INTERNATIONAL JOURNAL OF MICROSIMULATION (2010) 3(2) 3-22

Methodological Issues in Spatial Microsimulation Modelling for Small Area Estimation

Azizur Rahman*, Ann Harding, Robert Tanton and Shuangzhe Liu
National Centre for Social and Economic Modelling (NATSEM), University of Canberra, ACT 2601, Australia; *email: Azizur.Rahman@natsem.canberra.edu.au

ABSTRACT: In this paper, some vital methodological issues of spatial microsimulation modelling for small area estimation have been addressed, with a particular emphasis given to the reweighting techniques. Most of the review articles in small area estimation have highlighted methodologies based on various statistical models and theories. However, spatial microsimulation modelling is emerging as a very useful alternative means of small area estimation. Our findings demonstrate that spatial microsimulation models are robust and have advantages over other type of models used for small area estimation. The technique uses different methodologies typically based on geographic models and various economic theories. In contrast to statistical model-based approaches, the spatial microsimulation model-based approaches can operate through reweighting techniques such as GREGWT and combinatorial optimization. A comparison between reweighting techniques reveals that they are using quite different iterative algorithms and that their properties also vary. The study also points out a new method for spatial microsimulation modelling

Keywords: Bayesian prediction approach; combinatorial optimisation; GREGWT; microdata; small area estimation; spatial microsimulation

1. INTRODUCTION

Small area estimation is the method of estimating reliable statistics at the small geographical area or a spatial micropopulation unit. Reliable statistics of interest at small area levels cannot be ordinarily and directly produced, due to certain limitations of the available data. For instance, a suitable sample that contains enough representative observations is typically not available for all small areas from national level survey data. A basic problem with national or state level surveys is that they are not designed for efficient estimation for small areas (Heady et al., 2003). In practice, small area estimates from these national sample surveys are statistically unreliable, due to sample observations being insufficient, or in many cases non-existent, where the domain of interest may fall outside the sample domains (Tanton, 2007). Given typical time and money constraints, it is usually impossible to conduct a sufficiently comprehensive survey to get enough data from every small area we are interested in.

Nowadays indirect modelling approaches of small area estimation, such as spatial microsimulation models (SMMs), have received much attention, due to their usefulness and the increasing demand for reliable small area statistics from both private and public level organisations. In these approaches, one uses data from similar domains to estimate the statistics in a particular small area of interest, and this 'borrowing of strength' is justified by assuming a model that relates the small area statistics (Meeden, 2003). Typically, indirect small area estimation is the process of using statistical models and/or geographic models to link survey outcome or response variables to a set of predictor variables known for small areas, in order to predict small area estimates. As a result of inadequate sample observations in small geographic areas, the conventional area-specific direct estimates may not provide enough statistical precision. In such a situation, an indirect model-based method can produce better results.

Most of the review articles in small area estimation have highlighted the methodologies, which are fully based on various statistical models and theories (for example, Ghosh and Rao, 1994; Rao, 1999; Pfeffermann, 2002; Rao, 2002; Rao, 2003a). However another type of technique called 'spatial microsimulation modelling' has been used in providing small area estimates during the last decade (for instance, Williamson et al., 1998; Ballas et al., 2003; Taylor et al., 2004; Brown and Harding, 2005; Chin et al., 2005; Ballas et al., 2006; Chin and Harding, 2006; Cullinan et al., 2006; Lymer et al., 2006; Anderson, 2007; Chin and Harding, 2007; King, 2007; Tanton, 2007) . The SMMs are based on geographic and economic theories, and their methodologies are quite different from other statistical approaches. Although these approaches are frequently used in social and economic analysis, and seem to be a robust and rational indirect modelling tool, the mechanisms behind them are not always well documented. Also there are some important methodological issues where more research should improve the performance of SMMs and help in the validation of their estimates.

This paper provides a brief synopsis of the overall methodologies for small area estimation and explicitly addresses some vital methodological issues of spatial microsimulation modelling, with a particular emphasis given to the reweighting techniques. It also proposes a new approach in the SMM methodologies. An application of the generalised regression based reweighting technique discussed in this article is studied by Tanton and Vidyattama under distinct features of the applicable data. This contribution is part of this special issue as well.

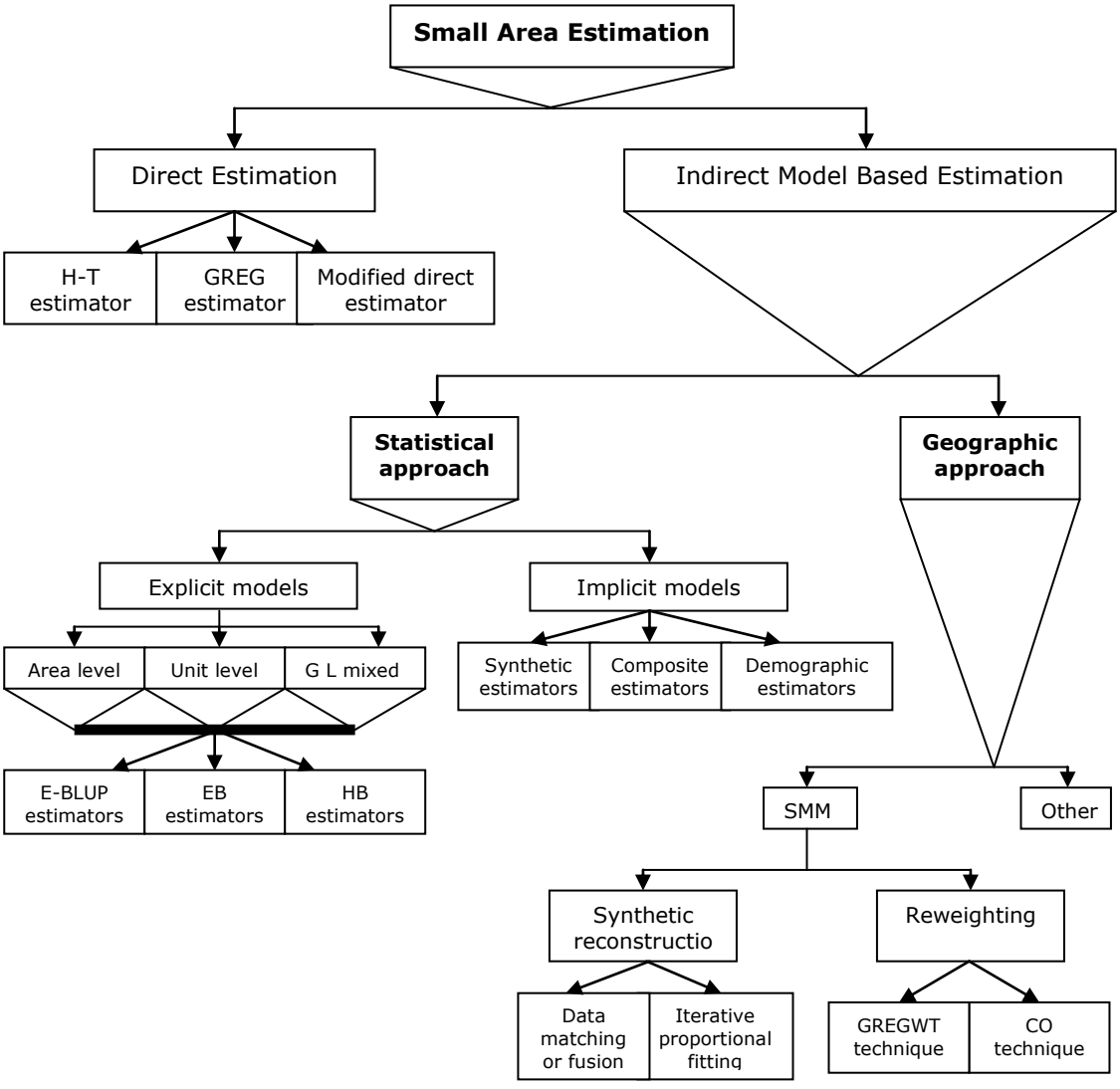
There are 5 sections within this paper. In Section 2, a diagramic representation of the overall methodologies in small area estimation is provided with a synopsis of various direct and indirect statistical model based estimations, and highlights of spatial microsimulation modelling. In Section 3, some vital methodological issues in spatial microsimulation modelling are addressed, which include theories and numerical solutions of different reweighting techniques. In Section 4, a comparison between two reweighting techniques is presented with a new methodology for generating small area microdata. Finally, conclusions are given in Section 5.

2. METHODS OF SMALL AREA ESTIMATION: AN OVERALL VIEW

A diagramic representation of the overall methodologies for small area estimation is depicted in Figure 1. Traditionally there are two types of small area estimation – direct and indirect estimation. Direct small area estimation is based on survey design and includes three estimators called the Horvitz-Thompson estimator, GREG estimator and modified direct estimator. On the other hand, indirect approaches of small area estimation can be divided into two classes – statistical and geographic approaches. The statistical approach is mainly based on different statistical models and techniques. However, the geographic approach uses techniques such as SMMs.

It is noted that implicit model based statistical approaches include three types of estimators, which are synthetic, composite and demographic estimators. Whereas, there are also three kinds of explicit models categorized as area level, unit level and general linear mixed models. Based on the type of study researchers are interested in, each of these models is widely applied to obtained small area indirect estimates by utilising the (empirical-) best linear unbiased prediction (E-BLUP), empirical Bayes (EB) and hierarchical Bayes (HB) methods. A very brief synopsis of different direct and indirect statistical model based small area estimation techniques is presented in Table 1 (for details, see Rahman, 2008a).

Figure 1 A summary of different techniques for small area estimation



(*after* Rahman, 2008a).

Table 1 A brief summary of different methods for direct and indirect statistical model based small area estimation.

