



Small Area Estimation Through Spatial Microsimulation Models:

SOME METHODOLOGICAL ISSUES

Paper Presented at the 2nd
International Microsimulation
Association Conference, Ottawa,
Canada, 8-10 June 2009

PREPARED BY

Azizur Rahman

Azizur.Rahman@natsem.canberra.edu.au

10 JUNE 2009

ABOUT NATSEM

The National Centre for Social and Economic Modelling was established on 1 January 1993, and supports its activities through research grants, commissioned research and longer term contracts for model maintenance and development.

NATSEM aims to be a key contributor to social and economic policy debate and analysis by developing models of the highest quality, undertaking independent and impartial research, and supplying valued consultancy services.

Policy changes often have to be made without sufficient information about either the current environment or the consequences of change. NATSEM specialises in analysing data and producing models so that decision makers have the best possible quantitative information on which to base their decisions.

NATSEM has an international reputation as a centre of excellence for analysing microdata and constructing microsimulation models. Such data and models commence with the records of real (but unidentifiable) Australians. Analysis typically begins by looking at either the characteristics or the impact of a policy change on an individual household, building up to the bigger picture by looking at many individual cases through the use of large datasets.

It must be emphasised that NATSEM does not have views on policy. All opinions are the authors' own and are not necessarily shared by NATSEM.

Director: Ann Harding

© NATSEM, University of Canberra 2009

All rights reserved. Apart from fair dealing for the purposes of research or private study, or criticism or review, as permitted under *the Copyright Act 1968*, no part of this publication may be reproduced, stored or transmitted in any form or by any means without the prior permission in writing of the publisher.

National Centre for Social and Economic Modelling
University of Canberra ACT 2601 Australia
170 Haydon Drive Bruce ACT 2617

Phone + 61 2 6201 2780

Fax + 61 2 6201 2751

Email natsem@natsem.canberra.edu.au

Website www.natsem.canberra.edu.au

CONTENTS

| | |
|---|-----------|
| About NATSEM | 2 |
| List of figures | 4 |
| List of tables | 4 |
| Author note | 5 |
| Acknowledgements | 5 |
| General caveat | 5 |
| Abstract | 6 |
| 1 Introduction | 7 |
| 2 Methods of small area estimation: A brief review | 8 |
| 2.1 A diagram of overall methodologies | 8 |
| 2.2 Direct estimation methods | 10 |
| 2.3 Statistical modelling for indirect estimation | 11 |
| 2.4 SMM estimation | 15 |
| 3 Methodological issues in SMM | 18 |
| 3.1 Creation of synthetic spatial microdata | 18 |
| 3.2 Reweighting: The GREGWT approach | 21 |
| 3.2.1 <i>How does GREGWT generate new weights? A theoretical view</i> | 23 |
| 3.2.2 <i>Explicit numerical solution for a hypothetical data</i> | 25 |
| 3.3 Combinatorial optimisation reweighting approach | 29 |
| 3.3.1 <i>The Simulated Annealing Method in CO</i> | 30 |
| 3.3.2 <i>An illustration of CO process for hypothetical data</i> | 33 |
| 3.4 A comparison between GREGWT and CO | 35 |
| 4 Some new possibilities in the SMM methodologies | 37 |
| 4.1 Reweighting via the Bayesian prediction theory: A prospective tool | 37 |
| 4.2 Statistical reliability measures of SMM estimates | 38 |
| 4.2.1 <i>Test statistic: A way to test statistical significance of SMMs estimates</i> | 39 |
| 4.2.2 <i>Confidence interval estimation for SMMs estimates</i> | 39 |
| 5 Conclusions | 40 |
| References | 42 |
| Appendix A: The Newton- Raphson iteration method | 48 |

LIST OF FIGURES

| | |
|---|----|
| Figure 2-1: A summary of different techniques for small area estimation | 9 |
| Figure 3-1: A comparison of absolute distance and Chi-squared distance measures | 28 |
| Figure 3-2: Plots of sampling design weights and new weights for specific cases | 28 |
| Figure 3-3: A flowchart of the simulated annealing algorithm | 32 |
| Figure 3-4: A simplified combinatorial optimisation process | 33 |
| Figure 4-1: A diagram of prospective tool for generating spatial microdata | 37 |
| Figure a-1: Graphical representation of the Newton-Raphson iteration process | 48 |

LIST OF TABLES

| | |
|---|----|
| Table 3-1: Synthetic reconstruction <i>versus</i> the reweighting technique | 20 |
| Table 3-2: New weights and its distance measures to sampling design weights | 27 |
| Table 3-3: A comparison of the GREGWT and CO reweighting methodologies | 36 |

AUTHOR NOTE

Azizur Rahman is a doctoral candidate and research assistant at the National Centre for Social and Economic Modelling at the University of Canberra, Australia.

