>Under-50</td>
<td>1</td>
<td>( 1 \times 5/1 = 5 )</td>
</tr>
</tbody>
</table>

The above process can then be used to reweight the individuals to fit another table. This reweighting procedure was adopted iteratively by Ballas, Clarke, Dorling, Eyre, Rossiter and Thomas (2005) to readjust the BHPS households weights so that they would fit the small area statistics tables described in the previous section. The generated weights for each household may represent estimates of the probabilities of the BHPS households of living in a given ward. (See also Ballas, Clarke, Dorling and Rossiter, 2007).

The deterministic method is also reweighting in a statistical sense, as it simply reweights (rather than selects) observations to match census values. In some areas there may not be a suitable observation to reweight, and the reweighting algorithm (which is essentially iterative proportional fitting – IPF) will then fail to produce the required weights. This however is relatively simply fixed, for example by reverting to a probabilistic reweighting approach to provide a suitable additional data point, and continuing the deterministic reweighting process. The statistical model is, like probabilistic reweighting, implicit rather than explicit so similar questions remain about methods and tests for selecting tables to be matched. One important difference however is that area and the variables of interest are not assumed independent given the selected table margins, provided respondents to the survey are only reweighted within the area they come from (unless no suitable respondent exists in area to exactly match census margins), when the (implicit) model is more robust to a poor choice of census tables for matching. With that exception however, the long-term (i.e. asymptotic ) result is that on average probabilistic reweighting and deterministic reweighting would produce the same microdata set and hence results. The algorithmic difference and concerns are consequently less important than the choice of variables from census tables (which collectively determine the statistical model).

### 2.2.6 Microsimulation approaches to small area population projection

As seen earlier in Section 2.2 there are numerous examples of microsimulation models that attempt to simulate populations dynamically and to update their socio-
economic attributes (including income). These include dynamic non-spatial microsimulation models such as CORSIM in the US, DYNACAN in Canada, DYNAMOD in Australia and LIFEMOD and PENSIM in the UK. van Imhoff and Post (1998) provide an extensive discussion of the advantages of microsimulation models and methods as population projection tools.

Dynamic spatial microsimulation allows projection of small area populations and their associated characteristics, including household and individual income. In addition to generating a complete pseudo-population at small area levels statically, microsimulation can also be used to project the characteristics of the small area populations in future years. A simple way of doing this may be to apply a crude projection model for aggregate small area populations and to use these totals as constraints in static spatial microsimulation procedures. In other words, the reweighting methodologies discussed in the previous model can be applied to readjust microdata records (e.g. BHPS households) so that they fit small area constraint data in any selected year.

Stillwell, Birkin, Ballas, Kingston and Gibson (2004) projected numbers of households and individuals for every year to 2021 by ward making the simple assumption that the annual rate of change between 1991 and 2001 will continue until 2021. They calculated for each ward the annual rates of change between 1991 and 2001 for households and individuals and applied these to successive years after 2001 to give ward-based population totals to 2021. They also projected spatially disaggregated counts of the population in future years according to marital status, socio-economic group, number of cars owned etc., by applying annual rates of change between 1991 and 2001 in exactly the same way as with the aggregate populations. So, for example, having projected the car ownership characteristics by household, the next step involved calculating the proportions of all car ownership categories in each ward. These proportions were applied to the projected numbers of households by ward in each future year. This approach ensured that the sum of all cars by household categories added up to the aggregate household projection. The same method was applied for all other household (e.g. tenure) and individual (e.g. ethnic group) variables. The projected counts were then used as constraints in a simulated annealing household reweighting procedure, such as those discussed in the previous section.

This projection model is simple, but there are more sophisticated models of projecting small area populations in a microsimulation framework. Such models rely heavily on the methods developed for non-spatial dynamic microsimulation models but seek to address geographical questions. They have been developed in many countries including Australia, the Netherlands, the UK and Sweden. For example, NATSEM (National Centre for Social and Economic Modelling) at the University of Canberra in Australia, the group which developed DYNAMOD, is developing a spatial microsimulation model to examine issues such as poverty and ageing in a spatial context (Harding, 2002). In particular, a regional microsimulation model has been developed in conjunction with Centrelink, the agency responsible for administering social benefit payments, to project regional demographics and likely use patterns for Centrelink services (King, McLellan and Lloyd, 2002). Also, as mentioned earlier in Section 2.2, Hooimejer (1996) described work in the Netherlands that adopted a spatial microsimulation approach to analyse the linkages between supply and demand in the housing market and labour market simultaneously using a life-course approach to the behaviour of households. Another example from the Netherlands is RAMBLAS, a regional planning model for the Eindhoven region based on the microsimulation of daily activity patterns (Veldhuisen, Kapoen and Timmermans, 2000). In the model, daily activity patterns are used as a basis for predicting the spatial distribution of the demand for various transport services in the urban system.

Other examples of comprehensive dynamic spatial microsimulation include the TOPSIM (Total Population Simulation Models) and SVERIGE (System for
Visualising Economic and Regional Influences Governing the Environment) in Sweden (Holm, Lindgren, Makila and Malmberg, 1996; Vencatasawmy, Holm and Rephann, 1999). These models were built at the Spatial Modelling Centre in Kiruna and are based on the CORSIM template adapted for the small area microdata available in Sweden (Holm, Lindgren, Makila and Malmberg, 1996; Vencatasawmy, Holm and Rephann, 1999). SVERIGE is the first national level spatial microsimulation model and is based on a longitudinal database of socio-economic information (including individual and household income from various sources) on every resident of Sweden between 1985 and 1995. It is aimed at studying the spatial consequences of various national, regional and local public policies, and relies heavily on the national register information which is not available in all countries. The SVERIGE database contains co-ordinates accurate to 100 meters for each resident in Sweden along with various social, economic and demographic characteristics (Vencatasawmy, Holm and Rephann, 1999). SVERIGE is designed to generate geographically detailed reports for policy makers and regional scientists.

It can be argued that in a spatial microsimulation context, the output of a static microsimulation model, such as those described in the previous sub-sections, provides the input for the dynamic microsimulation model. An example is the Spatial Model for the Irish Local Economy (SMILE) which involves the projection of static population forward through time by simulating the processes of mortality, fertility and internal migration (Ballas, Clarke and Wiemers, 2005). In the SMILE model the probability of an individual surviving within the five-year simulation period is assumed to be a function of age, gender and location. Table 9 shows a stylised version of the dynamic spatial microsimulation procedure adopted by the SMILE model and details the method by which mortality is assessed. This is a similar process to the static modelling procedure shown in Table 1. The first synthetic household in Table 9 has the following characteristics: male, aged 25, single, at work and living in the first District Electoral Division (DED) of Leitrim County, Republic of Ireland. As shown in Table 9, the estimated probability that an individual with these characteristics will survive in the period is 0.80. The next step in the procedure is to generate a random number to see if the synthetic individual is predicted to survive. The random number in this example is 0.5 and falls within the 0.001 - 0.80 range so survival is the prediction.

Table 9: A simple example of the microsimulation procedure for mortality (after Ballas, Clarke and Wiemers, 2005).

<table>
<thead>
<tr>
<th>Steps</th>
<th>1st</th>
<th>2nd</th>
<th>...</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, sex and marital status, employment status and location (DED level) (given)</td>
<td>Age: 25</td>
<td>Age: 76</td>
<td>...</td>
<td>Age: 30</td>
</tr>
<tr>
<td></td>
<td>Sex: Male</td>
<td>Sex: Female</td>
<td></td>
<td>Sex: Male</td>
</tr>
<tr>
<td></td>
<td>Marital Status: Single</td>
<td>Marital Status: Married</td>
<td></td>
<td>Marital Status: Married</td>
</tr>
<tr>
<td></td>
<td>Employment Status: At work</td>
<td>Employment Status: Other (e.g. Retired)</td>
<td></td>
<td>Employment Status: At work</td>
</tr>
<tr>
<td>Probability (conditional on sex, age and county location) of survival</td>
<td>0.80</td>
<td>.10</td>
<td>...</td>
<td>0.80</td>
</tr>
</tbody>
</table>
This procedure was followed to estimate mortality, fertility and migration. Every synthetic female in the database is tested for eligibility to give birth. Monte Carlo sampling against the fertility probabilities is used to determine which females give birth. If a birth is deemed to occur, the model creates a new individual. The new individual’s attributes are set as follows: age is zero, sex is determined probabilistically (50 percent probability of each sex), marital status is single, social class and location are that of the mother and all other attributes are left blank. In the next simulation period, the new individual is simulated along with the other individuals in the location. Migration is modelled on the basis of random sampling from calculated migration probabilities derived from the 1991 and 1996 Census of Population data at county level. Probabilities of migrating from one county to another are calculated by age, gender and county location. Migration was also modelled on the basis of Monte Carlo sampling from migration probabilities. In particular, in SMILE the probability of a migration moving to a particular DED is determined by the share of the county population currently residing in every small area. Areas with the biggest populations have the highest probability of attracting migrants. It should be noted that the internal migration modelling capability of this kind of models could be enhanced with the incorporation of more sophisticated procedures such as spatial interaction modelling methods (Fotheringham, Nakaya, Yano, Openshaw and Ishikawa, 2001; Nakaya, 2003; Nakaya, Fotheringham, Hanaoka, Clarke, Ballas and Yano, 2007).

Nevertheless, the assumptions used to model mortality in this kind of modelling do not account for all of the factors known to influence survival. There is evidence that mortality is associated with the quality of life of individuals (Dorling 1997; Mitchel, Dorling and Shaw, 2002). Mitchel, Dorling and Shaw (2000) point out that age, gender, social class and employment status play a very important role in producing geographical inequalities in mortality. However, only age- and location-specific mortality rates were used. Data availability was the primary factor governing their assumptions; however mortality rates could be adjusted for different socio-economic groups. It was also assumed, because of limited data availability, that the mortality rate for individuals over 85 years of age is constant.

Other dynamic simulations use more sophisticated projection models. For example, for projecting the population of Britain for 2001, 2011 and 2021 using data from previous censuses, Ballas, Rossiter, Thomas, Clarke and Dorling (2005) considered a ‘gravity’ model for projecting constraint variables of the form:

\[ A = \exp(\ln W * (\ln w)^2 * \ln u / (\ln v)^3) \]

where \( u, v \) and \( w \) are the smoothed proportions in 1971, 81 and 91 respectively, \( W \) is the observed ward proportion in 1991 and \( A \) is the projected ward proportion in 2001. Ballas, Rossiter, Thomas, Clarke and Dorling (2005) also considered exponential smoothing.