Small Area Estimation			Formula/model[1]	Methods/Comments	Advantages	Disadvantages	Applications
Direct	Horvitz-Thompson estimator		$\hat{Y}_i = \sum_{k \in s_i} d_{ik} y_{ik}$	Only based on real sample units.	Easy to calculate and unbiased for large samples.	It is unreliable and can not use auxiliary data.	Only if sample size is large enough.
	Generalised regression or GREG estimator		$\hat{Y}_i = X_i'\hat{\beta} + \sum_{k \in s_i} d_{ik}(y_{ik} - x_{ik}'\hat{\beta})$	Based on real data and weighted least square (WLS) estimate of regression coefficient.	Can use auxiliary data at small area level, and approximately design and model unbiased.	It could be negative in some cases and not a consistent estimator due to high residuals.	When sample size is large and reliable auxiliary data are available at small area level.
	Modified direct estimator		$\hat{Y}_i = \hat{Y}_i + (X_i - \hat{X}_i)'\hat{\beta}$	Based on real sample, auxiliary data and WLS estimate of regression.	Design unbiased and uses overall aggregated data for coefficient estimation.	Borrows strength from the overall data but can not increase effective sample.	When the overall sample size is large and reliable.
Indirect modelling	Implicit models	Synthetic estimator	$\hat{Y}_i = \sum_j (X_{ij} / X_{.j}) \hat{Y}_{.j}$	Requires actual sample and auxiliary data for a large scale domain.	Straightforward formula and very easy and inexpensive to calculate.	All small areas are similar to large area assumption is not tenable & estimate is biased.	Used in various areas in government and social statistics.
		Composite estimator	$\hat{Y}_i = \phi_i \hat{Y}_{i[D]} + (1-\phi_i)\hat{Y}_{i[S]}$	Based on direct and synthetic estimators.	Have choices of balancing weight at small areas.	Biased estimator; depends on the chosen weight.	If direct and synthetic estimates are possible.
		Demographic estimator	$\hat{p}_{it} = \frac{1}{2}\left(\frac{b_{it}}{\hat{r}_{it}^b} + \frac{d_{it}}{\hat{r}_{it}^d}\right)$	Rooted in data from census, and with time dependent variable.	Easy to estimate, and the underlying theory is simple and straightforward.	Only used for population estimates and affected by miscounts in census data.	Used to find birth and death rates and various population estimates.
	Explicit models[2]	Area level	$\hat{\theta}_i = x_i'\beta + \varepsilon_i + e_i$	Based on a two stages model and known as the Fay-Herriot model.	Can use area specific auxiliary data and direct estimator.	Assumptions of normality with known variance may untenable at small sample.	Various areas in statistics fitting with assumptions of the model.
		Unit level	$y_{ij} = x_{ij}'\beta + \varepsilon_i + e_{ij}$	Based on unit level auxiliary data and a nested error model.	Useful for continuous value variables, two stage and multivariate data.	Validating is quite complex and unreliable.	Used successfully in many areas of agricultural statistics.
		General linear mixed model	$y = X\beta + Z\varepsilon + e$	A general model, which encompasses all other small area models.	Can allow correlation between small areas, AR(1) and time series data.	Calculation and structure of matrix for area random effects are very complex.	In all areas of statistics where data are useful for the general model.

[1] All usual notations are utilized (see Rahman, 2008a for details).

[2] Methods such as empirical best linear unbiased predictor (E-BLUP), empirical Bayes (EB) and hierarchical Bayes (HB) are frequently used in explicit model based small area estimation. An excellent discussion of each of these complex methods is given in Rao (2003), and also in Rahman (2008a).

However, in contrast to statistical approaches, the geographic approach is based on microsimulation models, which are essentially creating synthetic/simulated micro-population data to produce 'simulated estimates' at small area level. To obtain reliable microdata at small area level is the key task for spatial microsimulation modelling. Synthetic reconstruction and reweighting are two commonly used methods in microsimulation, and each of them is stimulated by different techniques to produce simulated estimators. As the main objective of this paper is to discuss the methodological issues of spatial microsimulation modelling, the subsequent sections will encompass those methods for explicit treatments.

2.1 SMM estimation

The Spatial Microsimulation Model (SMM) approach to small area estimation harks back to the microsimulation modelling ideas pioneered in the middle of last century by Guy Orcutt (1957, 2007). This approach is fully based on SMMs and also known as the geographic method. During the last two decadesmicrosimulation modelling has become a popular, cost-effective and accessible method for socioeconomic policy analysis, with the rapid development of increasingly powerful computer hardware; the wider availability of individual unit record datasets (Harding, 1993, 1996); and with the growing demand (Harding and Gupta, 2007) for small area estimates at government and private sectors.

Microsimulation modelling was originally developed as a tool for economic policy analysis (Merz, 1991). Clarke and Holm (1987) provide a thorough presentation on how microsimulation methods can be applied in regional science and planning analysis. According to Taylor et al. (2004), spatial microsimulation can be conducted by re-weighting a generally national level sample so as to estimate the detailed socio-economic characteristics of populations and households at a small area level. This modelling approach combines individual or household microdata, currently available only for large spatial areas, with spatially disaggregate data to create synthetic microdata estimates for small areas (Harding et al., 2003). Various microsimulation models such as *static*, *dynamic* and *spatial* microsimulation models are discussed in the literature (Harding, 1996; Harding and Gupta, 2007).

Although microsimulation techniques have become useful tools in the evaluation of socioeconomic policies, they involve some complex subsequent procedures. An overall process involved with spatial microsimulation is described in detail by Chin and Harding (2006). They classified two major steps within this process, which are first, to create household weights for small areas using a reweighing method and, second, to apply these household weights to the selected output variables to generate small area estimates of the selected variables. Further, each of these major steps involve several sub-steps (Chin and Harding 2006). Ballas et al. (2005) outline four major steps involved with a microsimulation process, which are:

- the construction of a 'microdata' set (when this is not available);
- Monte Carlo sampling from this data set to 'create' a micro level population (or a 'synthetic' population (see, Chin and Harding, 2006)) for the interested domain;
- *what-if* simulations, in which the impacts of alternative policy scenarios on the population are estimated; and
- dynamic modelling to update a basic microdata set.

The starting point for microsimulation models is a microdata file, which provides comprehensive information on different characteristics of individual persons, families or households in the file. In Australia, microdata are generally available in the form of confidentialised unit record files (CURFs) from the Australian Bureau of Statistics (ABS) national level surveys. Typically, the survey data provide a very large number of variables and an adequate sample size to allow statistically reliable estimates for only large domains (such as only at the broad level of the state or territory). Small area estimates from these national sample surveys are statistically unreliable because of sample observations being insufficient or in many cases non-existent where the domain of interest may fall out of the sample areas. For example if a land development agency wants to develop a new housing domain/suburb, then this new small domain should be out of the sample areas. Also, in order to protect the privacy of the survey respondents, national microdata often lack a geographical indicator which, if present, is often only at the wide level of the state or territory (Chin and Harding, 2006). Therefore spatial microdata are usually unavailable and they need to be synthesized (Chin et al., 2005). The lack of spatially explicit microdata has in the past constrained of SMM for modelling of social policies and human behaviour.

One advantage of SMMs relative to the more traditional statistical small area estimation approaches is that the microsimulation models can be used for estimating the local or small area effects of *policy change* and future small area estimates of population characteristics and service needs (Williamson et al., 1998; Ballas et al., 2003; Taylor et al., 2004; Brown and Harding, 2005; Chin et al., 2005; Ballas et al., 2006; Chin and Harding, 2006; Cullinan et al., 2006; Lymer et al., 2006; Anderson, 2007; Chin and Harding, 2007; King, 2007; Tanton, 2007). For instance, spatial microsimulation may be of value in estimating the distributions of different population characteristics such as income, tax and social security benefits, income deprivation, housing unaffordability, housing stress, housing demand, care needs, etc. at small area level, when contemporaneous census and/or survey data are unavailable (Taylor et al., 2004; Chin et al., 2005; Lymer et al., 2006; Anderson, 2007; Tanton, 2007; Lymer et al., 2008; Harding et al., 2009).

This type of model is mainly intended to explore the relationships among regions and sub-regions and to project the spatial implications of economic development and policy changes at a more disaggregated level. Moreover spatial microsimulation modelling has some advanced features, which can be highlighted as:

- spatial microsimulation models are flexible in terms of the choice of spatial scale;
- they can allow data from various sources to create a microdata base file at the small area level;
- the models store data efficiently as lists of objects;
- spatial microdata have the potential for further aggregation or disaggregation; and
- models allow for updating and projecting.

Thus, from some points of view, spatial microsimulation exploits the benefits of object-oriented planning, both as a tool and a concept. Spatial microsimulation frameworks use a list-based approach to microdata representation where a household or an individual has a list of attributes that are stored as lists rather than as occupancy matrices (Williamson et al., 1996). From a computer programming perspective, the list-based approach uses the tools of object-oriented programming because the individuals and households can be seen as objects with their attributes as associated instance variables. Alternatively, rather than using an object orientated programming approach, a programming language like SAS can also be used to run spatial microsimulation. For a technical discussion of the SAS-based environment used in the development of the STINMOD model and adapted to run other NATSEM regional level models, readers may refer to the technical paper by Chin and Harding (2006). Furthermore, by linking spatial microsimulation with static microsimulation we may be able to measure small area effects of policy changes, such as changes in government programs providing cash assistance to families with children (Harding et al., 2009). Another advantage of SMMs is the ability to estimate the geographical distribution of socio-economic variables, which were previously unknown (Ballas, 2001).

However spatial microsimulation adds to the simulation a spatial dimension, by creating and using synthetic microdata for small areas, such as SLA levels in Australia (Chin et al., 2005). There is often great difficulty in obtaining household microdata for small areas, since spatially disaggregate reliable data are not readily available. Even if these types of data are available in some form, they typically suffer from severe limitations – in either lack of characteristics or lack of geographical detail. Therefore, spatial microdata should be simulated, and that can be achieved by different probabilistic as well as deterministic methods.

3. METHODOLOGICAL ISSUES IN SMM

As mentioned calculating statistically reliable population estimates in a local area using survey microdata is challenging, due to the lack of enough sample observations. To create a synthetic spatial microdata set is one of the possible solutions. Simulation based methods can deal with such a problem by (re)weighting each respondent in the survey data, to create the synthetic spatial microdata. However, it is not easy process to create reliable spatial microdata. Complex methodologies are associated with the process. This section presents some of the vital methodological issues in spatial microsimulation modelling.

3.1 Creation of synthetic spatial microdata

Methods for creating synthetic spatial microdata are mainly classified into the synthetic reconstruction and reweighting methods. Synthetic reconstruction is an older method, which attempts to construct synthetic micro-populations at the small area level in such a way that all known constraints at the small area level are reproduced. There are two ways of undertaking synthetic reconstruction - data matching or fusion (Moriarity and Scheuren, 2003; ABS, 2004; Tranmer et al., 2005) and iterative proportional fitting (Birkin and Clarke, 1988; Duley, 1989; Williamson, 1992; Norman, 1999). In contrast, the reweighting method, which is a relatively new and popular method, mainly calibrates the sampling design weights to a set of new weights based on a distance measure, by using the available data at spatial scale.