ACKNOWLEDGEMENTS

The author would like to gratefully acknowledge the funding from an E-IPRS award provided by the *Commonwealth of Australia* through the University of Canberra and the ACT-LDA Postgraduate Research Scholarship provided by the *Australian Capital Territory Land Development Agency* through the AHURI/RMIT-NATSEM housing research centre at the National Centre for Social and Economic Modelling, University of Canberra. Acknowledgements also go to Professor Ann Harding, Dr Shuangzhe Liu and Robert Tanton for their valuable comments on earlier versions of this manuscript. Finally the author heartily thanks the supports and stimulus obtained from Professor Ann Harding, as well as the conference funding provided by the NATSEM/University of Canberra.

GENERAL CAVEAT

NATSEM research findings are generally based on estimated characteristics of the population. Such estimates are usually derived from the application of microsimulation modelling techniques to microdata based on sample surveys.

These estimates may be different from the actual characteristics of the population because of sampling and nonsampling errors in the microdata and because of the assumptions underlying the modelling techniques.

The microdata do not contain any information that enables identification of the individuals or families to which they refer.

ABSTRACT

Small area estimation has received much attention in recent decades due to increasing demand for reliable small area estimates from both public and private sectors. Traditional direct estimation requires the domain-specific sufficiently large sample. But, in reality, domain-specific sample data are usually not large enough for all small areas (even zero for some small areas) to provide adequate statistical precision of their estimates. This makes it necessary to “borrow strength” from data on related multiple characteristics and/or auxiliary variables from other neighbouring areas through appropriate models, leading to indirect or model-based estimates. This paper describes some vital methodological issues of spatial microsimulation modelling for small area estimation, with a particular emphasis given to the reweighting techniques.

Most of the review articles in small area estimation have highlighted methodologies known as “statistical approaches” - which are based on various statistical models and theories. However another type of technique called “spatial microsimulation models” has also provided small area estimates during the last decade. These models are based on economic theory and using quite different methodologies. A thorough overview on various microsimulation models shows that spatial microsimulation models are robust and have advantages over others. In contrast to statistical approaches, the spatial microsimulation model-based approaches can operate through different reweighting techniques such as GREGWT and combinatorial optimization. A comparison between reweighting techniques reveals that they are using quite different algorithms and that their properties also vary. However their performances are fairly similar according to the advantages of spatial microsimulation modelling. Finally the study points out some new possibilities in the spatial microsimulation methodology.

Key words

Combinatorial optimisation; GREGWT; methodologies; prediction theory; reweighting techniques; small area estimation; small area models; spatial microsimulation model; synthetic microdata; test statistic.

1 INTRODUCTION

Small area estimation is the method of estimating reliable statistics at small geographical area or a spatial micropopulation unit. The reliable statistics of an 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 not available for all small areas from the 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 level 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 out of the sample areas (Tanton 2007). However, depending on the type of the study and the time and money constraints, it can be also impossible to conduct a sufficiently comprehensive sample survey to obtain an adequate sample from every small area we are interested in.

Nowadays indirect modelling approaches of small area estimation such as spatial microsimulation modelling (SMM) have received much attention due to its usefulness and the increasing demand for reliable small area statistics from both the 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 which 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-level 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.

Small area estimation methodologies are beneficial for business organisations, policy makers and researchers who are interested in estimates for regional small domains but who lack adequate funds for a large-scale survey that could produce precise, direct survey estimates for the small domains. For instance, population estimates of a small area may be used in a range of purposes, such as business organisations using them to develop profiles of customers, to identify market clusters and to determine optimal site locations for their business. In addition, state and local governments use them to establish political boundaries, to monitor the impact of public policies and to estimate the need for schools, roads, parks, public transportation and fire protection. Also, researchers use them to study urban sprawl, environmental conditions and social trends. Such estimates are used as denominators for calculating many types of rates and to determine the allocation of money from public funds each year (Smith et al. 2002). Therefore it is clear that the inference of precise small area estimates is of significance for many reasons.

Most of the review articles in small area estimation have highlighted the methodologies which are fully based on various statistical models and theories (see 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 (see 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 others. 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 interest should require improving the performance of SMMs and validation of estimates. This paper briefly reviews the methodologies of small area estimation and explicitly describes some vital methodological issues of spatial microsimulation modelling, with a particular emphasis given to the reweighting techniques. It also proposes some possibilities for new advancements in methodologies.