Another recent example of spatial microsimulation modelling work that aims at updating small area population attributes, including household income, is an ongoing project for the UK Department for Transport led by David Simmonds Consultancy (for more details, see http://www.davidsimmonds.com/). This project involves the modelling of individual demographic processes (ageing, death/survival, giving birth), household formation/reformation (beyond the changes in household composition resulting from individual processes), relocation of existing households, and location of new/in-migrant households as well as changes in employment status and location (and hence commuting). In addition, household income is dynamically assigned to all simulated households (at the electoral ward level) in each simulation year on the basis of data from the BHPS.
The results and outputs of using dynamic microsimulation techniques clearly hold great attraction for planners and policy makers, but some notes of caution are warranted. In addition to the complications of static microsimulation, in particular the risk of poorly specified and untested underlying models, there is the additional problem of projecting or predicting data. Again models are at the core in projection, and the time series available are almost inevitably short, making explicit model extrapolation and testing fraught. Sophistication in projection models is a commendable aim, but fitting (let alone testing) such models is far from simple, and the projections remain very dependent on the type of projection model chosen. There is also the temptation to ask quite reasonable policy and other questions that go beyond the ability of the method to answer, given the strong assumption inherent in its construction and fitting. Finally, such models are used to generate microdata which give the superficial appearance of being a census, but is not. Producing multiple datasets for every scenario considered would at least have the advantage of allowing assessment of accuracy conditional on the model being correct, and consequently this multiple imputation approach is highly recommended.

2.3 Mass imputation

Mass imputation is a technique that was trialled and used by Statistics Canada (Colledge, Johnson, Paré and Sande, 1978; Whitridge, Bureau and Kovar, 1990a, 1990b; Kovar and Whitridge, 1995) and Statistics Netherlands (Kooiman, 1998; de Waal, 2000) but has fallen out of favour at both institutions. Mass imputation covers a wide range of techniques, or rather imputation models, with the common features that a high or very high percentage of the data is imputed. Not only a particular variable but a complete record for a respondent may be imputed. It has most often been applied to survey data, usually to supplement it where there is substantial item non-response, or to create a pseudo-census.

At Statistics Canada, mass imputation was first used in the context of two phase sampling of administrative records. An efficient design was used to select a first phase sample from which additional information was collected. As an alternative to using sample survey weighting, imputation was used for the missing parts of the non-sampled primary units to produce a complete, rectangular file. This technique is what Statistics Canada call mass imputation (Kovar and Whitridge, 1995 p. 413). Statistics Canada first applied mass imputation to its Census of Construction data (Colledge et al., 1978), where the imputation rate was approximately 70%. It has also used mass imputation for agricultural income tax data to produce balance sheet estimates, where farmers’ records are missing for operational reasons rather than at random.

At Statistics Netherlands, mass imputation has been largely superseded by iterative reweighting methods that match survey data to a range of consistent tables, some from other surveys, some from administrative registers, without (as for deterministic, but different from probabilistic spatial microsimulation) creating a full pseudo-census. The Dutch call this technique (or perhaps more strictly the data generated from it) a “virtual census”. See van der Laan (2000), Linder (2004) and Houbiers (2004), for example. Although sometimes quoted in support of mass imputation, Houbiers (2004, p. 56-57) actually notes:

In principle, mass imputation offers a simple alternative to estimation by weighting to achieve numerical consistency between estimates from the [Social Statistical Database] SSD. By using some suitable imputation strategy, all missing fields in the SSD can be imputed. Tables can then simply be “counted” from the resulting complete data set. Although imputation models are better when more register information is available, these models are never sufficiently rich to account for all
significant data patterns between sample and register data, and may easily lead
to oddities in the estimates (see Kooiman 1998). Therefore, traditional estimation
by weighting is favored over mass imputation at Statistics Netherlands.

The Dutch virtual census approach is possible because of extensive register data
in Holland (as in Scandinavia); available registers and sizes at 2001 included the
Population Register (16,000,000 records), the Jobs File of employees (6,500,000
records), the Fiscal Administration database (Jobs: 7,200,000 records, and pensions
and life insurance benefits: 2,700,000 records), Social Security Administration
(2,000,000 records), and surveys included the Survey of Employment and Earnings
(3,000,000 records – working hours, place of work) and the Labour Force Survey (2
years, 230,000 records: education, occupation and economic activity). Together
these provide a very rich data source, which makes the virtual census method
feasible, but in many other countries such extensive information is not available, and
this limits general utility of the virtual census method for Official Statistics.

Kovar and Whitridge (1995) make a number of insightful remarks about mass
imputation and its use:
- For many imputation methods there is a corresponding weight adjustment:
  For simple random sampling using the sample mean for imputation is
equivalent to the direct expansion estimator. Using the ratio estimator with
auxiliary data to mass impute for subsampled variables is equivalent to using
a ratio estimator.
- Nearest neighbour imputation is implicitly equivalent to an expansion
  estimator with variable weights corresponding to the number of times each
sampled record is used as a donor.
- Weighting rather than mass imputation is recommended for more complicated
  statistics, such as variances, covariances and correlations.
- Mass imputation “has a place” where “quick, ad hoc estimates are needed, or
  where second-phase sample weights are difficult to calculate…as when
information is missing for operational reasons”.
- For non-random ignorable non-response, mass imputation by nearest
  neighbour methods may be preferable to weighting, since it makes more
extensive use of auxiliary information and multivariate relationships, and may
help attenuate bias.
- Mass imputation performs very poorly where there is non-ignorable non-
  response.
- The choice of imputation method is important.
- Significant bias can be introduced by variables that are not controlled in the
  imputation process.
- Imputed values should be flagged.
- Evaluation of the effect of mass imputation is critical.

Note that using the sample mean for imputation under simple random sampling
will give the same mean as weighting responses, but not the same variance unless
only non-imputed values are used for variance calculation. A parallel situation applies
to using the ratio estimator. That mass imputation may be better when there is non-
random ignorable non-response (Colledge et al. (1978); Michaud, 1987) can also be
interpreted as a comment about the benefit of using auxiliary information and
multivariate relationships in imputation models, and the importance of that modelling
being unbiased – these aspects will be discussed further in Sections 4 and 5. As we
indicate there, the choice of imputation method and its underlying model and fit is not
just important, it is critical. As noted by Kovar and Whitridge, even where the
percentage imputed is much lower than for mass imputation, most imputation
methods perform poorly under non-random ignorable non-response (Rancourt, Lee
and Särndal,1992); further, mass imputation is known to work in some circumstances.
but not in others (Cox and Folsom, 1981; Williams and Folsom, 1981; Cox and Cohen, 1985).

Given these caveats, mass imputation has nevertheless more recently been under discussion and/or in use, in agriculture statistics (Fetter, 2001), at the Australian Bureau of Statistics (ABS, 2003), at Statistics Norway (Gåsemyr, Børke and Andersen, 2007), at INSEE in France (Brion, 2008), and in results given at various conferences (e.g. Kozak, 2005; Kroti, Black and Creel, 2005). Other earlier uses of mass imputation included US corporate tax returns (Hinkins and Scheuren, 1986), and industry and occupation classification (Clogg, Rubin, Schenker, Schultz and Weidman, 1991).

Regardless of whether multiple imputation is used to allow estimates of accuracy given the model, the success of the mass imputation technique itself depends critically on the adequacy of the imputation model, both in terms of model type and the variables included in it. Simple hot-decking, i.e. using a single pass through the data and replacing missing records from imputation classes formed from cross-tabulations of the data, is not generally suitable for mass imputation. Simple reweighting is generally preferred because it removes a random element due to random record choice, it is computationally less intensive, and it has a large body of theory detailing its properties. This is why Statistics Netherlands prefers its virtual census, which at core is simply a sample reweighting method under a calibration model with new margins specified, in the main, via administrative data.

Nearest neighbour techniques have been popular with mass imputation. This imputation method has the benefit of greater control in the choice of imputed records and part records, but again the theory is only partially developed. A more fruitful approach, in business surveys at least, seems to be using an update of the same respondent’s information from a previous period, which has the advantage that it reflects any ongoing business characteristics, which are often highly correlated over time. Again however, the theory is not fully developed. In all situations, multiple imputation is required for accuracy estimates, but even when this is used, variance estimates remain conditional on the mass imputation model being correct.

### 2.4 Associated techniques

Mass imputation is only one of a variety of imputation techniques. Others include multiple imputation, fractional imputation, various varieties of hot deck ing, deterministic and stochastic imputation. These methods are not mutually exclusive. In particular mass imputation, even when appropriate, is best done many times to produce not one simulated dataset but many, since this allows some assessment of standard errors due to the imputation, and this generation of many such datasets is called multiple imputation. A very useful reference for mass imputation linked to business surveys remains Kovar and Whitridge (1995).

As part of the procedure, to impute for missing data, fractional imputation adds a fraction of a randomly chosen residual to a regression-based predictor, where the fraction is a function of independent variables in a regression. Although (unlike ELL) fractional imputation is usually applied to survey data, it has a connection to ELL where the residuals used in the bootstrapping are also scaled or “unshrunk”. The link between fractional imputation and hot-decking is discussed for example in Kim and Fuller (2004) and also in Qin, Rao and Ren (2008). More general references include Wolter (2007) and Särndal (2005).

Inverse sampling (Hinkins, Oh and Scheuren, 1997; Rao, Scott and Benhin, 2002; 2003) rather than trying to recreate a complete census from survey data instead subsamples the survey data (perhaps many times) to produce a dataset (or datasets) that can be analysed as though they were simple random samples. A
More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics

simple example is for a cluster sample with equal cluster sizes and sample size within sampled clusters, where each subsample would contain one unit selected randomly from each sampled cluster. This technique while it can produce multiple datasets is not (like mass imputation, ELL-type small area estimation, or spatial microsimulation) intended to produce a ‘complete’ dataset, even for a survey.

Record linkage methods also combine data sources to create a single dataset in the absence of unique identifiers, but unlike the methods considered here the datasets are usually of similar size. Generalized regression (GREG) estimation methods are usually applied to sample survey data only, rather than used to combine datasets.

M-quantile estimation (Chambers and Tzavidis, 2006; Tzavidis, and Chambers, 2005; Tzavidis, Salvati, Pratesi and Chambers, 2006; Betti and Neri, 2007) has been suggested as an alternative to ELL, but has not been assessed in this study.

Reweighting of data, essentially using calibration, raking, or the iterative proportional fitting algorithm (which are essentially equivalent techniques) is an inherent part of post-data collection procedures for survey data in many government statistical agencies. The aim is to adjust and thus match different tables to be consistent, i.e. to have the same margins for the same variables (or to match survey totals with census ones where known by adjusting survey weights). Of particular relevance to mass imputation are the reweighting techniques preferred and used by Statistics Netherlands to produce cross-tabulation tables that could also be constructed (although with lower accuracy given the available models) using mass imputation. See for example, Linder (2003), Schulte (2005), Houbiers (2004), van der Laan (2000), and de Waal (2000).

IPF and SPREE (Structure PREserving Estimation) methods are also inherently related to spatial microsimulation where survey microdata are chosen or rescaled to match census margins, and hence to loglinear models (see, for example, Noble, Haslett and Arnold, 2002), the generalised version of SPREE, GSPREE (Zhang and Chambers, 2004) and its extended version, ESPREE (Isidro, 2009). As the reviewer has noted, there are also very close links to ecological inference and data fusion, except there is then no sample model linking Y and X. Recent developments in the ecological inference literature (see Steel, Beh and Chambers, 2004) have identified the considerable gains from having even a small amount of linked data, while researchers in data fusion are now (somewhat belatedly) coming to the same conclusion.
3 A Unifying Theoretical Framework for ELL-Type-Small-Area Estimation, Spatial Microsimulation, and Mass Imputation

ELL-type small-area estimation, spatial microsimulation and mass imputation are all techniques that use survey and partial census or administrative data to create a pseudo-census. While there are important, if not critical differences between them that affect their utility, these tend to hinge on the model chosen, how it is selected, fitted and tested, and whether the method, as currently used, provides estimated standard errors conditional on the model.