Data matching or fusion is a multiple imputation technique often useful to create complementary datasets for microsimulation models. Data collected from two different sources may be matched using variables (such as name and address or different IDs), which uniquely identify an individual or household. This type of data matching is commonly known as 'exact matching'. But, due to data confidentiality constraints, these unique identifier variables may not be available in all cases (for example, sample units or households in microdata such as CURFs of the ABS used in NATSEM cannot be identified because of the existence of data privacy legislation when gathering data from the population). For such a case, records from different datasets can also be 'matched' if they share a core set of common characteristics. In general, the data matching technique involves a few empirical steps:

- adjusting available data files and variable transformations;
- choosing the matching variables;
- selecting the matching method and associated distance function; and
- validating.

A description of these empirical steps and theories behind them are available elsewhere (Alegre et al., 2000; Rassler, 2002). Details about data matching techniques are given by Rodgers (1984). Moreover, this tool is used to create microdata files by researchers in many countries, such as Moriarity and Scheuren (2001, 2003) in the USA;

Table 2 Synthetic reconstruction *versus* the reweighting technique

Synthetic reconstruction	Reweighting technique
○ It is based on a sequential step by step process – where the characteristics of each sample unit are estimated by random sampling using a conditional probabilistic framework.	○ It is an iterative process – where a suitable fitting between actual data and the selected sample of microdata should be obtained by minimizing distance errors.
○ Ordering is essential in this process (each value should be generated in a fixed order).	○ Ordering is not an issue. However convergence is achievable by repeating the process many times or by some simple adjustment.
○ Relatively more complex and time consuming.	○ The technique is complex from a theoretical point of view, it is comparatively less time consuming.
○ The effects of inconsistency between constraining tables could be significant for this approach due to a mismatch in the table totals or subtotals.	○ Reweighting techniques can allow the choice of constraining tables to match with researcher and/or user requirements.

Liu and Kovacevic (1997) in Canada; Alegre et al. (2000), Tranmer et al. (2005), Rassler (2004) in Europe; and ABS (2004) in Australia, among many others.

Besides, the iterative proportional fitting (IPF) tool initially proposed by Deming and Stephan (1940). The authors developed the method for adjusting cell frequencies in a contingency table based on sampled observations subject to known expected marginal totals. This method has been used for several decades to create synthetic microdata from a variety of aggregate data sources. The theoretical and practical considerations behind this method have been discussed in several studies (Fienberg, 1970; Evans and Kirby, 1974; Norman, 1999), and the usefulness of this approach in spatial analysis and modelling has been revealed by Birkin and Clarke (1988), Wong (1992), Ballas et al. (1999) and Simpson and Tranmer (2005). The study by Wong (1992) also considers the reliability issues of using the IPF procedure and demonstrates that the estimates of individual level data generated by this process using data of equal-interval categories other than equal-size categories are more reliable, and the performance of the estimation can be improved by increasing sample size.

Previous to the development of 'reweighting' techniques, the iterative proportional fitting procedure was a very popular tool to generate small area microdata. A summary of literature using this technique has been provided by Norman (1999). It appears from the study that almost all of the researchers in the United Kingdom were devoted to using the iterative proportional fitting procedure in microsimulation modelling. But nowadays most of the researchers are claiming that reweighting procedures have some advantages over the synthetic reconstruction approach (Williamson et al., 1998; Huang and Williamson, 2001; Ballas et al., 2003). A summary of the key issues associated with the two approaches is shown in Table 2.

Moreover, reweighting is a procedure used throughout the world to transform information contained in a sample survey to estimates for the micro population (Chin and Harding, 2006). For example, the Australian Bureau of Statistics calculates a weight (or 'expansion factor') for each of the 6,892 households included in the 1998-99 Household Expenditure Survey sample file (ABS, 2002). Thus if household number 1 is given a weight of 1000 by the ABS, it means that the ABS considers that there are 1000 households with comparable characteristics to household number 1 in Australia. These weights are used to move from the 6,892 households included in the HES sample to estimates for the 7.1 million households in Australia (Chin and Harding, 2006).

There are two reweighting techniques for SMM, which are a generalised regression technique known as the GREGWT approach (Bell, 2000; Chin and Harding, 2006) and the combinatorial optimisation technique (Williamson et al., 1998; Huang and Williamson, 2001; Ballas et al., 2003; Williamson, 2007). These techniques are widely used to create synthetic spatial microdata for the spatial microsimulation modelling approach of small area estimation. However, they have a different methodology. Details of these two reweighting methodologies are given in the following subsections

3.2 GREGWT theory: How does it generate new weights?

The GREGWT approach of reweighting is an iterative generalised regression algorithm written in SAS macros. Let us assume that a finite population is denoted by $\Omega = \{1,2,..., k,..., N\}$, and a sample s ($s \subseteq \Omega$) is drawn from Ω with a given probability sampling design $p(.)$. Suppose the inclusion probability $\Pi_k = \Pr(k \in s)$ is a strictly positive and known quantity. Now for the elements $k \in s$, let (y_k, x_k) be a set of sample observations; where y_k is the value of the variable of interest for the k^{th} population unit and x'_k =

($x_{k,1}$...k,$x_{k,j}$,...,$x_{k,p}$) is a vector of auxiliary information associated y'_k. Note that data for a range of auxiliary variables should be available for each unit of a sample *s*. In a particular case, suppose for an auxiliary variable *j*, the element $x_{k,j}$ = 1 in x_k if the k^{th} individual is not in workforce, and $x_{k,j}$ = 0 'otherwise'. Thus the number of individuals in the sample who are not in the workforce is given by

$$\sum_{k\in s} x_{k,j} \ .$$

If the given sampling design weights are $d_k = 1/\Pi_k$ ($k \in s$) then the sample based population totals of auxiliary information,

$$\hat{t}_{x,s} = \sum_{k\in s} d_k x_k$$

can be obtained for a *p*-elements auxiliary vector x_k. But the *true* value of the population total of the auxiliary information T_x should be known from some other sources such as from the census or administrative records. In practice, $\hat{t}_{x,s}$ is far from T_x when the sample *s* is a bad or poorly representative of the population.

For obtaining a more reliable small area estimate of population total of the variable of interest, we have to generate a new set of weights w_k for $k \in s$, for which the calibration equation

$$\sum_{k\in s} w_k x_k = T_x \qquad (1)$$

must be fitted and the new weights w_k will be as close as possible to d_k.

The distance measure used in the GREGWT algorithm is known as truncated Chi-squared distance function and it can be defined as

$$\overset{\chi^2}{G}_k = \frac{(w_k - d_k)^2}{2d_k}; \text{ for } L_k \le \frac{w_k}{d_k} \le U_k \qquad (2)$$

where $L_k$and $U_k$are pre specified lower and upper bounds respectively for each unit $k \in s$.

For a simple special case the total of this type of distance measure can be defined as

$$D = \frac{1}{2}\sum_{k\in s} \frac{(w_k - d_k)^2}{d_k}.$$

Hence the Lagrangean for the Chi-squared distance function is

$$L = \frac{1}{2}\sum_{k\in s} \frac{(w_k - d_k)^2}{d_k} + \sum_{j=1}^{p}\left(T_{x,j} - \sum_{k\in s} w_k x_{k,j}\right) \qquad (3)$$

where λ_j (j=1,2,...,p) are the Lagrange multipliers, and $T_{x,j}$ is the j^{th} element of the vector of true values of known population total for the auxiliary information, T_x.

By differentiating (3) with respect to w_k and then applying the first order condition, we have

$$\frac{\partial L}{\partial w_k} = \left(\frac{w_k - d_k}{d_k}\right) - \sum_{j=1}^{p} \lambda_j x_{k,j} = 0 \qquad (4)$$

for $k \in s \subseteq \Omega$, along with the p^{th} (j=1,2,...p) constraints conditions in equation (1). As earlier, for a simple representation it is convenient to write

$$x'_k\lambda = \sum \lambda_j x_{k,j} \ .$$

Hence the new weights can be formulated as

$$w_k = d_k + d_k x'_k \lambda \ . \qquad (5)$$

To find the values of the Lagrange multipliers, the equation (5) can be rearranged in a convenient form. After multiplying the equation by x_k and then summing over *k* it can be written as

$$\sum_{k\in s} w_k x_k = \sum_{k\in s} d_k x_k + \sum_{k\in s} d_k x_k x'_k \lambda.$$

Now since $\sum_{k\in s} d_k x_k = \hat{t}_{x,s}$ and $\sum_{k\in s} w_k x_k = T_x$ are known, the above equation can be expressed as

$$\left(\sum_{k\in s} d_k x_k x'_k\right)\lambda = T_x - \hat{t}_{x,s} \qquad (6)$$

where the summing term in brackets is a *p* x *p* symmetric-square matrix. If the inverse of this matrix exists, the vector of Lagrange multipliers can be obtained by the following equation

$$\lambda = \left(\sum_{k\in s} d_k x_k x'_k\right)^{-1}\left(T_x - \hat{t}_{x,s}\right); \text{ for } \left|\sum_{k\in s} d_k x_k x'_k\right| \neq 0 \ . \qquad (7)$$

Hence using the resulting values of Lagrange multipliers, λ, one can easily calculate the new weights w_k from equation in (5). Moreover to minimize the truncated Chi-squared distance function in (2), an iterative procedure known as the Newton-Raphson method (appendix A) is used in the GREGWT program (Bell, 2000). It adjusts the new weights in such a way that minimises equation (2) and produces generalised regression estimates or synthetic estimates of the variable of interest.

Explicit numerical solution for a hypothetical data

An explicit numerical solution of the above very simple case theory is given here. Let $x_{k,j}$ is the j^{th} auxiliary variable linked with k^{th} sample unit for which true population values T_x are available from census or other administrative records. Suppose in a hypothetical dataset, observations of 25 sample units for a set of 5 auxiliary variables such as age (1=16-30 years and 0= 'otherwise'), sex (1=female and 0=male), employment (1=unemployed and 0= 'otherwise'), income from unemployment benefits (in real unit values 0, 1, 2, 3, 4 and 5) and location (1=rural and 0= urban) are available, and its associated auxiliary information matrix, sample design weights and the known population values vector are accordingly given as –

$$X = [x'_{k,j}] = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 1 & 1 & 1 & 5 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 4 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 3 & 1 \\ 1 & 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 5 & 1 \\ 0 & 1 & 1 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 3 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 4 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 5 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad d = (d_k) = \begin{pmatrix} 4 \\ 5 \\ 6 \\ 5 \\ 3 \\ 4 \\ 6 \\ 4 \\ 5 \\ 3 \\ 5 \\ 4 \\ 3 \\ 6 \\ 4 \\ 5 \\ 6 \\ 3 \\ 6 \\ 4 \\ 5 \\ 3 \\ 5 \\ 4 \\ 3 \end{pmatrix}, \quad \text{and} \quad T_x = \begin{pmatrix} 50 \\ 45 \\ 70 \\ 200 \\ 65 \end{pmatrix}$$

Note: the 1st row of matrix X represents a sample unit of *age* between 16 to 30 years, *female*, in *'otherwise'* employment categories that is may be in labour force or employed, with a real unit value of *income* from unemployment is 0 dollar, and living in an *urban* area.