There are 5 sections within this paper. Section 2 provides a brief review of the methods of small area estimation, with a diagramic representation of overall methodologies. Then some important methodological issues in spatial microsimulation modelling are described in Section 3. This section includes a detailed description of different reweighting techniques, with an explicit numerical solution for hypothetical data. A comparison between two reweighting techniques is also given in Section 3. Some new possibilities in methodologies for spatial microsimulation modelling are introduced in Section 4. Finally, Conclusions are given in Section 5.

2 METHODS OF SMALL AREA ESTIMATION: A BRIEF REVIEW

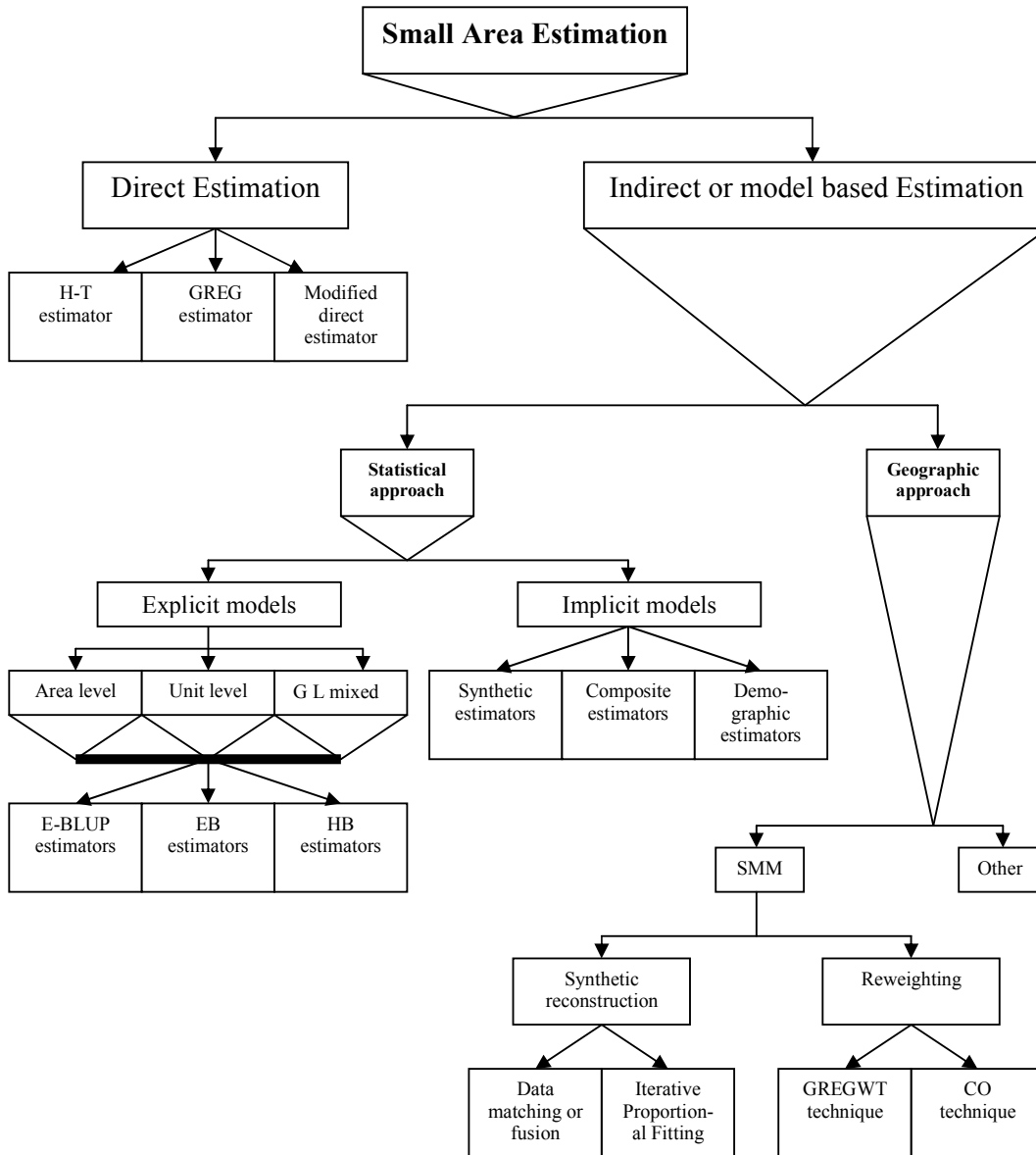
2.1 A DIAGRAM OF OVERALL METHODOLOGIES

Traditionally there are two types of small area estimation – direct and indirect estimation. The direct small area estimation is based on survey design and includes three estimators called the H-T 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 microsimulation modelling. A summary diagram of overall methodologies for small area estimation is depicted in Figure 2-1.