In this section we do not focus on these differences, although as we show later, they do help explain why the three methods do not work equally well. Instead we consider why and in what ways the three methods are fundamentally similar.

We assume that the object of interest is a (possibly nonlinear) function of the complete census data, say \( \phi(C) \). In general, the operator \( \phi(.) \) will act on a target variable or variables \( Y \) contained in the census; for example in small area estimation \( \phi(.) \) will typically produce subpopulation means of \( Y \) – a linear function – whereas in small-area estimation of poverty “incidence” the object is the subpopulation proportion of \( Y \) (income or expenditure) values below a threshold – a nonlinear function. In some uses of spatial microsimulation, \( \phi(.) \) may involve the application of a simulation model to household - or individual- level data; this too can be regarded as a nonlinear function of the census observations.

All three methodologies have been developed to cope with situations in which the complete census data is unobserved. We write formally:

\[
C = C_o + C_u
\]

where \( C_o, C_u \) denote respectively the observed and unobserved portions of the full census data. In small-area estimation situations, for example, we often have complete census data for “auxiliary variables” \( X \), but only partial, survey-derived, data for the target variable \( Y \). In mass imputation and spatial microsimulation, some records may be more incomplete than others, and many census observations may be missing most of the auxiliary data in addition to the variable(s) of interest.

We seek to replace the target:

\[
\phi(C) = \phi(C_o + C_u)
\]

by an estimate:

\[
\phi(C^*) = \phi(C_o + C_u^*)
\]

in which the missing data \( C_u \) is replaced by a surrogate \( C_u^* \). This surrogacy is or should be informed by a “model”, i.e. a set of assumptions or a fitted statistical structure that tells us what to expect for \( C_u \) based on \( C_o \).

In small-area estimation with an explicit linear model, \( C_u^* \) will be the expected value of \( C_u \) conditional on the observed data:

\[
\phi(C^*) = \phi(C_o + E[C_u \mid C_o])
\]
In poverty estimation, where incidence is a nonlinear function of income or expenditure, the ELL method is targeted at:

\[ \varphi(C^*) = E[\varphi(C_o + C_U) | C_o] \]

In stochastic microsimulation, the actual households in an area for which complete data is unavailable are replaced by a set of households with complete data, chosen to match some of the characteristics of the actual households. This again implies a model, since it assumes that the characteristics matched are useful in predicting the variable(s) of interest. Again denoting the variable(s) of interest by \( Y \), and the matching characteristics by \( X \), the area-level summaries \( \varphi(C) \) are approximated by draws from the distribution of:

\[ \varphi(C^*) = \varphi(C_o + C_U | C_o) \]

If deterministic microsimulation is used (via iterative proportional fitting) or if multiple stochastic microsimulation estimates are averaged, the situation is then exactly the same as that for small-area estimation as detailed above.

In mass imputation, incomplete records in the census have their missing portions replaced using complete records that match on the non-missing portions. Here again we can regard \( C_U \) as a random draw from the distribution of \( C_U | C_o \). If multiple imputations are used, these can be averaged to again give:

\[ \varphi(C^*) = E[\varphi(C_o + C_U) | C_o] \]

Because some of the auxiliary data \( X \) may be missing, in addition to some of the \( Y \) values, this process is equivalent to a model-based small-area estimation in which noise has been added to some of the \( X \) variables, the amount of noise being determined by the amount of missingness in \( X \) and the size of the model errors in the imputation of the missing \( X \)s.

This framework forms the basis for the simulation study presented in the next chapter.
4. Simulation Study

An extensive simulation study, with a number of components, was used to study and compare small area estimation (ELL), spatial microsimulation and mass imputation. These components are discussed in the following subsections.

Simulation was based on the New Zealand government projects mentioned in Section 1.1 (some of which are detailed more fully of Section 1.3) rather than using Official Statistics data directly. The principle reason for this choice is that by using simulation it is possible construct a census for which the true values of parameters, and means within small areas (which are the main output of the simulations, rather than more complex statistics), are known. It is also possible to know the true standard deviations of small area means, and to repeat the process many times.

Census data was generated based on a known model. In different contexts, a single fixed census, or a collection of censuses under the same superpopulation model, were generated. Several different superpopulation models were used, since this provided additional statistical modelling context for the fixed census, as the fixed census becomes a sample (of one) census from the superpopulation. Having generated a given census, data were deleted selectively. The pattern of deletion depended on which of the three techniques was being considered, as in part they are used to ‘re-complete’ census datasets in slightly different situations. From each census, a sequence of samples each under a specified design was drawn. This design structure for the simulation allowed the effect of census, sample, and characteristics of the three techniques to be studied.

For small area estimation, the deletions were for values of one variable only for those observations that were not selected from the census observations using a well-specified sample survey. The sample was selected many times using the same sample design, and a range of designs were used. Letting the variable for which there were deletions (and for which samples were drawn) be denoted $Y$, and the other census variables be $X$, an additional research issue was the effect of errors in $X$ on ELL small area estimation (for example from projection into the future). A core question for ELL small area estimation is whether its estimated standard errors for small area means are very much too small; this is also considered as part of the simulations.

Probabilistic reweighted spatial microsimulation generates missing census observations by reusing known complete observations, i.e. it is an imputation scheme. However, for such spatial microsimulation, part of all census records is already known, since these known variables are what are used (as an implicit model via imputation classes) to limit the selection to a subclass of complete records with the same values of the known variables. Superficially, spatial microsimulation and small area estimation seem unrelated, but if we consider the case where there is only one unknown variable, and where complete records are derived from a sample survey then the similarities become clearer, especially since ELL-type small area estimation can in principle be extended to multivariate $Y$. (See Jones, Haslett and Parajuli, (2006), for example, for some further discussion of multivariate $Y$ in the context of stunting, underweight and wasting in children under five years of age.).

Deterministic reweighted spatial microsimulation is a little different, since observations are not necessarily replicated by imputation but instead reweighted. However for integer weights, the effect on averages (but not necessarily their standard errors) is the same as for replicating observations, and (in concept at least ) non-integer weights are only a small departure from this situation.

For mass imputation, complete or near complete records are generally replaced, but if records used for replacement are not to be selected completely at random (which would be a poor imputation strategy anyway) some $X$ variables are again required known for all census observations to limit the selection. Again, as for
spatial microsimulation, the model is implicit and usually involves imputation classes, and \( Y \) is multivariate.

The sections that follow consider ELL-type small area estimation, its comparison with spatial microsimulation, the effect of model uncertainty, and mass imputation and its links to the other two methods. Given Section 3, and the unifying model that it contains and details for these three methods, strong similarities between them will be noted. The similarities mean it is possible to estimate standard errors (conditional on the statistical model used) for all three methods, an aspect that has not been noted at all in the spatial microsimulation literature, and which is not the direct focus of the mass imputation literature either (except where multiple imputation has been used). The simulations also tease out differences between the three techniques, especially in how they are currently used, and how well and clearly their underlying models are specified and tested.

Note that spatial microsimulation, seemingly deliberately, does not include small area level effects in its models (even though it could), instead choosing the sample values to use within a given small area from the entire sample. Mass imputation does not generally include area effects either. Since the main aim of the research was to compare these two techniques and ELL, small area effects have deliberately been left out of the ELL and other simulations.

As an aside, Jones, Haslett and Parajuli (2006) have tried including small area level random effects in a version of ELL in the small area estimation of poverty study for WFP in Nepal. They found (to the extent it is possible to assess using existing methods), that the effects at the small area level in Nepal were so small that they had negligible effect on the estimated standard errors for the small area estimates. What drives the success of ELL, at least in relation to the question of small area level random effects, is that the ratio of variance components at small area and cluster level is small, and the number of clusters in the population in each small area is large.

4.1 Investigation of small area estimation – ELL method

The first phase of the simulation study considered only small area estimation by the ELL method. It involved generating a census under a known statistical model, selecting a sample using a known sampling scheme, estimating a mixed linear model from survey data, using the model to predict for census households, and then amalgamating to Areas.

The census data was generated under a model that assumed a structure made up of Country, Area (i.e. small area), Cluster (each of which constituted a primary sampling unit - psu - when sampling), and Household, so that households were nested in clusters, clusters in areas, areas within country.

Two census / sample design combinations were used:

i. Census: 25 Areas; 250 Clusters per Area, 200 Households per Cluster;
   Sample: 50 Clusters per Area, 12 Households per sampled Cluster.

ii. Census: 25 Areas; 25 Clusters per Area, 2000 Households per Cluster;
   Sample: 5 Clusters per Area, 120 Households per sampled Cluster.

All the samples drawn were self-weighting. This decision simplified the simulation, but is not critical, especially since most applications of the ELL method use household surveys where, even if not equal, sample weights are similar by design. Some methods for estimating and predicting are easily adapted to include unequal weights, others not so easily.

The simulation for testing the ELL small area estimation method did not include an Area-level effect, i.e. in a superpopulation context, the expected mean of \( Y \) for an area, taken over censuses drawn from the superpopulation, did not differ from the
country average. This does not however indicate that there are no area effects for a given (fixed) census drawn from this superpopulation.

The census data comprises one $Y$ and nine $X$ variables, originally multivariate normal with moderate correlations at household level, with added cluster-level multivariate random effects; and with some $X$ variables then rounded or changed into categories. This census generation method does not assume a strong superpopulation model, essentially only specifying that variables are correlated and that there are “country-wide” cluster effects. The ELL small area model is more tightly specified: it fits a linear mixed model to the sample data for $Y$, so that for $Y$ there are cluster level effects in the fitted model. The link to multivariate cluster-level random effects generated in ($Y: X$) is that, strictly speaking, unless the model is fitted conditional on $X$, a measurement error model is required to fit the survey data for $Y$ rather than a linear mixed model, since the measurement error model would allow explicitly for the model errors (from the cluster level effects) in ($Y: X$), not just in $Y$. However the ELL-type small area estimation method does not use a measurement error model, so strictly speaking the model is misspecified. However this is also the situation in practice, so the simulation replicated the real-life situation and assessed whether the ELL-type model worked sufficiently well and gave adequate estimated standard errors (and/or mean square errors) in this situation.

For each sample taken, a linear mixed model for $Y$ was fitted to this sample and used to predict for all households in the census; these were used to calculate estimates and confidence intervals of the true average $Y$ for each Area. This was repeated 1000 times to give 1000 estimates for each Area, from which estimates of confidence interval coverage, bias etc. could be calculated. See Tables 10a-d.

Two methods were used for estimating $\beta$ of equation (3.1): weighted least squares (survey regression – sr) or mixed model with random cluster effects (lm) with weights taken as one, since the sample design was self-weighting.