Now we have to estimate $\hat{t}_{x,s}$ and the inverse matrix of $\left(\sum_{k\in s} d_k x_k x'_k\right) =$ A (*say*).

By using mathematical formulas one can easily obtain,

$$\hat{t}_{x,s} = \left(\sum_{k=1}^{25} d_k x_{k,1}, \sum_{k=1}^{25} d_k x_{k,2}, \sum_{k=1}^{25} d_k x_{k,3}, \sum_{k=1}^{25} d_k x_{k,4}, \sum_{k=1}^{25} d_k x_{k,5}\right)' = (46\ \ 42\ \ 69\ \ 206\ \ 64)' \quad \text{and}$$

$$\text{A} = \begin{bmatrix} A11 & A12 & A13 & A14 & A15 \\ A21 & A22 & A23 & A24 & A25 \\ A31 & A32 & A33 & A34 & A35 \\ A41 & A42 & A43 & A44 & A45 \\ A51 & A52 & A53 & A54 & A55 \end{bmatrix} = \begin{bmatrix} 46 & 18 & 31 & 108 & 24 \\ 18 & 42 & 22 & 62 & 12 \\ 31 & 22 & 69 & 206 & 39 \\ 108 & 62 & 206 & 750 & 120 \\ 24 & 12 & 39 & 120 & 64 \end{bmatrix}$$

where Ajj = $\sum_{k=1}^{25} d_k x_{k,j} x'_{k,j} = \sum_{k=1}^{25} d_k x_{k,j}^2$ and Aij = $\sum_{k=1}^{25} d_k x_{k,i} x'_{k,j}$; for all i,j (=1,2,3,4,5) and i ≠ j.

The inverse matrix of A = $\left(\sum_{k\in s} d_k x_k x'_k\right)$ can be obtained as $A^{-1} = \left(\sum_{k\in s} d_k x_k x'_k\right)^{-1}$

$$= \begin{bmatrix} 0.03661582 & -0.00901288 & 0.00228602 & -0.00429437 & -0.00538212 \\ -0.00901288 & 0.03088625 & -0.01214961 & 0.00183273 & 0.00155596 \\ 0.00228602 & -0.01214961 & 0.09100053 & -0.02239201 & -0.01204764 \\ -0.00429437 & 0.00183273 & -0.02239201 & 0.00794951 & 0.00000656 \\ -0.00538212 & 0.00155596 & -0.01204764 & 0.00000656 & 0.02468079 \end{bmatrix}$$

Then by using the results in relationship (7), the Lagrange multipliers should be calculated for this simple particular example as: $\lambda' = (0.14209475, 0.03501717, 0.18600019, -0.08176176, -0.00426682)$.

Table 3 New weights and its distance measures to sampling design weights

d_k	w_k	$w_k - d_k$	$\overset{\chi^2}{G_k}$
4	**4.70844769**	0.70844769	0.06273727
5	**5.39271424**	0.39271424	0.01542245
6	**6.10925911**	0.10925911	0.00099480
5	**4.77151662**	-0.22848338	0.00522047
3	**3.09225105**	0.09225105	0.00141838
4	**4.41695372**	0.41695372	0.02173130
6	**5.97439907**	-0.02560093	0.00005462
4	**4.00419164**	0.00419164	0.00000220
5	**5.15375174**	0.15375174	0.00236396
3	**3.41348379**	0.41348379	0.02849481
5	**5.69627800**	0.69627800	0.04848031
4	**4.45424007**	0.45424007	0.02579175
3	**3.48091381**	0.48091381	0.03854635
6	**4.63754748**	-1.36245252	0.15468974
4	**3.57588131**	-0.42411869	0.02248458
5	**5.00000000**	0	0
6	**6.47125708**	0.47125708	0.01850694
3	**3.10505151**	0.10505151	0.00183930
6	**6.10925911**	0.10925911	0.00099480
4	**4.00419164**	0.00419164	0.00000219
5	**4.97866589**	-0.02133411	0.00004551
3	**2.31877374**	-0.68122626	0.07734487
5	**5.88555961**	0.88555961	0.07842158
4	**4.55702240**	0.55702240	0.03878424
3	**3.41348379**	0.41348379	0.02849481
		TAD = 9.21152591	D **=** 0.67286721

Now using this result in equation (5), the new weights or calibrated weights for the Chi-squared distance measure can be easily obtained. The calculated new weights and its distance measures to the sample design weights are given in Table 3. For the 16th unit of our hypothetical data, the new weight remains unchanged to the sampling design weight due to the fact that all entries for this unit are zero. However this is very rare in GREGWT reweighting.

In addition, the total absolute distance (TAD) indicates higher quantity. While *absolute distance* has a higher value, the corresponding *Chi-squared distance* measure also indicates a higher value. However the fluctuations within absolute distances are remarkable compared to Chi-squared distance measures (see in Figure 2).

Furthermore, when the TAD will zero the total Chi-squared distance will also be zero, and in that situation the calibrated weights will remain same as the sampling design weights which indicates the sample data are fully representative to the small area population. Moreover, it is interesting to note that the values of a set of *new weights* vary greatly with the changing values of vector for differences between $\hat{t}_{x,s}$ and T_x. These differences can come from the poorly representative data and/or the chosen benchmarks. Four random alternative cases of difference vectors

$C\,1 = [4,3,1,-6,1]'$,

$C\,2 = [8,3,1,-6,1]'$,

$C\,3 = [12,3,1,-6,1]'$

and $C\,4 = [4,3,1,2,1]'$,

where $C_j=[T_x - \hat{t}_{x,s}]$ for j = 1,2,3,4 ,

have been considered and the resulting sets of new weights are plotted in Figure 3. The results show that the case $C4$ generates a more consistent set of new weights compared to the other cases. It is obvious that when the auxiliary information matrix provides quite rich sample data then the resulting difference vector between $\hat{t}_{x,s}$ and T_x will be fairly close. Hence the resulting set of calibrated weights will produce more accurate estimates.

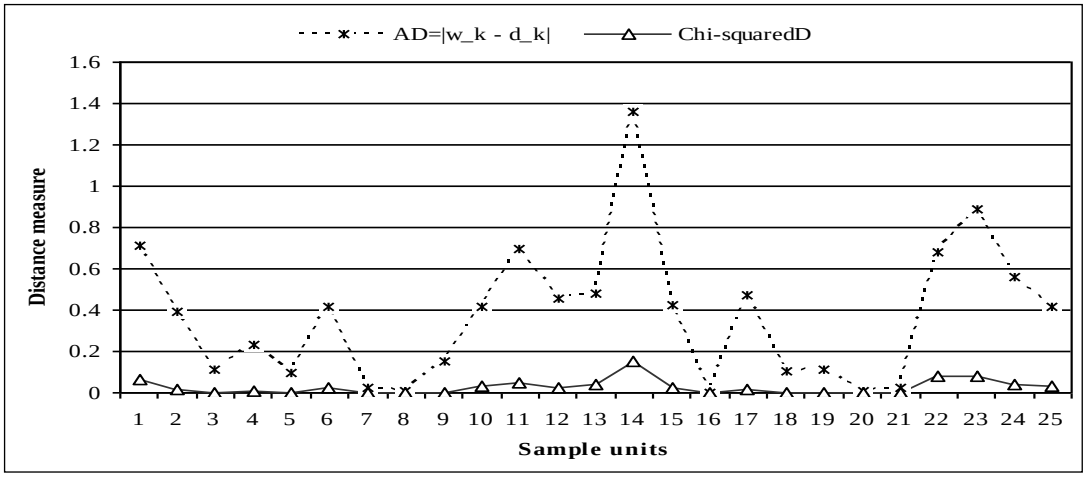


Figure 2 A comparison of absolute distance and Chi-squared distance measures

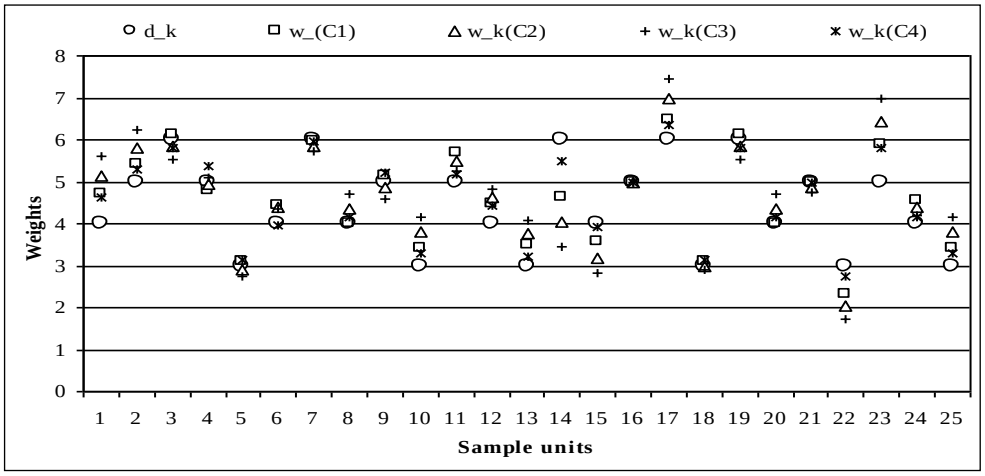


Figure 3 Plots of sampling design weights and new weights for specific cases

3.3 Combinatorial optimisation reweighting approach

The combinatorial optimisation (CO) reweighting approach was first suggested in Williamson et al. (1998) as a new approach to create synthetic micro-populations for small domains. This reweighting method is mainly motivated towards selecting an appropriate combination of households from survey data to attain the known benchmark constraints at small area levels using an optimization tool. In the combinatorial optimisation algorithms, an iterative process begins with an initial set of households randomly selected from the survey data, to see the fit to the known benchmark constraints for each small domain. Then a random household from the initial set of combinations is replaced by a randomly chosen new household from the remaining survey data to assess whether there is an improvement of fit. The iterative process continues until an appropriate combination of households that best fits known small area benchmarks is achieved (Williamson et al., 1998; Voas and Williamson, 2000; Huang and Williamson, 2001; Tanton et al., 2007). The overall process involves five steps which are as follows:

1. collect a sample survey microdata file (such as CURFs in Australia) and small area benchmark constraints (for example, from census or administrative records);
2. select a set of households randomly from the survey sample which will act as an initial combination of households from a small area;
3. tabulate selected households and calculate total absolute difference from the known small area constraints;
4. choose one of the selected households randomly and replace it with a new household drawn at random from the survey sample, and then follow step 3 for the new set of households combination; and
5. repeat step 4 until no further reduction in total absolute difference is possible.