Notice that implicit model based approaches include three types of estimations, which are synthetic, composite and demographic estimations; whereas explicit models are categorized as area level, unit level and general linear mixed models. Based on the type of

study of interest, each of these models is widely studied to obtain small area indirect estimates using the (empirical-) best linear unbiased prediction (E-BLUP), empirical Bayes (EB) and hierarchical Bayes (HB) methods.

Figure 2-1: A summary of different techniques for small area estimation



(after Rahman 2008a)

On the other hand the geographic approach is based on microsimulation models, which are essentially creating synthetic/simulated micro-population data to produce 'simulated estimates'. To obtain reliable microdata at small area level is the key task for spatial microsimulation modelling. Synthetic reconstruction and reweighting are commonly used methods in microsimulation, and each of them is stimulated by different techniques to produce simulated estimators (see Figure 2-1). A brief discussion of different methods of small area estimation is given in the following sections.

2.2 DIRECT ESTIMATION METHODS

Direct estimators only rely on the sample obtained from the survey. For direct estimation, all small areas must be sampled in order to be able to produce these kinds of estimates. Although it is rare, when survey samples are large enough to cover all the study areas with sufficient data in each area, different direct estimators can be developed. A brief description of direct small area estimation under the sample design based methods and the model based methods has been given by Rao (2003). Note that in a model-based approach inferences are involved with statistical models, whereas in a design-based approach inferences are fully based on sampling design.

The common direct estimators include the *Horvitz-Thompson estimator*, *generalised regression estimator*, *modified direct estimator* or *survey regression estimator* etc. The definitions and a detail accounts of these estimators have been given elsewhere (see, for example Rahman 2008a). Some of the basic properties of these direct estimators are summarized below.

In principle the Horvitz-Thompson estimator is not designed to use auxiliary information or covariates. However it is possible to consider auxiliary information to evaluate this estimator (see, Sarndal et al. 1992). When the inclusion probability of a sampling unit is positive and there is sufficient sample observations available at small area, the Horvitz-Thompson estimator is unbiased, but not efficient. However in real context of small area estimation, with an inadequate sample, this estimator can be biased and more unreliable.

Besides the generalised regression estimator is approximately design-unbiased for small area estimation but not consistent because of high residuals. When only area specific auxiliary information is available this estimator is model-unbiased under the linearity assumption. As well this estimator can take a more general form to provide estimates of all target variables under different assumptions and at different domains - and hence this estimator ensures consistency of results at different areas when aggregated over different variables of interest (see, Rao 2003). However the generalised regression estimator could be negative in some small areas when the linear regression overestimates the variable of interest. In addition this estimator does not ensure consistency with the estimate of target variable at aggregated level while it ensures consistency with the covariates totals.

Furthermore the modified direct estimator or survey regression estimator is approximately design-unbiased as the overall sample size increases, even if the regional sample size is small. This estimator uses the overall sample data at the aggregated level to calculate the overall regression coefficients. It is remarkable that although the modified direct estimator

borrow strength for estimating the overall regression coefficients, it does not increase the effective sample size, unlike indirect small area estimators (for an example see, Rao 2003).

In the context of small area estimation, direct estimators lead to unacceptably large standard errors because of disproportionately small samples from the small area of interest; in fact, no sample units may be selected from some small domains. Also the model based direct estimation approaches can perform poorly by model misspecification for increasing sample size (Rao 2003). This poor performance is mainly caused by asymptotic design inconsistency of the model based estimator with respect to the stratified random sampling. Therefore, direct small area estimators typically have the following limitations:

- estimates can only be computed for a subset of all areas which contain respondents to the sample survey;
- for those small sampled areas the achieved sample size will usually be very small or not large enough and the estimator will thus have low precision - and this low precision will be reflected in rather wide confidence intervals for the direct estimates, thus making them statistically unreliable; and
- direct estimates can perform poorly by model misspecification and asymptotic design inconsistency of the sampling structure.

2.3 STATISTICAL MODELLING FOR INDIRECT ESTIMATION

Indirect or model-based small area estimators rely on different statistical or geographic models to provide estimates for all small areas. Once the model is chosen, its parameters are estimated using the data obtained in the survey. An important issue in indirect small area estimation is that additional auxiliary information or covariates are needed. The statistical approach of small area estimation mainly uses two types of statistical models: a) the implicit models and b) the explicit models. The implicit models provide a link to related small areas through supplementary data from census and/or administrative records; whereas the explicit models account for small area level variations through supplementary data and they are termed 'small area models' in the literature.

a) Implicit models approach:

This approach includes three statistical techniques of indirect estimation - which are synthetic, composite and demographic estimations. Detailed descriptions about each of these estimations is provided elsewhere (see, Rao 2003; Rahman 2008a). A brief summary is illustrated here.