For a fixed census, both regression fitting methods will have some Areas where confidence interval coverage based on standard errors is much lower than nominal. This is essentially a consequence of the differences between Areas being fixed for a fixed census, for every sample drawn, even though values within Area are generated randomly from the superpopulation (and that, taken over many realisations of the census, Area effects are zero). For a fixed census, coverage for the survey regression (i.e. sr or WLS) estimates tends to be lower than for the linear mixed model (lm or LME) fit. However, this does not indicate that sr is not performing as well as lm - in fact to the contrary - since coverage problems are a direct consequence of the fixed census having fixed area effects that unless explicitly included introduce bias. When area effects are incorporated with the standard errors to calculate mean square errors, coverage by either method is no longer an issue. In most practical situations (c.f. Haslett, Noble and Zabala, 2008) there is only one census available, so that the census, and consequently areas, must be treated as fixed. In practice, this is precisely what is done in ELL: the model of equation (3.1) fitted to the survey data may be extended to incorporate random area effects, but the prediction of area effects for each particular area is what is sought for planning and policy purposes. In terms of coverage, even without incorporating fixed area effects to adjust for what otherwise is bias, there was better performance when survey design has many clusters per area; poorer when there are few; See Figure 1.

For a superpopulation for the census, LME is close to nominal coverage, WLS is more conservative with coverage being a little less than the nominal level.

Although not required for estimation of the area means of $Y$, ELL uses bootstrapping of residuals so it can estimate of standard errors of the area averages. Two methods of bootstrapping for standard errors were considered for the simulation (as in ELL) – nonparametric using residuals from fit or parametric using variance decomposition:
In terms of coverage, the parametric bootstrap was better than the nonparametric (in terms of confidence interval coverage) when there were small numbers of clusters, presumably because of “graininess” in the empirical distribution function of the latter; otherwise there was not much difference.

Two results of the simulation study for ELL-type small area estimation deserve comment. The first is that there seems to be benefit in considering and testing for inclusion of Area-level random effects in equation (3.1) to improve fixed-census properties. As specified by Elbers, Lanjouw and Lanjouw (2003) and in most consequent applications, the model does not include such random area-level effects only cluster-level and household-level ones. Elbers, Lanjouw and Lanjouw (2003) favour fitting different models to different strata which may mitigate the need for random area-level effects, but this strategy raises complications because such models are then fitted to rather small samples within strata, losing information about the overall design and making overfitting (see for example Miller, 2002) an additional complication. In practice, at least for small area estimation of poverty in Nepal (Jones, Haslett and Parajuli, 2006) estimation of variance components suggested that the area-level effects would be small, resulting in negligible change to area estimates and their estimated standard errors, even for a country-wide model. The second is that estimated standard errors from ELL-type small area estimation are very similar to their true values, not markedly smaller. This is an important issue, since underestimating standard errors would lead to concluding results from ELL are much more accurate than they actually are, and misallocating resources (e.g. food aid) programmes as a result. This important issue is discussed in more detail in Section 5.
Table 10a: Coverage of nominal 95% confidence intervals

<table>
<thead>
<tr>
<th>Area</th>
<th>sr.p</th>
<th>sr.np</th>
<th>lm.p</th>
<th>lm.np</th>
<th>Census 1</th>
<th>sr.p</th>
<th>sr.np</th>
<th>lm.p</th>
<th>lm.np</th>
<th>Census 2</th>
<th>sr.p</th>
<th>sr.np</th>
<th>lm.p</th>
<th>lm.np</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.946</td>
<td>0.946</td>
<td>0.978</td>
<td>0.971</td>
<td>0.938</td>
<td>0.938</td>
<td>0.965</td>
<td>0.950</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.952</td>
<td>0.952</td>
<td>0.980</td>
<td>0.973</td>
<td>0.940</td>
<td>0.941</td>
<td>0.967</td>
<td>0.950</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.945</td>
<td>0.946</td>
<td>0.974</td>
<td>0.974</td>
<td>0.937</td>
<td>0.938</td>
<td>0.963</td>
<td>0.951</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.904</td>
<td>0.902</td>
<td>0.953</td>
<td>0.944</td>
<td>0.933</td>
<td>0.933</td>
<td>0.966</td>
<td>0.952</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.947</td>
<td>0.947</td>
<td>0.978</td>
<td>0.975</td>
<td>0.943</td>
<td>0.941</td>
<td>0.969</td>
<td>0.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.949</td>
<td>0.948</td>
<td>0.976</td>
<td>0.974</td>
<td>0.940</td>
<td>0.941</td>
<td>0.964</td>
<td>0.954</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.945</td>
<td>0.945</td>
<td>0.975</td>
<td>0.973</td>
<td>0.941</td>
<td>0.937</td>
<td>0.961</td>
<td>0.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.919</td>
<td>0.918</td>
<td>0.956</td>
<td>0.954</td>
<td>0.935</td>
<td>0.935</td>
<td>0.967</td>
<td>0.954</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.824</td>
<td>0.825</td>
<td>0.909</td>
<td>0.902</td>
<td>0.934</td>
<td>0.934</td>
<td>0.964</td>
<td>0.948</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.937</td>
<td>0.937</td>
<td>0.971</td>
<td>0.965</td>
<td>0.938</td>
<td>0.937</td>
<td>0.960</td>
<td>0.953</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.937</td>
<td>0.937</td>
<td>0.975</td>
<td>0.964</td>
<td>0.939</td>
<td>0.938</td>
<td>0.967</td>
<td>0.957</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.938</td>
<td>0.938</td>
<td>0.974</td>
<td>0.968</td>
<td>0.936</td>
<td>0.936</td>
<td>0.959</td>
<td>0.950</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.797</td>
<td>0.798</td>
<td>0.884</td>
<td>0.861</td>
<td>0.916</td>
<td>0.913</td>
<td>0.944</td>
<td>0.935</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.952</td>
<td>0.952</td>
<td>0.977</td>
<td>0.974</td>
<td>0.942</td>
<td>0.942</td>
<td>0.961</td>
<td>0.949</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.920</td>
<td>0.920</td>
<td>0.957</td>
<td>0.952</td>
<td>0.937</td>
<td>0.935</td>
<td>0.960</td>
<td>0.947</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.950</td>
<td>0.950</td>
<td>0.980</td>
<td>0.974</td>
<td>0.942</td>
<td>0.942</td>
<td>0.966</td>
<td>0.953</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.861</td>
<td>0.860</td>
<td>0.923</td>
<td>0.910</td>
<td>0.924</td>
<td>0.925</td>
<td>0.952</td>
<td>0.938</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.919</td>
<td>0.918</td>
<td>0.959</td>
<td>0.952</td>
<td>0.937</td>
<td>0.936</td>
<td>0.957</td>
<td>0.944</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.931</td>
<td>0.931</td>
<td>0.968</td>
<td>0.962</td>
<td>0.938</td>
<td>0.937</td>
<td>0.961</td>
<td>0.951</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.948</td>
<td>0.948</td>
<td>0.980</td>
<td>0.974</td>
<td>0.940</td>
<td>0.940</td>
<td>0.961</td>
<td>0.953</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.721</td>
<td>0.722</td>
<td>0.825</td>
<td>0.819</td>
<td>0.924</td>
<td>0.921</td>
<td>0.955</td>
<td>0.937</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.937</td>
<td>0.937</td>
<td>0.971</td>
<td>0.959</td>
<td>0.937</td>
<td>0.935</td>
<td>0.965</td>
<td>0.955</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.750</td>
<td>0.749</td>
<td>0.841</td>
<td>0.827</td>
<td>0.909</td>
<td>0.907</td>
<td>0.947</td>
<td>0.930</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.936</td>
<td>0.936</td>
<td>0.974</td>
<td>0.969</td>
<td>0.940</td>
<td>0.938</td>
<td>0.963</td>
<td>0.957</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.950</td>
<td>0.951</td>
<td>0.981</td>
<td>0.970</td>
<td>0.939</td>
<td>0.938</td>
<td>0.963</td>
<td>0.951</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

sr.p and sr.np: survey regression with parametric and non-parametric bootstrap respectively
lm.p and lm.np: mixed model regression with parametric and non-parametric bootstrap respectively
More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics

Table 10b: Average length of nominal 95% confidence intervals

<table>
<thead>
<tr>
<th>Area</th>
<th>Census 1</th>
<th></th>
<th></th>
<th>Census 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sr.p</td>
<td>sr.np</td>
<td>lm.p</td>
<td>lm.np</td>
<td>sr.p</td>
<td>sr.np</td>
</tr>
<tr>
<td>1</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
<tr>
<td>2</td>
<td>0.0228</td>
<td>0.0228</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>3</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0714</td>
</tr>
<tr>
<td>4</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0714</td>
</tr>
<tr>
<td>5</td>
<td>0.0228</td>
<td>0.0228</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>6</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0259</td>
<td>0.0716</td>
<td>0.0714</td>
</tr>
<tr>
<td>7</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
<tr>
<td>8</td>
<td>0.0228</td>
<td>0.0228</td>
<td>0.0268</td>
<td>0.0259</td>
<td>0.0717</td>
<td>0.0715</td>
</tr>
<tr>
<td>9</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
<tr>
<td>10</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0713</td>
</tr>
<tr>
<td>11</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
<tr>
<td>12</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0259</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
<tr>
<td>13</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>14</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>15</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>16</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>17</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
<tr>
<td>18</td>
<td>0.0228</td>
<td>0.0228</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>19</td>
<td>0.0228</td>
<td>0.0228</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0717</td>
<td>0.0716</td>
</tr>
<tr>
<td>20</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0259</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>21</td>
<td>0.0228</td>
<td>0.0228</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0717</td>
<td>0.0715</td>
</tr>
<tr>
<td>22</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
<tr>
<td>23</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
<tr>
<td>24</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0716</td>
<td>0.0715</td>
</tr>
<tr>
<td>25</td>
<td>0.0227</td>
<td>0.0227</td>
<td>0.0268</td>
<td>0.0260</td>
<td>0.0715</td>
<td>0.0714</td>
</tr>
</tbody>
</table>
Table 10c: Bias = Average of (Estimate – True Value)

<table>
<thead>
<tr>
<th>Area</th>
<th>Census 1</th>
<th>Census 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sr.p</td>
<td>sr.np</td>
</tr>
<tr>
<td>1</td>
<td>-0.0009</td>
<td>-0.0009</td>
</tr>
<tr>
<td>2</td>
<td>-0.0004</td>
<td>-0.0004</td>
</tr>
<tr>
<td>3</td>
<td>-0.0011</td>
<td>-0.0011</td>
</tr>
<tr>
<td>4</td>
<td>0.0038</td>
<td>0.0038</td>
</tr>
<tr>
<td>5</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td>6</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td>7</td>
<td>-0.0011</td>
<td>-0.0011</td>
</tr>
<tr>
<td>8</td>
<td>0.0032</td>
<td>0.0032</td>
</tr>
<tr>
<td>9</td>
<td>0.0058</td>
<td>0.0058</td>
</tr>
<tr>
<td>10</td>
<td>-0.0018</td>
<td>-0.0018</td>
</tr>
<tr>
<td>11</td>
<td>0.0018</td>
<td>0.0018</td>
</tr>
<tr>
<td>12</td>
<td>-0.0016</td>
<td>-0.0016</td>
</tr>
<tr>
<td>13</td>
<td>-0.0066</td>
<td>-0.0066</td>
</tr>
<tr>
<td>14</td>
<td>-0.0003</td>
<td>-0.0003</td>
</tr>
<tr>
<td>15</td>
<td>-0.0027</td>
<td>-0.0027</td>
</tr>
<tr>
<td>16</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td>17</td>
<td>-0.0051</td>
<td>-0.0051</td>
</tr>
<tr>
<td>18</td>
<td>-0.0028</td>
<td>-0.0028</td>
</tr>
<tr>
<td>19</td>
<td>-0.0021</td>
<td>-0.0021</td>
</tr>
<tr>
<td>20</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td>21</td>
<td>0.0080</td>
<td>0.0080</td>
</tr>
<tr>
<td>22</td>
<td>0.0021</td>
<td>0.0021</td>
</tr>
<tr>
<td>23</td>
<td>-0.0075</td>
<td>-0.0075</td>
</tr>
<tr>
<td>24</td>
<td>0.0019</td>
<td>0.0019</td>
</tr>
<tr>
<td>25</td>
<td>-0.0007</td>
<td>-0.0007</td>
</tr>
</tbody>
</table>
### Table 10d: Standard Error (Standard Deviation of Estimate)