Note that when an array based survey data set contains a finite number of households it is possible to calculate all possible combinations of households. In theory, it may also be possible to find the set of households' combination that best fits the known small area benchmarks. But, in practice, it is almost unachievable, due to computing constraints for a very large number of all possible solutions. For example, to select an

appropriate combination of households for a small area with 150 households from a survey sample of 215789 households, the number of possible solutions greatly exceeds a billion (Williamson et al., 1998).

To overcome this difficulty, the combinatorial optimisation approach uses several ways of performing 'intelligent searching', effectively reducing the number of possible solutions. Williamson et al. (1998) provide a detailed discussion about three intelligent searching techniques: hill climbing, simulated annealing and genetic algorithms. Later on, to improve the accuracy and consistency of outputs, Voas and Williamson (2000) developed a 'sequential fitting procedure', which can satisfy a level of minimum acceptable fit for every table used to constrain the selection of households from the survey sample data. The following section will address the simulated annealing method only.

The simulated annealing method in CO

Simulated annealing, an intelligent searching technique for optimisation problems, has been successfully used in the CO reweighting process to create spatial microdata. The method is based on a physical process of annealing – in which a solid material is first melted in a heat bath by increasing the temperature to a maximum value at which point all particles of the solid have high energies and the freedom to randomly arrange themselves in the liquid phase. The process is then followed by a cooling phase, in which the temperature of the heat bath is slowly lowered. When the maximum temperature is sufficiently high and the cooling is carried out sufficiently slowly then all the particles of the material eventually arrange themselves in a state of high density and minimum energy. Simulated annealing has been used in various combinatorial optimisation problems (Kirkpatrick et al., 1983; van Laarhoven and Aarts, 1987; Williamson et al., 1998; Pham and Karaboga, 2000; Ballas, 2001).

The simulated annealing algorithm used in the CO reweighting approach was originally based on the Metropolis algorithm, which had been proposed by Metropolis et al. (1953). To simulate the evaluation to 'thermal equilibrium' of a solid for a fixed value of the temperature T the authors introduced an iterative method, which generates sequences of states of the solid in the following way. As mentioned in the book Simulated Annealing: Theory and Applications by van Laarhoven and Aarts (1987, p. 8):

> "Given the current state of the solid, characterized by the position of its particles, a small, randomly generated, perturbation is applied by a small displacement of a randomly chosen particle. If the difference in energy, ∂E, between the current state and the slightly perturbed one is negative, that is, if the perturbation results in a lower energy for the solid, then the process is continued with the new state. If $\partial E \geq 0$, then the probability of acceptance of the perturbed state is given by $\exp(\partial E / K_B T)$. This acceptance rule for new states is referred to as the *Metropolis Criterion*. Following this criterion, the system eventually evolves into thermal equilibrium, that is, after a large number of perturbations, using the aforementioned acceptance criterion, the probability distribution of the states approaches the Boltzmann distribution, given as
>
> $$p(\partial E) = \frac{1}{c(T)} \exp\left(-\frac{\partial E}{K_B T}\right)$$
>
> where $c(T)$ is a normalizing factor depending on the temperature T and K_B is the Boltzmann constant."

To search an appropriate combination of households from a survey dataset that best fits the benchmark constraints at small area levels is a combinatorial optimisation problem, and solutions in a combinatorial optimisation problem are equivalent to states of a physical annealing process. In the process of CO reweighting by simulated annealing algorithm, a combination of households takes the role of the states of a solid while the total absolute distance (TAD) function and the control parameter (for example, rate of reduction) take the roles of energy and temperature respectively. According to Williamson et al. (1998), change in energy becomes potential change in households' combination performance (assessed by TAD) to meet the benchmarks, and temperature becomes a control for the maximum level of performance degradation (% of reduction) acceptable for the change of one element in a combination of households by a random element picks from the sample data. The control parameter is then lowered in steps, with the system being allowed to approach equilibrium for each step by generating a sequence of combinations by obeying the Metropolis criterion.

In addition, the algorithm is terminated for some small value of the control parameter, for which practically no deteriorations are accepted. Hence the normalizing constant which is depending on the controlling factor as well as Boltzmann constant can be dropped from the probability distribution. In this particular case we have the equation:

$$p(\partial E) = \exp\left(-\frac{\partial E}{T}\right).$$

There are two important features of this probability equation described by Williamson et al. (1998). One is that the smaller the value of difference in energy, ∂E, the greater is the likelihood of a potential replacement being made in a combination. Another feature is that the smaller the value of controlling factor *T*, the smaller the change in performance likely to be accepted.

A typical simulated annealing algorithm is depicted in Figure 4. The overall process consists of a series of iterations in which random shifting is occurring from an existing solution to a new solution among all possible solutions. To accept a new solution as the base solution for further iteration, a test of goodness-of-fit based on TAD is consistently

checked. The rules of the test are if the change of the difference in energy is negative, the newly simulated solution is accepted unconditionally; otherwise it is accepted satisfying the abovementioned Metropolis criterion. The simulated annealing algorithm may be able to avoid deceiving at local extremum in the solutions. Moreover, a solution or selected combination of households by this algorithm can generate real individuals living in actual households in a sense that individuals are from modelled outputs and not synthetically reconstructed (Ballas, 2001).

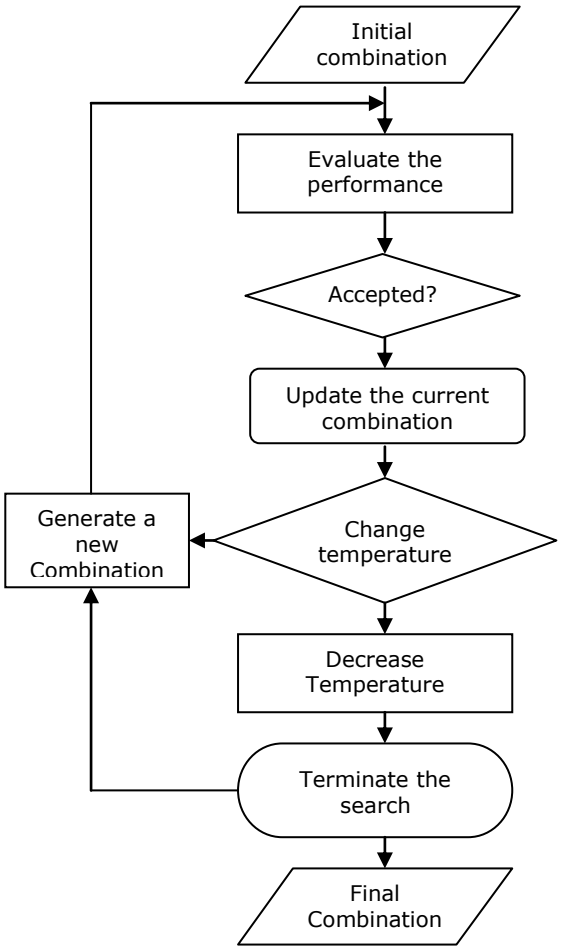


Figure 4 A flowchart of the simulated annealing algorithm

(*after* Pham and Karaboga, 2000)

An illustration of CO process for hypothetical data
A simplified combinatorial optimisation process is depicted in Figure 5. It is noted that in this process when the total absolute difference (aforementioned TAD) in *gregwt* section) is equal to zero, the selection of households' combination indicates the best fit. In other words, in this case the new weights give the actual households units from the survey sample microdata, which are the best representative combination. Thus it is a selection process of an appropriate combination of sample units, rather than calibrating the sampling design weights to a set of new weights.

4. COMPARISON OF REWEIGHTING TECHNIQUES AND A NEW APPROACH

In this section, a comparison of the two reweighting methodologies is given with a new approach to the creation of synthetic spatial microdata.

4.1 Comparison of GREGWT and CO

Although both the reweighting approaches are widely used in the creation of small area synthetic microdata, the methodology behind each approach is quite different. For instance, GREGWT is typically based on generalised linear regression and attempts to minimize a truncated Chi-squared distance function subject to the small area benchmarks. Combinatorial optimisation, on the other hand, is based on 'intelligent searching' techniques and attempts to select a combination of appropriate households from a sample that best fits the benchmarks.

Tanton et al. (2007) provide a comparison of these two approaches using a range of performance criteria. The study also covers the advantages and disadvantages of each method. Using the data of the 1998-99 Household Expenditure Survey from Australia, the study reveals that the GREGWT algorithm seems to be capable of producing good results. However the GREGWT algorithm has some limitations compared to the combinatorial optimisation algorithm. One of the drawbacks of GREGWT approach is that for some small areas, 'convergence' does not exist.

That means that the GREGWT algorithm is unable to produce estimates for those small areas, while the combinatorial optimisation algorithm is able to do so. In addition, the GREGWT algorithm takes more time to run compared to combinatorial optimisation, and it is still unclear whether that extra time is due to the different programming language (GREGWT is written in SAS code and CO uses compiled FORTRAN code) or the relative efficiencies of the underlying algorithms. Moreover the combinatorial optimisation routine has a tendency to include fewer households but give them higher weights – and, conversely, the GREGWT routine has a tendency to select more households but give them smaller weights.

A comparison of the GREGWT and CO reweighting approaches is summarized in Table 4. The focus is here mostly on methodological issues. However some entries are consistent with Tanton et al. (2007).

Figure 5 A simplified combinatorial optimisation process

Step 1: Obtain sample survey microdata and small area constraints.

Survey Sample Microdata

Household	Characteristics		
	size	adult	children
a	2	2	0
b	2	1	1
c	4	2	2
d	1	1	0
e	3	2	1

Known small area constraints

1. Household size

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

2. Age of occupants

Type of person	Frequency
adult	3
child	2
Total	**5**

Step 2: Randomly select two households from survey sample (for example, **a** & **e**) to act as an initial small area microdata estimate.

Step 3: Tabulate selected households and calculate absolute difference from known constants.

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

Age	Estimated frequency (1)	Observed frequency (2)	Absolute difference |(1)-(2)|
adult	4	3	1
child	1	2	1
	Sub-total:		***2***

Total absolute difference = 4+2 = 6

Step 4: Randomly select one of the selected households (a or e). Then replace with another household selected at random from the survey sample, provided this leads to a reduced total absolute difference.