The idea of synthetic estimation and its application was first introduced in the United States by the National Center for Health Statistics (1968). They had used this indirect estimation technique to calculate state level disability estimates. The synthetic estimators can be derived by partitioning the whole population (for example, national or state-wide data) into a series of mutually exclusive and exhaustive cells (for example, age, sex, ethnicity, income, etc.) and deriving the estimate as a sum of products. Gonzalez (1973)

provides an excellent definition of synthetic estimator – “an estimator should be synthetic when a reliable direct estimator for a large area is used to derive an indirect estimator for a small area belonging to the large area under the assumption that all small areas have the same characteristics as the large area”. In addition, Levy (1979) and Rao (2003) provide extensive overviews on various synthetic estimation approaches and its applications in small area estimation.

This type of estimator can be classified as a ratio-synthetic estimator and regression-synthetic estimator. Note that the synthetic estimator is essentially a biased estimator. Even so, when the small area does not exhibit strong individual effects with respect to the regression coefficients, the synthetic estimator will be efficient, with small mean squared error. Moreover the synthetic estimates are generally easy and inexpensive to obtain, since the independent variables are easily available from census or other administrative data and the regression coefficients are obtainable from national level surveys.

Besides composite estimation is a kind of balancing approach between the synthetic and direct estimators. It is rational that as the sample size in a small area increases, a direct estimator becomes more desirable than a synthetic estimator. This is true whether or not the surveys are designed to produce estimates for small areas. In other words, when area level sample sizes are relatively small, the synthetic estimator outperforms the traditional simple direct estimator - whereas when the sample sizes are large enough, the direct estimator outperforms the synthetic estimator. A weighted sum of these two estimators would be an alternative to choosing one over the other to balance their degree of bias, and this type of estimator is commonly known as a ‘composite estimator’.

According to Ghosh and Rao (1994), composite estimation is a natural way to balance the potential bias of a synthetic estimator against the instability of a direct estimator by choosing an appropriate weight. A number of the estimators proposed in the literature have the form of synthetic estimators - for example the James-Stein estimator and the shrinkage estimator. Rao (2003) provides an excellent account of composite estimation (with a brief review of the James-Stein methods) as well as examples of its practical applications in the context of small area estimation. Composite estimators are biased and they may have improved precision but depend on the selection of weight.

Moreover demographic estimation is another way to obtain indirect estimators based on implicit models. This approach mainly uses data from the recent census in conjunction with demographic information derived from various administrative record files. Three approaches are very popular in demographic estimation: the vital rates technique, component method and sample regression method. The general description about these techniques is given by Rao (2003) and also more references are therein. A summary of different demographic techniques in small area estimation is given in Rahman (2008a).

It is worth noting that the demographic approach is only used for population estimates. The Australian Bureau of Statistics has been used this method for their small area population estimates (for example, see ABS 1999). As the size and composition of the residents in a geographical area may change over time, postcensal or noncensal estimates

of population are essential for using a variety of purposes such as the determination of fund allocations, calculation of social and economic indicators (for example, vital rates, unemployment rates, poverty rates, etc. in which the population count serves as the denominator), calculation of survey weights etc. In demographic methods of small area estimation several regression symptomatic procedures, their properties and significant applications are discussed by Rao (2003). The advantages of demographic estimation are that it is an easy estimation and the theory behind this method is very simple and easily understandable. However underestimate of population count due to omission, duplication and misclassification in census is a big concern.

b) Explicit models approach:

This class of small area estimation approaches is mainly using different explicit models and in the literature it is termed 'small area models'. Available small area models can be classified as the basic area level models and the basic unit level models (for example, see Rao 1999). In the first type of models, information on the response variable is available only at the small area level, and in the second type of models, data are available at the unit or respondent level. A short summary of these small area models is given below.

At first to estimate the per capita income of small areas with the population size of less than 1000, Fay and Herriot (1979) used a two-level Bayesian model which is currently well-known as the Fay-Herriot model or basic area level mixed model. The Fay-Herriot model can be expressed as:

Linking model: $\theta_i \sim^{iid} N(x_i' \beta, \sigma_\varepsilon^2)$, where $i = 1, 2, \dots, n$;

Matching sampling model: $\hat{\theta}_i | \theta_i \sim^{iid} N(\theta_i, \omega_i^2)$, where $i = 1, 2, \dots, n$.