<table>
<thead>
<tr>
<th>Area</th>
<th>Census 1 sr.p</th>
<th>Census 1 sr.np</th>
<th>Census 1 lm.p</th>
<th>Census 1 lm.np</th>
<th>Census 2 sr.p</th>
<th>Census 2 sr.np</th>
<th>Census 2 lm.p</th>
<th>Census 2 lm.np</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00570</td>
<td>0.00570</td>
<td>0.00568</td>
<td>0.00568</td>
<td>0.01904</td>
<td>0.01904</td>
<td>0.01906</td>
<td>0.01905</td>
</tr>
<tr>
<td>2</td>
<td>0.00572</td>
<td>0.00572</td>
<td>0.00569</td>
<td>0.00568</td>
<td>0.01908</td>
<td>0.01908</td>
<td>0.01909</td>
<td>0.01909</td>
</tr>
<tr>
<td>3</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00569</td>
<td>0.00567</td>
<td>0.01903</td>
<td>0.01904</td>
<td>0.01896</td>
<td>0.01904</td>
</tr>
<tr>
<td>4</td>
<td>0.00571</td>
<td>0.00570</td>
<td>0.00570</td>
<td>0.00567</td>
<td>0.01903</td>
<td>0.01903</td>
<td>0.01911</td>
<td>0.01904</td>
</tr>
<tr>
<td>5</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00569</td>
<td>0.00567</td>
<td>0.01906</td>
<td>0.01906</td>
<td>0.01907</td>
<td>0.01906</td>
</tr>
<tr>
<td>6</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00569</td>
<td>0.00569</td>
<td>0.01903</td>
<td>0.01903</td>
<td>0.01901</td>
<td>0.01904</td>
</tr>
<tr>
<td>7</td>
<td>0.00570</td>
<td>0.00570</td>
<td>0.00566</td>
<td>0.00566</td>
<td>0.01902</td>
<td>0.01902</td>
<td>0.01907</td>
<td>0.01903</td>
</tr>
<tr>
<td>8</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00569</td>
<td>0.00568</td>
<td>0.01907</td>
<td>0.01906</td>
<td>0.01916</td>
<td>0.01906</td>
</tr>
<tr>
<td>9</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00568</td>
<td>0.00568</td>
<td>0.01902</td>
<td>0.01902</td>
<td>0.01909</td>
<td>0.01902</td>
</tr>
<tr>
<td>10</td>
<td>0.00569</td>
<td>0.00569</td>
<td>0.00566</td>
<td>0.00568</td>
<td>0.01900</td>
<td>0.01901</td>
<td>0.01900</td>
<td>0.01901</td>
</tr>
<tr>
<td>11</td>
<td>0.00569</td>
<td>0.00570</td>
<td>0.00568</td>
<td>0.00567</td>
<td>0.01901</td>
<td>0.01901</td>
<td>0.01903</td>
<td>0.01902</td>
</tr>
<tr>
<td>12</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00567</td>
<td>0.00568</td>
<td>0.01904</td>
<td>0.01904</td>
<td>0.01908</td>
<td>0.01903</td>
</tr>
<tr>
<td>13</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00570</td>
<td>0.00568</td>
<td>0.01904</td>
<td>0.01904</td>
<td>0.01906</td>
<td>0.01903</td>
</tr>
<tr>
<td>14</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00570</td>
<td>0.00569</td>
<td>0.01905</td>
<td>0.01905</td>
<td>0.01907</td>
<td>0.01905</td>
</tr>
<tr>
<td>15</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00567</td>
<td>0.01903</td>
<td>0.01903</td>
<td>0.01909</td>
<td>0.01903</td>
</tr>
<tr>
<td>16</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00568</td>
<td>0.00569</td>
<td>0.01907</td>
<td>0.01906</td>
<td>0.01904</td>
<td>0.01908</td>
</tr>
<tr>
<td>17</td>
<td>0.00570</td>
<td>0.00570</td>
<td>0.00571</td>
<td>0.00569</td>
<td>0.01903</td>
<td>0.01902</td>
<td>0.01902</td>
<td>0.01902</td>
</tr>
<tr>
<td>18</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00570</td>
<td>0.00569</td>
<td>0.01905</td>
<td>0.01906</td>
<td>0.01907</td>
<td>0.01906</td>
</tr>
<tr>
<td>19</td>
<td>0.00572</td>
<td>0.00572</td>
<td>0.00569</td>
<td>0.00571</td>
<td>0.01908</td>
<td>0.01908</td>
<td>0.01916</td>
<td>0.01909</td>
</tr>
<tr>
<td>20</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00570</td>
<td>0.00569</td>
<td>0.01903</td>
<td>0.01904</td>
<td>0.01907</td>
<td>0.01907</td>
</tr>
<tr>
<td>21</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00570</td>
<td>0.00570</td>
<td>0.01907</td>
<td>0.01907</td>
<td>0.01909</td>
<td>0.01907</td>
</tr>
<tr>
<td>22</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00568</td>
<td>0.00569</td>
<td>0.01905</td>
<td>0.01905</td>
<td>0.01904</td>
<td>0.01905</td>
</tr>
<tr>
<td>23</td>
<td>0.00570</td>
<td>0.00570</td>
<td>0.00568</td>
<td>0.00568</td>
<td>0.01902</td>
<td>0.01902</td>
<td>0.01906</td>
<td>0.01902</td>
</tr>
<tr>
<td>24</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00571</td>
<td>0.00570</td>
<td>0.01905</td>
<td>0.01904</td>
<td>0.01912</td>
<td>0.01904</td>
</tr>
<tr>
<td>25</td>
<td>0.00570</td>
<td>0.00570</td>
<td>0.00569</td>
<td>0.00568</td>
<td>0.01903</td>
<td>0.01903</td>
<td>0.01900</td>
<td>0.01903</td>
</tr>
</tbody>
</table>
Figure 1: Coverage versus Bias (difference between true value and estimate) for the two censuses.

+ indicates design i (many sampled clusters per Area)
* indicates design ii (few sampled clusters per Area).

4.2 Comparing spatial microsimulation and ELL-type small area estimation

The spatial microsimulation technique matches survey households to census margins (for small areas) of 1-, 2- or 3-way tables; i.e. it uses only categorical $X$ variables. Hence to compare spatial microsimulation and ELL-type small area estimation, the census design was changed to give four categorical $X$s:

- $X_1$: 2 levels – 0.4:0.6;
- $X_2$: 2 levels – 0.2:0.8;
- $X_3$: 3 levels – 0.2:0.3:0.5;
- $X_4$: 5 levels – 0.1:0.2:0.3:0.3:0.1.
$X$ variables were generated as before by generating multivariate normals with added cluster-level components, then using cut-points to convert to categories.

$Y$ values were generated from a linear model of the form:

$$Y \sim A + B + C + D + A:B + A:C + A:D$$

so that microsimulation using the correct model adjusts survey-derived statistics to the Area-level two-way tables $A \times B$, $A \times C$ and $A \times D$. This seems to be typical of a number of current applications of microsimulation. Household- and cluster-level random errors were added to the $Y$s. Thus in the variable of interest there is both “explained” cluster-level variation (because of the cluster-level variance component in the $X$s) and “unexplained” cluster-level variation (because of the cluster-level random errors added to the $Y$s).

Here we used a superpopulation perspective, simulating 100 censuses and estimating Area-level means of $Y$ by:

- $Y.s$ – direct survey estimate;
- $Y.ms$ – deterministic microsimulation estimate by rescaling the survey households to match the appropriate table summaries from the census data (using iterative proportion fitting);
- $Y.lm$ – fitting a linear model that incorporates the survey weights and using it to predict the Area-level mean;
- $Y.lme$ – fitting a mixed model with random cluster effects to allow for the unexplained cluster-level variation in $Y$, then predicting Area-level means.

To explore the effect of different census structures and sampling designs, we used two different census structures, and four different sampling designs for each. See Table 11a. Some typical results are summarized graphically in Figure 2.
Figure 2: Graph showing variability in the Census values (Y.c) the sample (Y.s) and the two estimates by iterative proportional fitting (Y.ipf) and small area estimation using a linear mixed effects model (Y.lme). These results are for one realization of design A1.

Table 11b gives the average mean squared error across Areas and 100 simulations. All the model-based methods are a vast improvement on the direct estimates from the survey. The results for microsimulation and the linear model are identical, which follows from Section 3 and the choice of the same model for ELL and spatial microsimulation in the simulation, the estimates from these two methods are in all cases identical. The mixed model seemed to have a marginally better performance when there were many Clusters per Area but only a few were sampled.

Table 11c explores the difference between the linear model and mixed model a little further, by calculating confidence intervals and increasing the number of simulations to 1000. The mixed model gives appreciably wider intervals in most cases, leading to better coverage. The simpler linear model worked reasonably well in some situations, but seemed to undercover badly when there were many Clusters per Area but few were sampled.
Table 11a: Design of census and sampling

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Nca</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>NhC</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>nca</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>20</td>
<td>20</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>nhc</td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>25</td>
<td>10</td>
<td>10</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>n</td>
<td>4000</td>
<td>10000</td>
<td>4000</td>
<td>10000</td>
<td>2500</td>
<td>10000</td>
<td>2500</td>
<td>10000</td>
</tr>
</tbody>
</table>

Na = number of Areas in census
Nca = number of Clusters per Area in census
Nha = number of Households per Cluster in census
Nca = number of Clusters sampled per Area in survey
Nha = number of Households sampled per Cluster in survey
n = survey sample size

Table 11b: Average mean square error (100 simulations)

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y.s</td>
<td>2.9837</td>
<td>1.151</td>
<td>1.3682</td>
<td>2.7486</td>
<td>1.2477</td>
<td>0.2914</td>
<td>2.8056</td>
<td>1.1176</td>
</tr>
<tr>
<td>Y.ms</td>
<td>0.0933</td>
<td>0.0839</td>
<td>0.0825</td>
<td>0.0892</td>
<td>0.0338</td>
<td>0.0249</td>
<td>0.0598</td>
<td>0.0332</td>
</tr>
<tr>
<td>Y.lm</td>
<td>0.0933</td>
<td>0.0839</td>
<td>0.0825</td>
<td>0.0892</td>
<td>0.0338</td>
<td>0.0249</td>
<td>0.0598</td>
<td>0.0332</td>
</tr>
<tr>
<td>Y.lme</td>
<td>0.0931</td>
<td>0.0837</td>
<td>0.0824</td>
<td>0.0894</td>
<td>0.0329</td>
<td>0.0248</td>
<td>0.0590</td>
<td>0.0335</td>
</tr>
</tbody>
</table>