Households selected: d & e (Household a replaced by d). Tabulate this new combination of households and calculate absolute difference from known constants.

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

Age	Estimated frequency (1)	Observed frequency (2)	Absolute difference |(1)-(2)|
adult	3	3	0
child	1	2	1
	Sub-total:		***1***

Total absolute difference = 2+1 = 3

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

Result: Final selected households are **c** &**d** (since this household combination best fits the small area benchmarks):

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

Age	Estimated frequency (1)	Observed frequency (2)	Absolute difference |(1)-(2)|
adult	3	3	0
child	2	2	0
	Sub-total:		***0***

Total absolute difference = 0+0 = 0

(*after* Huang and Williamson, 2001)

Table 4 A comparison of the GREGWT and CO reweighting methodologies

GREGWT	CO
o An iterative process.	o An iterative process.
o Use the Newton-Raphson method of iteration.	o Use a stochastic approach of iteration MCMC.
o Based on a distance function.	o Based on a combination of households.
o Attempt to minimize the distance function subject to the known benchmarks.	o Attempt to select an appropriate combination that best fits the known benchmarks.
o Use the Lagrange multipliers as minimisation tools for minimising the distance function.	o Use different techniques as intelligent searching tools in optimizing combinations of households.
o Weights are in fractions.	o Weights are in integers (but could be fractions).
o Boundary condition is applied to new weights for achieving a solution.	o There is no boundary condition to new weights.
o The benchmark constraints at small area levels are fixed for the algorithms.	o The algorithm is designed to optimize fit to a selected group of tables, which may or may not be the most appropriate ones. Hence there may be a choice of benchmark constraints.
o Typically focus on simulating microdata at small area levels and aggregation is possible at larger domains.	o Offers a flexibility and collective coherence of microdata, making it possible to perform mutually consistent analysis at any level of aggregation or sophistication.
o All estimates have their own standard errors obtained by a group jackknife approach.	o No information about this in literature. May be possible in theory but nothing available in practice.
o In some cases convergence does not exist and this requires readjusting the boundary limits or a proxy indicator for this nonconvergence.	o There are no convergence issues. However, the finally selected household combination may still fail to fit user-specified benchmark constraints.
o There is no standard index to check the statistical reliability of the estimates.	o There is no standard index to check the statistical reliability of the estimates.
o The iteration procedure can be unstable near a horizontal asymptote or at local extremum.	o The iteration algorithm may able to avoid deceiving at local extremum in the solutions.

4.2 Bayesian prediction approach of small area microdata simulation

A new system for creating synthetic spatial microdata is offered in this subsection. It is noted that after the sample survey, a finite population usually has two parts - which are observed units in the sample called data and unobserved sampling units in the population (Figure 6). Suppose Ω represents a finite population in which Ω_i (say) is the subpopulation of small area i. Now if s_i denotes the observed sample units in the i^{th} area then we have

$s_i \cup \overline{s}_i = \Omega_i \subseteq \Omega$ for $\forall_i$,

where $\overline{s}_i$ denotes the unobserved units in the small area population. Let y_{ij} represents a variable of interest for the j^{th} characteristic of the population at i^{th} small area. Thus we always have the estimate of population total at i^{th} small area

$$t_{y_{i.}} = \sum_{j \in s_i} y_{ij} + \sum_{j \in \overline{s}_i} y_{ij} \ .$$

The main challenge in this process is to establish the link of observed data to the unobserved sampling units in the population. It is a kind of prediction problem, where a modeller tries to find a probability distribution of unobserved responses using the observed sample and the auxiliary data. The Bayesian methodology (see Ericson, 1969; Lo, 1986; Little, 2007; Rahman, 2008b) can deal with such a prediction problem.

The Bayesian prediction theory is very straightforward and mainly based on the Bayes's posterior distribution of unknown parameters (Rahman 2008c). Let y be a set of observed units from a model with a joint probability density $p(y|\theta)$, in which θ is a set of model parameters. If a prior density of unknown parameters θ is g, the posterior density of θ for given y can be obtained by Bayes's theorem and defined as $p(y|\theta) \propto p(y|\theta) g(\theta)$.

Now, if $\overline{y}$ is the set of unobserved units in a finite population, then under the Bayesian methodology its prediction distribution can be obtained by solving the integral

$$p(\overline{y} \mid y) \propto \int_{\theta} p(\theta \mid y) p(\overline{y} \mid \theta) d\theta \,,$$

where $p(\overline{y}|\theta)$ is the probability density of unobserved units. Details of the Bayesian prediction theory for various regression models are given in Rahman (2008c).

For an i^{th} small area, let a multivariate linear model for the observed sample units

$$Y_i = (y_1, y_2, ..., y_{n_i})'$$

be $Y_i = X_i\beta + E_i$

with errors distribution

$$E_i \sim T_{n_i p}(0, I_{n_i \times n_i}, \Sigma_{p\times p}, \nu),$$

and for unobserved population units let

$$\overline{Y}_i = (\overline{y}_1, \overline{y}_2, ..., \overline{y}_{N_i - n_i})'$$

be $\overline{Y}_i = \overline{X}_i\beta + \overline{E}_i$

with errors distribution

$$\overline{E}_i \sim T_{(N_i-n_i)p}(0, I_{(N_i-n_i)\times(N_i-n_i)}, \Sigma_{p\times p}, \nu),$$

where all the notations are as usual (see Rahman 2008c).

Applying the Bayesian prediction theory under a prior distribution

$$p(\beta, \Sigma) \propto |\Sigma|^{-(p+1)/2},$$

we can derive the distribution of unobserved population units as

$$f(\overline{Y}_i \mid Y_i, \beta, \Sigma) = C\left[S_{Y_i} + (\overline{Y}_i - \overline{X}_i\hat{\beta})'H(\overline{Y}_i - \overline{X}_i\hat{\beta})\right]^{-\frac{N_i-k}{2}}$$

where

$\hat{\beta}$ is the OLS of β,

$$S_{Y_i} = (Y_i - X_i\hat{\beta})'(Y_i - X_i\hat{\beta})$$

$$H = [I - \overline{X}_i(X_i'X_i + \overline{X}_i'\overline{X}_i)^{-1}\overline{X}_i],$$

And

$$C = \frac{(\pi)^{-\frac{(N_i-n_i)p}{2}}\Gamma_p\left(\frac{n_i-k}{2}\right)|H|^{-\frac{p}{2}}}{\Gamma_p\left(\frac{N_i-k}{2}\right)|S_{Y_i}|^{-\frac{n_i-k}{2}}}$$

is the normalizing constant.

The joint posterior density of parameters (*say*) $\theta = (\beta, \Sigma)'$ for the observed sample units Y_i and unobserved population units $\overline{Y}_i$ can be determined as

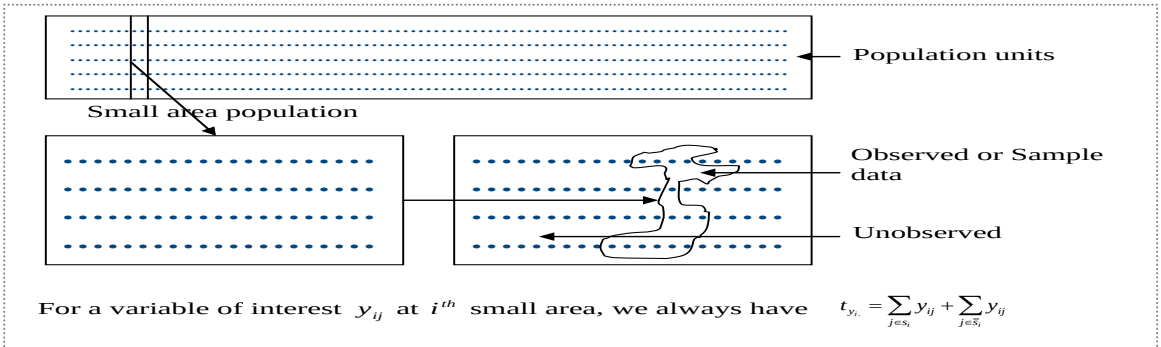
$$f(\beta, \Sigma \mid Y_i, \overline{Y}_i) \propto |\Sigma|^{-\frac{N_i+p+1}{2}}\left|I_p + \Sigma^{-1}Q\right|^{-\frac{\nu+p+N_i-1}{2}} \quad (8)$$

where

$$Q = (Y_i - X_i\beta)'(Y_i - X_i\beta) + (\overline{Y}_i - \overline{X}_i\beta)'(\overline{Y}_i - \overline{X}_i\beta).$$

Now by using the Markov Chain Monte Carlo simulation method to equation (8), we can obtain simulated copies of the micropopulation data for the i^{th} small area.

Figure 6 A diagram of a prospective tool for generating spatial microdata



The main steps involved with this process of spatial microdata simulation are as follows:

1. obtain a suitable joint prior distribution of the event under research E_i, say housing stress in the population at i^{th}small area, that is $p(E_i)$ for $\forall_i$;
2. find the conditional distribution of unobserved sampling units, given the observed data, that is $p(y_{ij} : j \in \bar{s}_i \mid y_{ij} : j \in s_i)$ for $\forall_i$;
3. derive the posterior distribution using Bayes theorem, that is $p(\theta \mid s, X)$; $E_i \subseteq \theta$, where θ is the vector of model parameters and X is an auxiliary information vector; and
4. get simulated copies of the entire population from this posterior distribution by the MCMC simulation technique.

The key feature of this new method is that it can simulate complete scenarios of the whole micro-population in a small area, which means it can produce more reliable small area estimates and their variance estimation. It is also able to create the statistical reliability measures (for example, the Bayes credible region or confidence interval) of spatial microsimulation models' estimates that are still unavailable in the literature.

However, in the new approach to find a suitable prior distribution for each interested event, as well as the appropriate model for linking between observed data and unobserved sampling units are difficult in practice.

5. CONCLUSIONS

This paper has briefly summarised the overall methods of small area estimation and explicitly described some methodological issues in the spatial microsimulation modelling arena. Review papers in small area estimation literature have regularly focused only methodologies on various statistical approaches including the area level and unit level modelling with E-BLUP, EB and HB methods. However, spatial microsimulation modelling has also been widely used in small area estimation, and recently classified as the geographic approach. Simulating a reliable synthetic spatial microdata is the key challenge in the indirect geographic approach of small area estimation. The review of different methodologies demonstrates that two reweighting methods – the GREGWT and CO are commonly used tools to produce small area microdata.