The first type of model is derived from area specific auxiliary data which are related to some suitable functions of the small area total to develop a linking model under normality assumptions with mean zero and variance σ_ε^2 . Then the linking model is combined with a matching sampling model to generate finally a linear mixed model. In this case the matching sampling model uses a direct estimator of the corresponding suitable function of the small area total and assumed normally distributed errors with mean zero and a known sampling variance ω_i^2 . However these two assumptions of the matching model have been considered as the limitations of the basic area level model (Rao 2003, 2003a). The authors argue that the assumption of known sampling variances for the matching sampling model is restrictive and the assumption of a zero mean may not be tenable if the small area sample size is very small and the relevant functional relationship is a nonlinear function of the small area total. The basic area level model has popularly used in the United States as a special case of the general linear mixed model. Note that the success of small area estimation through a basic area level model approach largely depends on getting good

auxiliary information (x_i) that leads to a small model variance σ_ε^2 relative to known or estimated ω_i^2 (see, Rao 1999).

On the other hand, the basic unit level model is based on unit level auxiliary variables. These are related to the unit level values of response through a nested error linear regression model, under the assumption that the nested error and the model error are independent to each other and normally distributed with common mean zero and common or different variances. For the unit responses y_{ij} , this type of model can be represented by the following mathematical equation

$$y_{ij} = x'_{ij}\beta + \varepsilon_i + e_{ij},$$

where x_{ij} represents unit-specific auxiliary data, which are available for areas $i = 1, 2, \dots, n$; and $j = 1, 2, \dots, N_i$ as N_i is the number of population units in the i^{th} area, and β represents the vector of regression parameters. The ε_i 's are normal, independent, and identically distributed with mean zero and variance σ_ε^2 . The e_{ij} 's are independent of ε_i 's and follow independent as well as identical normal distribution with mean zero and variance σ_e^2 . The abovementioned equation is known as a nested error regression model.

The nested error unit level regression model was first used to model county crop areas in the United States (Battese et al. 1988). This type of model is a particular case of the general linear mixed model which is appropriate for continuous value response variables. Various extensions of this type of model have been proposed in literature to handle the binary responses, two-stage sampling within areas, multivariate responses, two-level logistic regression model and others (see Rao 2003, 2003a). Nonetheless it is important to note that area level models have extensive scope in comparison to unit level models, because area level auxiliary data is more readily available than unit-specific auxiliary data (Rao 2003a).

A variety of approaches such as (empirical-) best linear unbiased prediction (E-BLUP), empirical Bayes (EB) and hierarchical Bayes (HB) are commonly used in explicit models based small area estimation. All of these methods are extensively discussed in the small area literature (see for example, Ghosh and Rao 1994; Pfeiffermann 2002; Rao 2003). Moreover an excellent summary of these methodologies and their comparison should be found in a recent manuscript (see, Rahman 2008d). The E-BLUP approach is applicable to the linear mixed models which are usually designed for continuous variables. But in practical fields there are many situations that require dealing with binary or count data where the E-BLUP method is not appropriate. Hence an advantage of the EB and HB approaches over E-BLUP is that they are applicable to linear mixed models, as well as models with binary or count data. In addition, under some conditions these approaches produce identical results. However they have quite different computational tools and techniques. For instance, in principle the EB approach is considered as a frequentist approach and does not depend on a prior distribution of the model parameters. In contrast, the HB approach essentially uses a prior distribution of model parameters and it can handle complex problems using the Markov Chain Monte Carlo (MCMC) technique. It is

worthwhile to mention that the techniques of maximum likelihood (ML), restricted or residual maximum likelihood (REML), penalized quasi-likelihood, etc. have been utilized for variance estimates of statistical model based estimators. Details of theoretical backgrounds for the estimation of parameters for different types of small area models are discussed by Rao (2003) and references therein. The EURAREA Consortium (2004) provides a thorough review of algorithms and computational tools and techniques for the estimation of parameters of small area explicit models.

Further discussion about of each of these statistical methods is beyond this paper, as the key objective of the paper was to address some methodological issues in spatial microsimulation modelling. However we may finally note that although these statistical approaches use data from different sources to obtain the estimators, they do not involve generating a base data file for the small area, that the SMMs do, and the base data file is a significant resource for various further analyses.

2.4 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 last two decades microsimulation 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 by policy makers (Harding and Gupta 2007) for small area estimates at government and private sector levels.

Microsimulation modelling was originally developed as a tool for economic policy analysis (see, Merz 1991). Clarke and Holm (1987) provide a thorough presentation on how microsimulation methods can be applied in regional science and planning analysis. In microsimulation models, researchers represent members of a population for the purpose of studying how individual behaviours generate aggregate results from the bottom up (Epstein 1999). This brings about a very natural instrument to anticipate trends in the environment through monitoring and early warning as well as to predict and value the short-term and long-term consequences of implementing certain policy measures (Saarloos 2006). 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). Description of different types of microsimulation models – such as *static*, *dynamic* and *spatial* microsimulation models is given in the literature (see, for example 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 reweighting method and, second, to apply these household weights to the selected output variables to generate small-domain estimates of the selected variables. Further, each of these major steps involve several sub-steps (for details see, 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 micro data set.