Y.s = direct survey estimate
Y.ms = deterministic microsimulation estimate (using iterative proportion fitting)
Y.lm = prediction from fitted linear model
Y.lme = prediction from fitted linear mixed model

Table 11c: Performance of nominal 95% confidence intervals (1000 simulations)

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cov.lm</td>
<td>0.930</td>
<td>0.938</td>
<td>0.944</td>
<td>0.931</td>
<td>0.877</td>
<td>0.937</td>
<td>0.757</td>
<td>0.86</td>
</tr>
<tr>
<td>Cov.lme</td>
<td>0.946</td>
<td>0.948</td>
<td>0.945</td>
<td>0.943</td>
<td>0.953</td>
<td>0.954</td>
<td>0.944</td>
<td>0.954</td>
</tr>
<tr>
<td>AvLen.lm</td>
<td>1.139</td>
<td>1.135</td>
<td>1.158</td>
<td>1.127</td>
<td>0.588</td>
<td>0.574</td>
<td>0.581</td>
<td>0.564</td>
</tr>
<tr>
<td>AvLen.lme</td>
<td>1.189</td>
<td>1.147</td>
<td>1.147</td>
<td>1.186</td>
<td>0.756</td>
<td>0.617</td>
<td>0.973</td>
<td>0.755</td>
</tr>
</tbody>
</table>

Cov.lm = coverage of confidence interval from linear model
Cov.lme = coverage of confidence interval from linear mixed model
AvLen.lm = average length of confidence interval from linear model
AvLen.lme = average length of confidence interval from linear mixed model

Several conclusions follow from this component of the simulations:
• Deterministic microsimulation (via IPF) is equivalent to predicting from the equivalent linear model (fitted by WLS with weights determined by IPF) and hence the same as SAE. This and Section 3 provide the link between spatial microsimulation and ELL-type small area estimation – they are essentially the same technique.
• An advantage of an explicit model specification (e.g. the linear model in the case of ELL) is the provision of standard errors (that can include random cluster effects). It is thus possible to get estimated standard errors from spatial microsimulation, by explicitly specifying the underlying model (which also has the advantage that it no longer need be set a priori, but can be tested statistically to decide what variables and interactions should be included);
• In addition, whole linear model machinery for model selection and diagnostics can be made available for spatial microsimulation as well as small area estimation;
• It is possible to extend spatial microsimulation models to include numerical covariates, not just categorical variables in the same way as for small area estimation (see Noble, Haslett and Arnold, 2003; Haslett, Noble and Arnold 2006), although this would need record-level census data;
• Stochastic microsimulation is linked to deterministic (hence small area estimation) by the idea of conditional expectation, where selection of replicated observations (for probability reweighted spatial microsimulation) and reweighting of observations (for deterministic reweighted spatial microsimulation) is simply selection or reweighting, conditioned on an underlying statistical model.

4.3 Effect of Model Uncertainty

In assessing model uncertainty, we compared area-level predictions from the “true model” (i.e. known $X$s but unknown $\beta$s) with those from a “selected model” ($X$s chosen from stepwise regression). The interest here was whether estimating the model had an appreciable affect on the predictions from it than were later aggregated into area-level estimates. The fundamental question was how robust area-level estimates are to incorrect choice of effects from the true model, and to variations in estimates even given the true model. Using the same census/sampling designs as before, Area-level estimates were produced by estimating, using the weighted least squares method, the true model $Y \sim A + B + C + D + A:B + A:C + A:D$ and a model chosen to minimize the AIC criterion.

Surprisingly and encouragingly, model uncertainty did not appreciably degrade the predictions in either situation (as assessed from average MSE across simulations).
See Table 12a-b. Table 12a shows the typical variability in model choice from 100 simulations, using a superpopulation approach. Table 12b compares the performances of the true and AIC-selected models.
Table 12a: Distribution of chosen model in 100 simulations (from design B1)

<table>
<thead>
<tr>
<th>Model</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+B+C+D+A:B+A:C+A:D*</td>
<td>21</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C</td>
<td>19</td>
</tr>
<tr>
<td>A+B+C+D+A:C+A:D</td>
<td>15</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+A:D+B:C</td>
<td>6</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+B:C</td>
<td>5</td>
</tr>
<tr>
<td>A+B+C+D+A:C+A:D+B:C</td>
<td>5</td>
</tr>
<tr>
<td>A+B+C+D+A:C</td>
<td>4</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+B:D</td>
<td>3</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:D</td>
<td>3</td>
</tr>
<tr>
<td>A+B+C+D+A:C+B:C</td>
<td>3</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+A:D+B:D</td>
<td>2</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+B:D+B:C</td>
<td>2</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+A:D+C:D</td>
<td>2</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+C:D</td>
<td>2</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+A:D+B+C</td>
<td>1</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+B+C+C:D</td>
<td>1</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+B:D+B:C</td>
<td>1</td>
</tr>
<tr>
<td>A+B+C+D+A:B+A:C+B:D</td>
<td>1</td>
</tr>
<tr>
<td>A+B+C+D+A:B+B:C</td>
<td>1</td>
</tr>
<tr>
<td>A+B+C+D+A:C+A:D+C:D</td>
<td>1</td>
</tr>
<tr>
<td>A+B+C+D+A:C+B:D+B:C</td>
<td>1</td>
</tr>
<tr>
<td>A+B+C+D+A:C+B:D+B:C+B:C</td>
<td>1</td>
</tr>
</tbody>
</table>

* indicates the true model

Table 12b: Comparison of performance of estimators using true and chosen models (1000 simulations)

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE.tr</td>
<td>0.0933</td>
<td>0.0839</td>
<td>0.0825</td>
<td>0.0892</td>
<td>0.0338</td>
<td>0.0249</td>
<td>0.0598</td>
<td>0.0332</td>
</tr>
<tr>
<td>MSE.cht</td>
<td>0.0934</td>
<td>0.0839</td>
<td>0.0825</td>
<td>0.0892</td>
<td>0.0337</td>
<td>0.0249</td>
<td>0.0599</td>
<td>0.0332</td>
</tr>
<tr>
<td>Cov.tr</td>
<td>0.930</td>
<td>0.938</td>
<td>0.944</td>
<td>0.931</td>
<td>0.877</td>
<td>0.937</td>
<td>0.757</td>
<td>0.860</td>
</tr>
<tr>
<td>Cov.ch</td>
<td>0.936</td>
<td>0.957</td>
<td>0.956</td>
<td>0.945</td>
<td>0.910</td>
<td>0.946</td>
<td>0.781</td>
<td>0.874</td>
</tr>
<tr>
<td>Len.tr</td>
<td>1.1390</td>
<td>1.1350</td>
<td>1.1580</td>
<td>1.1270</td>
<td>0.5880</td>
<td>0.5740</td>
<td>0.5810</td>
<td>0.5640</td>
</tr>
<tr>
<td>Len.ch</td>
<td>1.1370</td>
<td>1.1360</td>
<td>1.1620</td>
<td>1.1230</td>
<td>0.5880</td>
<td>0.5740</td>
<td>0.5810</td>
<td>0.5670</td>
</tr>
</tbody>
</table>

MSE.tr average mean squared error using true model
MSE.ch average mean squared error using model chosen by AIC
Cov.tr average coverage of nominal 95% confidence interval using true model
Cov.ch average coverage of nominal 95% confidence interval using model chosen by AIC
Len.tr average length of nominal 95% confidence interval using true model
Len.ch average length of nominal 95% confidence interval using model chosen by AIC
More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics
4.4 Updating via projections

This component of the simulation was designed to address the future scenario situation encountered in the use of spatial microsimulation. It is also relevant to small area estimation where there is a prediction of the census values contained in X to a future period (e.g. via a time series model) before the small area method is applied to the projected X for the census, and either new sampled values for Y or projections for them as well.

We consequently assumed a unit-level model for updating the census values (X: Y), but with errors in the updated values. To do this we generated a census as before, then added noise to simulate the effect of projecting rather than knowing the census values.

The conclusion was that adding noise at cluster- and household-level has little effect on area-level estimates when there is a reasonably large number of clusters in each area, provide the noise has zero mean (or equivalently the projection model used for the census is unbiased). See Table 13 and Figure 3. Noise, representing errors in the updating model, was added to the Y values as a proportion of the variances used in the simulated census values. The increase in mean square error resulting from the addition of noise was measured and found to be quite small even when the noise variance was up to 10% of the error variance for Y.

“Fixed-effect” noise, e.g. changing the βs, would obviously have a significant effect, since this would mean that the projections would now not be unbiased, and where this unbiasedness carried over to area level, it would not average out in particular small areas, and would consequently introduce unacceptable inaccuracies into the techniques..

Table 13: Effect on estimation of accuracy of additional noise added to the variable of interest, using census/sample design B2

<table>
<thead>
<tr>
<th>NVf</th>
<th>Δ MSE (×10^{-4})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.4938</td>
</tr>
<tr>
<td>0.03</td>
<td>0.7337</td>
</tr>
<tr>
<td>0.01</td>
<td>0.4726</td>
</tr>
<tr>
<td>0.003</td>
<td>0.3546</td>
</tr>
<tr>
<td>0.001</td>
<td>0.3069</td>
</tr>
<tr>
<td>0.0003</td>
<td>0.2819</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.2703</td>
</tr>
</tbody>
</table>

NVf – added noise as a fraction of the overall variances
Δ MSE – increase in average mean squared error resulting from the added noise
Figure 3: Increase in MSE resulting from errors in updating or projection for different values of the Noise Variance Factor

4.5 Mass Imputation

The final component of the simulations considered mass imputation. Now there may be missingness in $X$ as well as $Y$, so that values of both are imputed. Nevertheless, there are links to the earlier components of the simulation study, because errors in $X$ are considered for spatial microsimulation and ELL-type small area estimation in Section 4.4, and Section 3 has specified the underlying similarities in terms of model for all three techniques.

We still take $Y$ as the “variable of interest” in the complete data set (“true census”). Now, in the conceptual framework, some values of $Y$ are imputed from their conditional distribution given a subset of the $X$s. This can be simulated by adding some noise to perturb the $X$ values (as in Section 4.4) – the resulting changes to the categories are then equivalent to the errors made when a missing $X$ category is imputed based on the observed data. The amount of noise used in the perturbation corresponds to the number of errors made in imputing the $X$s, which corresponds in turn to how much missingness there is in the $X$ data.

If the true model is:
More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics

\[ Y_{ij} = X_{ij} \beta + c_i + e_{ij} \]

the imputed value can be thought of as

\[ Y_{ij} = X^*_{ij} \beta + c_i' + e_{ij}' \]

where \( X^*_{ij} \) represents the imputed (or perturbed) \( X \) data and \( c_i' + e_{ij}' \) is the cluster- and unit-level random effects respectively from the unit used to impute. These will not in general be the correct ones; the cluster effect in particular will probably be wrong, and different for the different units imputed in the same cluster. Thus for the area-level average, \( c_i \) will be replaced by the average of a number of imputed \( c_i' \); and this will be a source of bias in a fixed-population framework. If we also have, as assumed here, cluster-level variations in the \( X \) data, the above argument extends, and \( X^*_{ij} \) will also be a source of bias.