The GREGWT technique utilizes a truncated Chi-squared distance function and generates a set of new weights by minimising the total distance with respect to some constraint functions. The minimisation tool Lagrange multipliers has been used in the GREGWT process to minimise the distance function and it is based on the Newton-Raphson iterative process. Results show that sets of new weights can vary substantially with changing values of the vector of difference between the benchmark totals and sample based estimated totals. On the other hand, the combinatorial optimisation technique uses an intelligent searching algorithm *simulated annealing* – which selects an appropriate set of households' from survey microdata that best fits to the benchmark constraints by minimising the total absolute error/distance with respect to the *Metropolis Criterion*. The new weights give the actual household units, which are the best representative combination. Thus, CO is a selection process to reach an appropriate combination of sample units rather than calibrating the sampling design weights to a set of new weights.

Findings reveal that the GREGWT and CO are using quite different iterative algorithms and their properties also vary, but their performances are fairly similar from the standpoint of use in SMM. The Chi-squared distance measures show more smooth fluctuations than the absolute distance measures. Besides, SMMs techniques are robust and have significant advantages. In particular, since the spatial microsimulation framework uses a list-based approach to microdata representation, it is possible to use the microdata file for further analysis and updating. Also, by linking spatial microsimulation models with static microsimulation models, it is possible to measure small area effects of policy changes.

Moreover, the study points out a new approach in the spatial microsimulation methodology. The new technique is based on the Bayesian prediction theory and can simulate complete scenarios of the whole population in each small area. As a result the process can yield more accurate and statistically reliable small area estimates compared to the estimates from the other reweighting techniques. Besides, the Bayesian prediction based microdata simulation is a probabilistic approach, which is quite different from the deterministic approach used in GREGWT and the intelligent searching tool *simulated annealing* used in CO. However, the new approach can use the generalised regression model operated in the GREGWT algorithm to link observed units in the sample and unobserved units in the population. In contrast, from the view point of the CO reweighting, it uses the MCMC simulation with a posterior density based iterative algorithm. Further account of the new approach and its practical applications on empirical data will be presented in our next manuscript. Future research may look into this option of methodological advancements by practicing it into all arenas of SMMs.

Acknowledgements

The authors are grateful to the editors and two anonymous referees for their stimulus suggestions and valuable comments. This paper comes from a part of the doctoral thesis of the first author. We would also like to acknowledge the PhD awards from *an Endeavour- International Postgraduate Research Scholarship* provided by the Commonwealth of Australia*, the ACT-LDA Postgraduate Research Scholarship* provided by an agency of the ACT Government and the AHURI, *the NATSEM Top-Up Scholarship* from the National Centre for Social and Economic Modelling at the University of Canberra, and the "Regional Dimensions" Australian Research Council Linkage Project (LP 775396).

REFERENCES

ABS 2002, The 1998-99 Household Expenditure Survey, Australia: Confidentialised Unit Record Files (CURF), Technical Manual (2nd ed.), cat. no. 6544.0, Canberra, Australian Bureau of Statistics.

ABS 2004, Statistical Matching of the HES and NHS: An Exploration of Issues in the use of Unconstrained and Constrained Approaches in Creating a Basefile for a Microsimulation Model of the Pharmaceutical Benefits Scheme, Canberra, Australian Bureau of Statistics.

Alegre, J., Arcarons, J., Calonge, S. and Manresa, A. 2000, Statistical matching between different datasets: An application to the Spanish household survey (EPF90) and the income tax file (IRPF90), http://selene.uab.es/mmercader/workshop/cuerpo.html, Accessed 15 April 2008.

Anderson, B. 2007, Creating small-area Income Estimates: spatial microsimulation modelling, http://www.communities.gov.uk/publications/

communities/creatingsmallareaincome, Accessed 3 April 2008.

Ballas, D., Clarke, G. and Turton, I. 1999, 'Exploring microsimulation methodologies for the estimation of household attributes', *paper presented at the 4th International conference on GeoComputation*, Virginia, USA, July 25-28.

Ballas, D. 2001, *A spatial microsimulation approach to local labour market policy analysis*, unpublished PhD thesis, School of Geography, University of Leeds, UK.

Ballas, D., Clarke, G.P. and Turton, I. 2003, 'A spatial microsimulation model for social policy evaluation' in B. Boots and R. Thomas, (eds), *Modelling Geographical Systems*, Kluwer, Netherlands, vol. 70, pp. 143-168.

Ballas, D., Rossiter, D., Thomas, B., Clarke, G.P. and Dorling, D. 2005, Geography Matters: Simulating the local Impacts of National Social Policies, York, Joseph Rewntree Foundation.

Ballas, D., Clarke, G. and Dewhurst, j. 2006, 'Modelling the socio-economic impacts of major job loss or gain at the local level: a spatial microsimulation framework', *Spatial Economic Analysis*, vol. 1, no. 1, pp. 127-146.

Bell, P. 2000, GREGWT and TABLE macros - User guide, ABS, Canberra, unpublished.

Bell, P. 2000a, Weighting and standard error estimation for ABS Household Surveys, Canberra, Australian Bureau of Statistics.

Birkin, M. and Clarke, M. 1988, 'SYNTHESIS- a synthetic spatial information system for urban and regional analysis: methods and examples', *Environment and Planning Analysis*, vol. 20, pp. 1645-1671.

Brown, L. and Harding, A. 2005, 'The new frontier of health and aged care: using microsimulation to assess policy options', Tools for Microeconomic Policy Analysis, Productivity Commission, Canberra.

Chin, S.F. and Harding, A. 2007, 'SpatialMSM' in A. Gupta and A. Harding, (eds), *Modelling our future: population ageing, health and aged care*, Amsterdam, North-Holland.

Chin, S.F., Harding, A., Lloyd, R., McNamara, J., Phillips, B. and Vu, Q.N. 2005, 'Spatial microsimulation using synthetic small area estimates of income, tax and social security benefits', *Australasian Journal of Regional Studies*, vol. 11, no. 3, pp. 303-335.

Chin, S.F. and Harding, A. 2006, Regional Dimensions: Creating Synthetic Small-area Microdata and Spatial Microsimulation Models, *Online Technical Paper - TP33*, NATSEM, University of Canberra.

Clarke, M. and Holm, E. 1987, 'Microsimulation methods in spatial analysis and planning', *Geografiska Annaler, Series B*, Human Geography, vol. 69, no. 2, pp. 145-164.

Cullinan, J., Hynes, S. and O'Donoghue, C. 2006, 'The use of spatial microsimulation and geographic information systems (GIS) in benefit function transfer - an application to modelling the demand for recreational activities in Ireland', *paper presented at the 8th Nordic Seminar on Microsimulation models*, Oslo, June 7-9.

Deming, W.E. and Stephan, F.F. 1940, 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *The Annals of Mathematical Statistics*, vol. 11, no. 4, pp. 427-444.

Duley, C.J. 1989, *A Model for Updating Census-Based Population and Household Information for Inter-Censal Years*, School of Geography, University of Leeds, UK.

Ericson, W.A. 1969, 'Subjective Bayesian models in sampling finite populations ', *Journal of the Royal Statistical Society. Series B*, vol. 31, no. 2, pp. 195-233.

Evans, S.P. and Kirby, H.R. 1974, 'A three dimensional furness procedure for calibarating gravity models', *Transportation Research*, vol. 8, pp. 105-122.

Fienberg, S.E. 1970, 'An iterative procedure for estimation in contingency tables', *The Annals of Mathematical Statistics*, vol. 41, pp. 907-917.

Ghosh, M. and Rao, J.N.K. 1994, 'Small area estimation: an appraisal', *Statistical Science*, vol. 9, no. 1, pp. 55-93.

Harding, A. 1993, Lifetime income distribution and redistribution: applications of a microsimulation model, Amsterdam, North-Holland.

Harding, A., (ed.). 1996, *Microsimulation and public policy, Contributions to economic analysis*, Amsterdam, North-Holland.

Harding, A., Lloyd, R., Bill, A. and King, A. 2003, 'Assessing poverty and Inequality at a detailed regional level: new advances in spatial microsimulation' in M. McGillivray and M. Clarke, (eds), *Understanding Human Well-Being*. Helsinki, United Nation University Press, vol. 1, pp. 239-261.

Harding, A. and Gupta, A., (eds), 2007, *Modelling our future: population aging, social security and taxation, International symposia in economic theory and econometrics*, Amsterdam, Elsevier.

Harding, A., Vu, Q.N., Tanton, R. and Vidyattama, Y. 2009, 'Improving work incentives and incomes for parents: the national and geographic impact of liberalising the family tax benefit income test', *The Economic Record*, vol. 85, no. SI, pp. 48 – 58

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

Huang, Z. and Williamson, P. 2001, A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata, *Working Paper 2001/2*, Population Microdata Unit, Department of Geography, University of Liverpool, UK.

King, A. 2007, 'Providing income support services to a changing aged population in Australia: Centrelink's Regional Microsimulation model' in A. Gupta and A. Harding, (eds), *Modelling our future: population ageing, health and aged care*, Amsterdam, North-Holland.

Kirkpatrick, S., Gelatt Jr., C.D. and Vecchi, M.P. 1983, 'Optimization by Simulated Annealing', *Science*, vol. 220, no. 4598, pp. 671-680.

Little, R. 2007, An objective Bayesian view of survey weights, O'Bayes 07, http://3w.eco.uniroma1.it/OB07/papers/little.ppt, Accessed 27 June 2008.

Liu, T.P. and Kovacevic, M.S. 1997, 'An empirical study on categorically constrained matching', *Proceedings of the Survey Methods Section*, Canada, Statistical Society of Canada.

Lo, A.Y. 1986, 'Bayesian statistical inference for sampling a finite population', *The Annals of Statistics*, vol. 14, no. 3, pp. 1226-1233.

Lymer, S., Brown, L., Harding, A., Yap, M., Chin, S.F. and Leicester, S. 2006, Development of CareMod/05, *Online Technical Paper - TP32*, NATSEM, University of Canberra.

Lymer, S., Brown, L., Yap, M. and Harding, A. 2008, 'Regional disability estimates for New South Wales in 2001 using spatial microsimulation', *Applied Spatial Analysis and Policy*, vol. 1, pp. 99-116.

Meeden, G. 2003, 'A noninformative Bayesian approach to small area estimation', *Survey Methodology*, vol. 29, no. 1, pp. 19-24.

Merz, J. 1991, 'Microsimulation- a survey of principles, developments and applications', *International Journal of Forecasting*, vol. 7, pp. 77-104.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. 1953, 'Equation of state calculations by fast computing machines', *Journal of Chemical Physics*, vol. 21, pp. 1087-1092.