The starting point for microsimulation models is a microdata file, which provides comprehensive information on different characteristics for every individual persons, families or households on 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 adequate sample size to allow statistically reliable estimates for only large domains (such as only at the broad level of the state or territory). In practice, small area level 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 out of the sample areas (Tanton 2007). For example if a land development agency wants to develop a new housing 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, they need to be synthesized (Chin et al. 2005), and the lack of spatially explicit microdata has in the past constrained spatial microsimulation modelling of social policies and human behaviour.

The SMMs 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). 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 small area level;
- the models store data efficiently as lists;
- 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-orientated 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-orientated 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, see 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. 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). As noted earlier, 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. Thus spatial microdata are not usually obtainable; they need to be simulated, and that can be achieved by different probabilistic as well as deterministic methods.

3 METHODOLOGICAL ISSUES IN SMM

In spatial microsimulation modelling, to calculate 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 to create reliable spatial microdata. Complex methodologies are associated with the process of creating synthetic microdata. This section presents some of the significant 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 a 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 (Moriarty and Scheuren 2003; ABS 2004; Tranmer et al. 2005) and iterative proportional fitting (Birkin and Clarke 1988; Duley 1989; Williamson 1992; Norman 1999). However, the reweighting method is relatively new and popular method, which mainly calibrates the sampling design weights to a set of new weights based on a distance measure, and 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 (for example, 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 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 (see, Alegre et al. 2000; Rassler 2002). Details about data matching techniques are given by Rodgers (1984). Moreover the data matching tool is used to create microdata base files by researchers in many countries, such as Moriarity and Scheuren (2001, 2003) in the USA; Liu and Kovacevic (1997) in Canada; Alegre et al. (2000), Tranmer et al. (2001, 2005), Rassler (2004) in Europe; and ABS (2004) in Australia, among many others.

Besides, the iterative proportional fitting technique initially proposed by Deming and Stephan (1940) is mainly based on the methods of contingency table analysis and probability theory. The authors developed the method for adjusting cell frequencies in a contingency table based on sampled observations subject to known expected marginal totals. To create synthetic spatial microdata from a variety of aggregate data sources, such as census or administrative records this method is in use during several decades. The theoretical and practical considerations behind this method have been discussed in several studies (see, 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 iterative proportional fitting 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.

Note that 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 summary 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 (see for example, Williamson et al. 1998; Huang and Williamson 2001; Ballas et al. 2003). A summary of the remarks pointed out by researchers is shown in Table 3-1.

Table 3-1: **Synthetic reconstruction versus the reweighting technique**

| Synthetic reconstruction | Reweighting technique |
|---|---|
| <ul style="list-style-type: none"> ○ 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. ○ Ordering is essential in this process that means each value should be generated in a fixed order. ○ Relatively more complex and time consuming. ○ The effects of inconsistency between the constraining tables could be significant for this approach due to a mismatch in the table totals or subtotals. | <ul style="list-style-type: none"> ○ It is an iterative process – where a suitable fitting between actual data and the selected sample of microdata should be obtained by minimizing some sorts of distance errors. ○ Ordering is not an issue in this process. However converge is achievable by repeating the process many times or by some simple adjustment. ○ Although the technique is complex from theoretical point of view, it is comparatively less time consuming ○ Reweighting techniques can allow the choice of constraining tables to match with researcher and/or user requirements |

Moreover, reweighting is a procedure used throughout the world to transform information contained in a sample survey to estimates for the whole 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). For small area estimation to create a synthetic spatial microdata by reweighting methods, there are two widely used techniques: a generalised regression technique known as GREGWT (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). A detail of these two reweighting techniques is given in the following subsections.

3.2 REWEIGHTING: THE GREGWT APPROACH

The GREGWT approach has been developed by the ABS for Generalised Regression and Weighting of sample survey results. It is an iterative generalised regression algorithm written in SAS macros to calibrate survey estimates to benchmarks. Calibration can be looked at either as a way of improving estimates or as a way of making the estimates add up to benchmarks (Bell 2000). That is, the grossing factors or weights on a dataset containing the survey returns are modified so that certain estimates agree with externally provided totals known as benchmarks. This use of external or auxiliary information typically improves the resulting survey estimates that are produced using the modified grossing factors.