Using the same census/sample designs as previously, we first perturbed the \( X \) data by adding random noise to the underlying continuous variables, then recategorising. The variance matrix of this multivariate noise was proportional to the variance used to create the original underlying continuous variables; we used fractions of 0.001, 0.01 and 0.1. The amount of perturbation in the categorical \( X \) values is shown in Table 14a: the proportion of households with unchanged \( X \) values ranged from 94\% for 0.001 to 51\% for 0.1. New \( Y \) values were imputed from the perturbed \( X \) values as described above. The “sample” here corresponds to those census records for which complete information is available; for these no imputation is required so their \( Y \) values are left unchanged.

Using a fixed census approach, we produced 100 perturbations, corresponding to 100 multiple mass imputations of the missing data. Table 14b shows the distribution across Areas of the true Area-level means \( Y.c \) for each design, and the performances of the mass imputation estimates \( Y.i \) as follows:

- for each Area the true value \( Y.c \) was subtracted from the mean of \( Y.i \) in the 100 imputations to give the bias in that Area; this was then summarized by the standard deviation across Areas;
- for each Area the variability in the 100 imputations of \( Y.i \) was summarized by its standard deviation; this was then averaged across Areas.

Some typical results, for design A1, are summarized graphically in Figure 4.

Surprisingly perhaps, the amount of zero mean noise added, corresponding to the amount of missing \( X \) data, had only a small effect on the accuracy as measured by sd(bias). For the designs we used it appears that the most important factor is the variability in the \( Y \) values themselves; since mass imputation is essentially replacing unobserved \( Y \) values with observed ones, the factors affecting the noise introduced by this process are the same ones governing the variability in the Area-level estimates \( Y.c \).
Figure 4: Graphs showing the Area-level Census means ($Y_c$), and the means ($Y_{bar.i}$) and standard deviations ($Isd$) of the corresponding multiple mass imputation estimates (from design A1).

Table 14a: Proportions of households with a given number of changes in the $X$ variables resulting from added noise

<table>
<thead>
<tr>
<th>NVf</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.936620</td>
<td>0.061841</td>
<td>0.001525</td>
<td>0.000014</td>
<td>0.000000</td>
</tr>
<tr>
<td>0.01</td>
<td>0.809436</td>
<td>0.176311</td>
<td>0.013776</td>
<td>0.000473</td>
<td>0.000004</td>
</tr>
<tr>
<td>0.1</td>
<td>0.509130</td>
<td>0.363293</td>
<td>0.112863</td>
<td>0.014103</td>
<td>0.000611</td>
</tr>
</tbody>
</table>

$NVf$ – added noise as a fraction of the overall variance
Table 14b: Performance of mass imputation estimator for fixed census (designs as in Table 10a).

<table>
<thead>
<tr>
<th></th>
<th>mean(Y.c)</th>
<th>sd(Y.c)</th>
<th>sd(bias)</th>
<th>mean(lsd)</th>
<th>sd(bias)</th>
<th>mean(lsd)</th>
<th>sd(bias)</th>
<th>mean(lsd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>4.140</td>
<td>0.318</td>
<td>0.286</td>
<td>0.283</td>
<td>0.291</td>
<td>0.287</td>
<td>0.320</td>
<td>0.296</td>
</tr>
<tr>
<td>A2</td>
<td>4.140</td>
<td>0.318</td>
<td>0.280</td>
<td>0.280</td>
<td>0.289</td>
<td>0.285</td>
<td>0.313</td>
<td>0.295</td>
</tr>
<tr>
<td>A3</td>
<td>4.140</td>
<td>0.318</td>
<td>0.286</td>
<td>0.282</td>
<td>0.297</td>
<td>0.287</td>
<td>0.330</td>
<td>0.295</td>
</tr>
<tr>
<td>A4</td>
<td>4.140</td>
<td>0.318</td>
<td>0.279</td>
<td>0.281</td>
<td>0.285</td>
<td>0.285</td>
<td>0.306</td>
<td>0.294</td>
</tr>
<tr>
<td>B1</td>
<td>4.130</td>
<td>0.148</td>
<td>0.125</td>
<td>0.141</td>
<td>0.132</td>
<td>0.144</td>
<td>0.143</td>
<td>0.149</td>
</tr>
<tr>
<td>B2</td>
<td>4.130</td>
<td>0.148</td>
<td>0.141</td>
<td>0.142</td>
<td>0.144</td>
<td>0.144</td>
<td>0.153</td>
<td>0.149</td>
</tr>
<tr>
<td>B3</td>
<td>4.130</td>
<td>0.148</td>
<td>0.126</td>
<td>0.141</td>
<td>0.140</td>
<td>0.144</td>
<td>0.165</td>
<td>0.149</td>
</tr>
<tr>
<td>B4</td>
<td>4.130</td>
<td>0.148</td>
<td>0.140</td>
<td>0.142</td>
<td>0.150</td>
<td>0.144</td>
<td>0.171</td>
<td>0.150</td>
</tr>
</tbody>
</table>

NVf – added noise as a fraction of the overall variance
Y.c – true value of Area-level mean of Y
bias – mean of imputed values from 100 imputations minus Y.c
Isd – standard deviation of imputed values from 100 imputations
5. Conclusions

5.1 Simulations

The simulations have indicated some underlying similarities between spatial microsimulation, small area estimation using the ELL method, and mass imputation, in line with the structural similarities from a more theoretical perspective apparent from Section 3. These similarities do not however extend to categorising all three methods together in terms of effectiveness, since in practice they tend to be used in different ways.

Mass imputation and spatial microsimulation have tended to be used with implicit, and perhaps too often untested, statistical models as their basis, with variables included in them being decided a priori or on the basis of the (often rather limited) number of variables available. In the case of spatial microsimulation, the information available has taken the form of various census cross-tabulations and variables, which implicitly define which effects are included in a loglinear model. For mass imputation the situation is even more opaque, as even the effect of the best imputation methods based on nearest neighbour techniques (even if made explicit) do not necessarily lead to a clearly specified statistical model.

Small area estimation, whether using ELL or not, has a longer history of specifying, fitting and testing explicit statistical models, and it is recommended that such specifying, fitting and testing provide a focus for further research in spatial microsimulation and mass imputation.

All three methods, under well-specified and fitted models without bias, are capable of producing reliable estimates, even where projections of data are required.

5.2.1 New Zealand - government projects: general comments on case studies

5.2.1 IBULDD imputation

The simulation study in the main did not consider skewed variables, which occur frequently in business survey and census data. In principle such variables in the LBD might be categorised (as for the simulation study), as this would allow technical adjustments to be made to compensate for skewness. In practice however, the open question is whether this would limit utility of the imputed dataset. Use of longitudinal data and nearest neighbour methods would generally provide more accurate imputation than using hot deck, and both these two methods (since they have models more correlated to what is missing) would provide better results than the imputation method used for mass imputation in the simulation study.

However, the LBD involves creation of only one dataset for rather than several using multiple imputation. Although this was a forced choice for computing time reasons, this lack of multiple imputed values for each respondent means that it is not possible to estimate standard errors, even assuming the implicit and untested imputation model is correct. If imputation has produced no bias, averages for subgroups – especially where missingness of both records and items is comparatively low – may be satisfactory, but caution is warranted if the imputed dataset were then used to estimate more complicated analysis or statistics. Particular caution would be needed for statistics derived from variables collected only in the sample surveys if these were imputed, since then missingness of complete records is high since not all businesses are surveyed, even though the response rate for the survey itself is also high. The imputation model currently used for the tax-based
variables is comparatively simple (and provided a priori on the basis of the limited number of variables available that have a small number of missing values, rather than on a model tested for statistical significance of its component parts).

5.2.2 Small area estimation in outcome evaluation of Youth Transition Services

The YTS study involves aspects of both spatial microsimulation and small area estimation. For both techniques, accuracy and lack of bias depends on the choice of statistical model for the time series projections. The main difference from the techniques for small area estimation is that the variables used for prediction also require projection, so it is not simply the dependent but also the independent variables in models that require projection and are consequently subject to error. In one sense then, it is all the data, not just the variable of particular interest that must be projected for modelling. This of course introduces additional error in area-based estimates in comparison with the more usual small area techniques.

The simulation results in Section 4 indicate that whether such a method will work depends on the quality of the projections and particularly on whether they are subject to bias. Generally the simulations indicate that, in the absence of bias, provided the aggregation to form estimates of proportions involves a sufficiently large number of youths (which will be indicated by the estimated standard errors conditional on the model used), projection causes few additional complications in comparison with the more usual small area estimation methods. Bias however is a possible complication – what is crucial is that, whatever its faults, the projection or prediction model used for the modelling does not introduce appreciable bias, since this cannot be compensated for by aggregation even into comparatively large areas.

5.2.3 Imputing Victim Forms from the New Zealand Crime and Safety Survey:

Imputation for NZCASS 2006 includes multiple imputation, in which each missing response is imputed a number of times, since this allows better assessment of the imputation procedure, at least in terms of its variance conditional on the model used for imputation. While commendable, this is not the only aspect of imputation which is important however.

Usually imputation techniques applied to survey data, whether multiple or not, are applied to a small percentage of the data. In the NZCASS however, the percentage of non-completed victim forms is very high at 67%, partly as a result of a deliberate sample design choice intended to limit interview time for those interviewees with more than three victimisation incidents. The survey imputation thus has a strong mass imputation flavour, since a high percentage of responses are imputed albeit limited to a few variables. The high imputation rate for responses, particularly given the very simple model used for number of incident imputation, warrants further research focussed on the NZCASS data itself, to determine whether mass imputation has had any of the deleterious effects that have been found (for example by Statistics Canada and Statistics Netherlands) in some other situations. The simulation studies in Section 4 indicate that adequacy of mass imputation hinges on lack of bias and sufficient accuracy in the imputation model, and that rigorous checking of statistical models used in imputation is necessary.

5.3 General conclusions: links and comparison of small area estimation (ELL), spatial microsimulation, and mass imputation

In practice, despite an underlying conceptual and theoretical similarity and that all are methods for ‘completing’ databases, there are both similarities and differences
between spatial microsimulation, mass imputation, and small area estimation using the ELL method.

For all three techniques, the generally common intention (either as an interim or final output) is to produce a dataset (or datasets) which is rectangular without missing values, which is essentially a pseudo-census created by substitution of missing information using an implicit or explicit statistical model. The attraction of this approach is that, superficially at least, the pseudo-census can substitute for the unavailable census, though caution is clearly warranted for complex statistics required accurately by small area. The three methods can be considered as variations on a theme, under the unifying framework outlined in Section 3. This unification has a number of consequences not limited to the theoretical. For example, the framework of Section 3 and the results of the simulations in Section 4 make it very clear that:

- regardless of the method used, there are major benefits in use of an explicit rather than implicit statistical model when imputing.
- the structure of the underlying statistical model (e.g. linear or non-linear, with or without random effects) needs to be determined on strong theoretical grounds.
- the model needs to be fitted and tested, and should explain a substantial part of the variation in records not requiring imputation, so that inference to incomplete records can be properly justified.
- imputing residuals only, rather than entire variables, has major advantages in terms of utility of the pseudo-census(es), since it is better able to control bias especially where average values (for example for areas) are required.
- estimation of standard errors conditional on the fitted model is possible not only for ELL-type small area estimation (where it is routine), but also by a simple theoretical extension under the common framework of Section 3 to spatial microsimulation and mass imputation.