Moriarity, C. and Scheuren, F. 2001, 'Statistical matching: A paradigm for assessing the uncertainty in the procedure', *Journal of Official Statistics*, vol. 17, no. 3, pp. 407-422.

Moriarity, C. and Scheuren, F. 2003, 'A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations', *Journal of Business and Educational Studies*, vol. 21, no. 1, pp. 65-73.

Norman, P. 1999, Putting iterative proportional fitting on the researcher's desk, *WP 99/03*, School of Geography, University of Leeds, UK.

Orcutt, H.G. 2007 , 'A new type of socio-economic system', Reprinted with permission in the *International Journal of Microsimulation*, Vol 1, Autumn (available from www.microsimulation.org/IJM), originally published in 1957 in *Review of Economics and Statistics*, vol. 39, no. 2, pp. 116-123.

Pfeffermann, D. 2002, 'Small area estimation - new developments and directions', International Statistical Review, vol. 70, no. 1, pp. 125-143.

Pham, D.T. and Karaboga, D. 2000, *Intelligent optimisation techniques: genetic algorithms, tabu search, simulated annealing and neural networks*, London, Springer.

Rahman, A. 2008a, A review of small area estimation problems and methodological developments, *Online Discussion Paper - DP66*, NATSEM, University of Canberra.

Rahman, A. 2008b, 'The possibility of using Bayesian prediction theory in small area estimation', *presentation to the ARCRNSISS/ANZRSAI Annual Conference*, Adelaide, Nov. 30 to Dec. 03.

Rahman, A. 2008c, *Bayesian Predictive Inference for Some Linear Models under Student-t Errors*, Saarbrucken, VDM Verlag.

Rao, J.N.K. 1999, 'Some current trends in sample survey theory and methods (with discussion)', *Sankhya: The Indian Journal of Statistics; Series B*, vol. 61, no. 1, pp. 1-57.

Rao, J.N.K. 2002, 'Small area estimation: update with appraisal' in N. Balakrishnan, (ed.) *Advances on Methodological and Applied Aspects of Probability and Statistics*, New York, Taylor and Francis, pp. 113-139.

Rao, J.N.K. 2003, *Small Area Estimation*, New Jersey, John Wiley and Sons, Inc.

Rao, J.N.K. 2003a, 'Some new developments in small area estimation', *JIRSS*, vol. 2, no. 2, pp. 145-169.

Rassler, S. 2002, *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Verlag, Springer.

Rassler, S. 2004, 'Data fusion: identification problems, validity, and multiple imputation', *Austrian Journal of Statistics*, vol. 33, no. 2, pp. 153-171.

Rodgers, W.L. 1984, An evaluation of statistical matching, *Journal of Business and Economic Statistics*, vol. 2, no. 1, pp. 91-102.

Simpson, L. and Tranmer, M. 2005, 'Combining sample and census data in small area estimates: iterative proportional fitting with standard software', *The Professional Geographer*, vol. 57, no. 2, pp. 222-234.

Tanton, R. 2007, 'SPATIALMSM: The Australian spatial microsimulation model', *The 1st General Conference of the International Microsimulation Association*, Vienna, August 20-21.

Tanton, R., Williamson, P. and Harding, A. 2007, 'Comparing two methods of reweighting a survey file to small area data - Generalised regression and Combinatorial optimisation', *The 1st General Conference of the International Microsimulation Association*, Vienna, August 20-22.

Taylor, E., Harding, A., Lloyd, R. and Blake, M. 2004, 'Housing unaffordability at the statistical local area level: new estimates using spatial microsimulation', *Australian Journal of regional Studies*, vol. 10, no. 3, pp. 279-300.

Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M., Martin, D., Steel, D. and Gardiner, C. 2005, 'The case for small area microdata', *Journal of the Royal Statistical Society: Series A*, vol. 168, no. 1, pp. 29-49.

van Laarhoven, P.J. and Aarts, E.H. 1987, *Simulated Annealing: Theory and Applications*, NY, Springer.

Voas, D. and Williamson, P. 2000, 'An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata', *International Journal of Population Geography*, vol. 6, pp. 349-366.

Williamson, P. 1992, *Community Health Care Policies for the Elderly: A Microsimulation Approach*, School of Geography, University of Leeds, UK.

Williamson, P., Clarke, G.P. and McDonald, A.T. 1996, 'Estimating small area demands for water with the use of microsimulation' in G. P. Clarke, (ed.) *Microsimulation for Urban and Regional Policy Analysis*, Pion, London.

Williamson, P., Birkin, M. and Rees, P. 1998, 'The estimation of population microdata by using data from small area statistics and sample of anonymised records', *Environment and Planning Analysis*, vol. 30, pp. 785-816.

Williamson, P. 2007, CO Instruction Manual, *Working Paper 2007/1*, Population Microdara Unit, Department of Geography, University of Liverpool, UK.

Wong, D.W.S. 1992, 'The Reliability of Using the Iterative Proportional Fitting Procedure', *The Professional Geographer*, vol. 44, no. 3, pp. 340.

Appendix A: The Newton-Raphson iteration method

The Newton-Raphson iteration method is a root-finding algorithm for a nonlinear equation. The method is based on the first few terms of the Taylor series of a function. Let for a single variable nonlinear equation $f(z)=0$, the Taylor series of $f(z)$ about the point $z=z_0+\varepsilon$ is expressed as

$$f(z_0+\varepsilon)=f(z_0)+f'(z_0)\varepsilon+f''(z_0)\varepsilon^2+\ldots \qquad \text{(a1)}$$

Where z_0 is an initial assumed root of $f(z)$, f' represents the first order derivative and ε is a small arbitrary positive quantity. Keeping terms only to first order derivative, we have

$$f(z_0+\varepsilon)\approx f(z_0)+f'(z_0)\varepsilon. \qquad \text{(a2)}$$

Now (a2) is the equation of tangent line to the curve of $f(z)$ at the point $\{z_0,f(z_0)\}$, and hence $(0,z_0)$ is the interval where that tangent line intersects the horizontal axis at z_1 (Fig. a-1).

The expression in (a2) can be used to estimate the amount of adjustment for ε should require to converge to the accepted root starting from an initial assumed root value, z_0. From the relation in (a2), after setting $f(z_0+\varepsilon)=0$ and considering an arbitrary quantity $\varepsilon=\varepsilon_0$ we get

$$\varepsilon_0=-\frac{f(z_0)}{f'(z_0)},$$

which is the first-order adjustment to the original root.

Now by considering $z_i = z_{i-1} + \varepsilon_{i-1}$ for $i = 1,2,\ldots,r$, we can subsequently obtain a new ε_i, for which

$$\varepsilon_i = -\frac{f(z_i)}{f'(z_i)};\forall i\,. \qquad \text{(a3)}$$

Let the process should be repeated until $(r+1)$ times when a value of the arbitrary quantity, ε is reached to the accuracy level. In other words, the process should be repeated until $(r+1)$ times when an estimated root of the function – (say) z_{r+1}, will converge to a precisely stable number or to an accepted root value. Hence the following algorithm can be applied iteratively to obtain an accepted root

$$z_{r+1}=z_r-\{f'(z_r)\}^{-1}f(z_r);\forall r=1,2,3,\ldots. \qquad \text{(a4)}$$

The method uses this iterative equation in (a4) to approach one root of a function. A well-chosen initial root value can lead the convergence quickly (Fig. a-1). However the procedure can be unstable near a horizontal asymptote or a local extremum. Besides, this iteration method is easily adapted to deal with a set of equations for a function with vector variables when its second order derivative also exists.

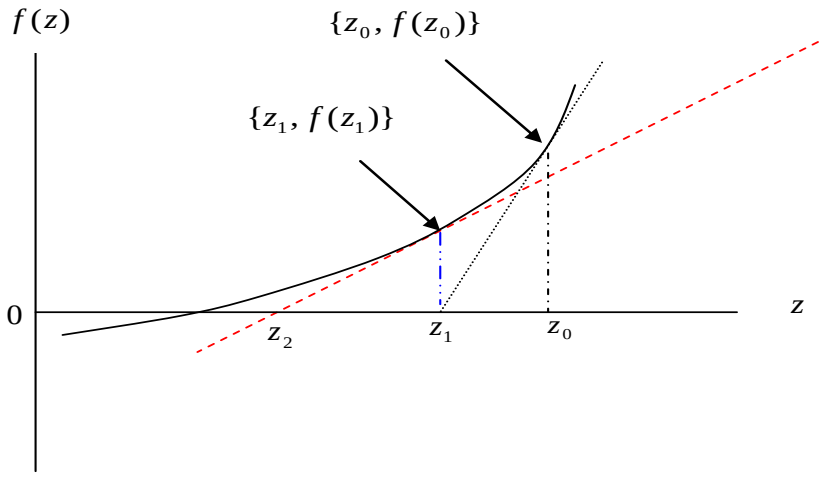


Figure a-1 Graphical view of the Newton-Raphson iteration process

Now equation (6) in GREGWT theory can be written as a function of the vector λ –

$$l_j(\lambda) = C_j - \sum_{k \in s} d_k \left\{ f^{-1}(x_k' \lambda) - 1 \right\} x_{k,j} = 0 \qquad \text{(a5)}$$

for $j = 1,2,\ldots, p$;

where

$C = T_x - \hat{t}_{x,s}$ is a known vector,

$d_k \left\{ f^{-1}(x_k' \lambda) - 1 \right\}$ is a scalar,

and the equation is nonlinear in the Lagrange multipliers vector, λ.

The equation (a5) can be solved by the above Newton-Raphson iterative procedure. Hence the iteration algorithm can be expressed as

$$\lambda_{[r+1]} = \lambda_{[r]} - \left[l'(\lambda) \right]^{-1}_{\lambda_{[r]}} \left[l(\lambda) \right]_{\lambda_{[r]}} ; \forall r = 1,2,3,\ldots \qquad \text{(a6)}$$

where

$\lambda_{[r]}$ is the value of the vector λ in the r^{th} iteration,

$l'(\lambda) = [\partial l_j(\lambda) / \partial \lambda_h]$

represents the Hessian matrix,

and

$[l'(\lambda)]_{[r]}$

defines the values of vector $l'(\lambda)$, which are determined by the r^{th} iteration values of vector $\lambda_{[r]}$.

Note that GREGWT stops iteration process when the condition

$\left| \lambda_{[r+1]} - \lambda_{[r]} \right| < \varepsilon_r = 0.0001$ is satisfied

or when a predefined maximum iteration has been reached.

However, the ε_r can take any suitable positive arbitrary value and the choice is fully depending on our desired accuracy.