ELL-type small area estimation currently has the advantage over the other techniques of an explicit statistical model which is not only specified, but also fitted and tested. It is also able to provide estimated standard errors for its area-level averages.

One issue that deserves further discussion about ELL, however, is that ELL has smaller (sometimes much smaller) estimated standard errors than many of the small area methods of Rao (2003) and Longford (2005). This is not a direct result of the modelling, since all these small area methods first fit models to the survey data and test them, but instead due to the levels at which the error structure of models are fitted and to differences in the way available census data is integrated into the small area estimates.

ELL does not include a small-area-level error in its models. Instead it includes cluster (within area) and household (within cluster) error terms in linear models that may contain a comparatively large number of predictor variables, fitted separately to each survey stratum. While this strategy limits omitted variables (and hence the need for a small-area-level error term), it runs the counter-risk of overfitting models, since the number of candidate variables (including interactions) is often close to the number of observations within strata. Not all ELL-type models run similar risk of overfitting however (see, for example, the more restricted models of Haslett and Jones, 2005a and 2005b, Jones and Haslett, 2003, and Jones, Haslett and Parajuli, 2006, which are fitted to the complete sample, rather than by stratum, using a limited number of candidates). One instance where ELL-type models with an added small-area-level error term and a limited number of predictor variables (both fitted and used as candidates) is Jones, Haslett and Parajuli (2006), where the small-area-level error has negligible effect on the standard error estimates of poverty, so an ELL-type approach seems useful in at least some circumstances. The required estimation of
the small area-level and the cluster-level variances in such models for complex sample survey data with stratification, clustering and unequal weights is however complex, especially since there is only one error prediction per small area, and there are consequently risks of allocating the error variance to the wrong level in the hierarchy, especially since estimation of the variance of the variance components for complex survey data is both a difficult and relatively unexplored research area. Note that if area-level-errors are comparatively large, whether they are fitted explicitly (as by Rao, 2003) or not (as in ELL), Rao-type estimated standard errors are essentially the correct ones. In practice, the situation is complicated however because many Rao-type models do not include a cluster-level error term, and standard errors for small area estimates can be over-estimated if cluster-level variability is instead counted toward small-area-level prediction errors.

Many of the models fitted to survey data by Rao (2003) and by Longford (2005) do not use census data at all or only census averages by small area, so their accuracy is determined by the fitted model (including any small-area-level errors) and limited (even if indirectly) by survey sample size. In comparison, for ELL every census observation can be and is predicted (multiple times, under the model that has been fitted to the survey data), based on the regression and added imputed residuals. ELL might be viewed as involving mass imputation (J. N. K. Rao - personal communication) but if so it is mass imputation of residuals only and usually for one variable (rather than a range of variables, as is more usual in mass imputation) under a comparatively well specified and tested model where the bulk of the predictor is fully based on a model, all predictor variables are available for all census observations and have been matched against their survey equivalents both in definition and in value, and where taken over small areas the expected value of the residual under the model is zero (which is a property that can be tested). So this is not mass imputation in the usual sense, since it is only for univariate residuals not multivariate observations, and, whatever it is called, it is comparatively small part of the prediction, especially after individual or household pseudo-census observations are aggregated to small area level. Nevertheless, ELL tends to produce much smaller estimated standard errors for the same small areas than mixed model methods that include small-area-level errors and do not integrate (or do not so fully integrate) known census information on key predictor variables. One explicit reason for the difference in standard errors is that the contribution of the estimated variance components in the ELL model (which are themselves similar if not identical to those estimated in a survey based small area method) are divided by the population size (e.g. number of clusters, or households) for estimated standard errors for small areas from ELL, rather than by the corresponding sample size as required when census predictions are not available or used. It is this stronger model assumption in ELL (which can be tested as part of the model fitting and only applies to the residuals anyway) together with the assumption that the model contains enough predictor variables that the small-area-level error is negligible, that gives ELL its markedly lower estimated standard errors than small area techniques not incorporating census data so directly. The point (as the simulations in Section 4 show) is not that ELL is wrong, but that it requires stronger assumptions, which must be rigorously tested as part of the model fitting process. Related issues are discussed in Haslett and Jones (2005a, 2005b), Jones and Haslett (2003), and Jones, Haslett and Parajuli (2006).

Unlike the other techniques, the subclass of deterministically reweighted spatial microsimulation methods do not necessarily produce a pseudo-census. If its weights for survey observations were integer, creating a pseudo-census using this method would involve only one additional step: simple replication of observations, so that the number of replications equalled each survey observation’s weight. More often however, weights from deterministic spatial microsimulation are non-integer. Deterministically reweighted spatial microsimulation (unlike its probabilistic reweighted relation) is a genuine reweighting technique, based on use of IPF or its...
equivalent to calibrate to various census and other tables. In fact, deterministically reweighted spatial microsimulation is substantially different from the Dutch “virtual census” only in the methods used for choosing calibration variables (which are rather more implicit for deterministically reweighted spatial microsimulation) and in the possibility of using observations from outside the small area in spatial microsimulation.

Spatial microsimulation and mass imputation may impute complete or near complete records. In practice though, for both techniques at least some variables are available for all records, usually as aggregate counts by small area from census or administrative data sources, and these are used to inform an implicit imputation model, which is usually decided a priori rather than tested statistically before adoption. For mass imputation, the technique used (e.g. nearest neighbour imputation) may be set, but the model is still usually implicit. One intriguing possibility is that, where the model is implicit, its performance may be testable using the techniques used in data mining, e.g. cross-validation.

Small area estimation using the ELL method does not impute complete records, but instead usually imputes only one variable at a time under a mixed linear model. For ELL it is only the residuals from the random components in the mixed model that are imputed; most of the structure in the imputation model is contained in a regression equation.

In summary, even though all three methods, spatial microsimulation, mass imputation, and small area estimation via ELL, show strong structural similarities, this does not mean that deficiencies in one are necessarily deficiencies in another. Of the three methods, the underlying model used for imputation is explicit only in ELL, and consequently ELL can really be considered the best of the techniques given sound model fitting and testing. Adding such explicit fitting and testing to spatial microsimulation and mass imputation would improve both techniques, without being theoretically burdensome. It would also allow their accuracy to be better assessed, as is already done for small area estimation using ELL, by creating multiple pseudo-censuses and estimating standard errors under the specified and tested model. From these points of view, rapid improvements to the accuracy and assessment of accuracy of spatial microsimulation models in particular should be plausible under the theoretical framework for all three techniques outlined in Section 3.
More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics

References


More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics


More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics


More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics


More for less? Using statistical modelling to combine data for more detailed, less expensive Official Statistics


Appendix 1

Guidelines for using mass imputation, spatial microsimulation, and ELL-type small area estimation

1. Consider carefully what the aims of the research are, in detail, and the range and quality of the data sources available. Be aware that more complicated and detailed aims will be more difficult to achieve than simpler less detailed ones.

2. Establish whether these data sources are compatible (e.g. same or similar time period, same definitions of variables, level of disaggregation), available and clean.

3. Assess what proportion of the complete (or full) data required is available from each of these sources, both in terms of the population under study and the variables available.

4. Consider carefully the structure of the full dataset you are trying to create, in terms of variables and level and numbers of observations.

5. Reassess what proportion of the data is missing on key variables, and establish the pattern of missingness (e.g. random from a sample survey, missing due to purposive administrative procedures).

6. Consider the range of types of statistical models (e.g. linear models with or without random effects, generalized linear models) that seem suitable for modelling, and make a preliminary choice.

7. List the candidate variable available to predict the key variables.

8. Clean and check all data to be used, and fit preliminary models to the that part of the data for which complete data is available using these candidate variables, and assess how much of the variation they explain. Do this for the time period for which data is available.

9. If projections are required, or future scenarios are needed, consider what models are suitable for projection and whether required variables are available, and whether a statistical time series model can be built from the available data.

10. Given steps 8 and 9 above, consider what prediction errors are likely to be at unit (e.g. individual) level

11. Given step 10, think about what aggregations of data you will need to use to get sufficient accuracy to be useful.

12. Go back to your aims, and ask whether the project is feasible given the likely accuracy.

13. If so, set up initial simulations, whether small area estimation, spatial microsimulation or mass imputation.

14. Run simulations many times, for every scenario if relevant. (This is multiple imputation).

15. Use your multiple simulation results to assess accuracy for key variables (conditional on the simulation model used).

16. Recheck data for errors, and reassess what types of statistical models are most appropriate.

17. Begin a more thorough search for models that fit well, checking them and their component parts for statistical significance. Loop back to step 14 as many times as necessary.

18. Reassess whether you can meet your aims.

19. Consider carefully at what level of aggregation (e.g. by what size of area) you can get results that are sufficiently accurate.

20. Produce results at this level or higher (i.e. more aggregated), and relate the results at this level back to your aims.
Appendix 2

Guidelines for designing surveys for mass imputation, spatial microsimulation, and ELL-type small area estimation

Most national sample surveys, unless a comprehensive list is available, are based on regional stratification and sampling clusters within strata. At the finer level there is usually subsampling within clusters and perhaps within households.

The core underlying characteristic of the small area estimation using census predictions, spatial microsimulation, and mass imputation is that a statistical model is implicitly or explicitly fitted to the available survey data. To be adequate, the models fitted or used a priori have to recognise the underlying survey structure and incorporate this into predictions at the finest level possible. The stratification and clustering then, being at the highest level of the design hierarchy are the most important.

For social surveys, fitted statistical models will then test, and likely incorporate, stratum and cluster effects, as well as exogenous variables such as household type. Since interactions between stratum and exogenous variables are reasonably common in models, having a sufficient sample size within each stratum to properly do the fitting is important. Unless designs are of the two units per stratum type, this is not usually an issue. What tends to be more crucial is the selection method for clusters and number of clusters sampled per stratum. Clusters are often selected with probability proportional to size (pps). While then sampling a fixed number of households per sampled cluster then gives a self-weighting sample of households, at least within strata, different cluster characteristics (e.g. their size) can cause problems when models need to incorporate and estimate cluster-level variances. The estimation of these variances is further complicated if the sampling scheme only samples a few clusters within strata, especially if the differences in cluster variance between strata need testing as part of the model and/or there are underlying differences related to size between clusters.

In business surveys the situation is different. For area based surveys of business the conclusions are rather similar. For samples selected from a list, for example by pps, because differences in size of businesses can be marked, the sample needs to be strongly weighted (inversely proportional to size) to produce unbiased estimates. The strong weighting can complicate model fitting to the survey data. It can also mean that subgroups of the population important to the modelling but not to the overall national figures can be under-represented were the primary consideration.

The general conclusion is that for modelling, designs which spread their choice of sample across the population are preferable. The situation has parallels to the trade-off necessary when direct subpopulation estimates as well as national totals or averages are required from a sample. Designing for sound direct subpopulation estimates and estimates based on statistical models have similar requirements, and meeting these requirements will to some extent always lower the accuracy of national-level estimates relative those possible to a more complicated design.

It is consequently important to decide the design of a survey based on the entire range of uses to which it will be put, and not to expect good statistical modelling will be possible simply as an add-on to a design constructed to meet other optimality criteria.