The GREGWT algorithm uses a constrained distance function known as the truncated Chi-squared distance function that is minimized subject to the calibration equations for each small area. The method is also known as linear truncated or restricted modified Chi-squared (see, Singh and Mohl 1996) or truncated linear regression method (see, Bell 2000). The basic feature of this method over the linear regression is that the new weights must lie within a prespecified boundary condition for each small area unit. The upper and lower limits of boundary interval could be constant across sample units or proportional to the original sampling weights.

Let us assume that a finite population is denoted by $\Omega = \{1, 2, \dots, k, \dots, N\}$, and a sample $s (s \subseteq \Omega)$ is drawn from Ω with a given probability sampling design $p(\cdot)$. 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 values of the variable of interest for the k^{th} population unit and $x'_k = (x_{k,1}, \dots, x_{k,j}, \dots, x_{k,p})$ 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 $\sum_{k \in s} x_{k,j}$ gives the number of individuals in the sample who are not in the workforce. 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.

To obtain a more reliable estimate of the population total of the interest variable, we can use this true population total T_x of the auxiliary information. To do so the main task is to compute new weights w_k for $k \in s$, such that

$$\sum_{k \in s} w_k x_k = T_x \quad (3.1)$$

and the new weights w_k are as close as possible to the sampling design weights d_k .

The equation (3.1) is known as the 'calibration equation' or the constraints function which is used to minimize the distance between two sets of weights. In a survey sampling calibration process, because the prime intention is to minimize the distance or to confirm the closeness of the two sets of weights, it is essential to identify an appropriate distance measure.

In usual notations, let $G_k(w_k, d_k)$ be the distance between w_k and d_k . Then the total distance over the sample s should be defined as

$$D = \sum_{k \in s} G_k(w_k, d_k). \quad (3.2)$$

Now the problem is to minimize the equation (3.2), subject to the constraints equation (3.1). Deville and Sarndal (1992) have accounted a class of well-known distance functions such as Healing distance, minimum entropy distance, Chi-squared distance, etc. They also propose a new distance measure widely known as the Deville-Sarndal distance. A discussion about those distance measures minimisation is provided elsewhere (see, Singh and Mohl 1996; Cai et al. 2004). It is notable that the Lagrange multiplier is commonly used as a minimisation tool of distance measures to calculate the new weights. The Lagrange equation or Lagrangean for this type of minimisation problem would be –

$$L = D + \sum_{j=1}^p \lambda_j \left(T_{x,j} - \sum_{k \in s} w_k x_{k,j} \right) \quad (3.3)$$

where $\lambda_j; \forall j$ are the Lagrange multipliers.

Now consider a special case, if the distance function defined in (3.2) has a property that the first derivative with respect to w_k can be expressed as a function $f(w_k/d_k)$ and inverse of this function f^{-1} exists, then after differentiating and applying the first order minimization condition in (3.3), we have

$$\frac{\partial L}{\partial w_k} = f\left(\frac{w_k}{d_k}\right) - \sum_{j=1}^p \lambda_j x_{k,j} = 0 \quad (3.4)$$

It is convenient to write $x'_k \lambda = \sum \lambda_j x_{k,j}$ for a simple representation. Hence from the above equation in (3.4) the new weights can be formulated as

$$w_k = d_k f^{-1}(x'_k \lambda); \text{ for } \forall k \in s. \quad (3.5)$$

When f^{-1} exists, and for a solution of the Lagrange multipliers vector, λ , the new set of weights can be easily obtained from the above equation in (3.5). However to obtain the values of λ , we should use the known relations $T_x = \sum_{k \in s} w_k x_k = \sum_{k \in s} d_k f^{-1}(x'_k \lambda) x_k$ and $\hat{t}_{x,s} = \sum_{k \in s} d_k x_k$. Hence these relations can be joined as the following form:

$$T_x - \hat{t}_{x,s} = \sum_{k \in s} d_k \{f^{-1}(x'_k \lambda) - 1\} x_k \quad (3.6)$$

where $T_x - \hat{t}_{x,s} = C(\text{say})$ is a known vector, $d_k \{f^{-1}(x'_k \lambda) - 1\}$ is a scalar, and the equation is nonlinear in the Lagrange multipliers vector, λ . Hence, (3.6) can be solved by an iterative procedure such as the Newton-Raphson method (see, *Appendix A*).

3.2.1 How does GREGWT generate new weights? A theoretical view

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

$$G_k^2 = \frac{(w_k - d_k)^2}{2d_k}; \text{ for } L_k \leq \frac{w_k}{d_k} \leq U_k \quad (3.7)$$

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

$$L = \frac{1}{2} \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k} + \sum_{j=1}^p \lambda_j \left(T_{x,j} - \sum_{k \in s} w_k x_{k,j} \right) \quad (3.8)$$

where λ_j ($j=1,2,\dots,p$) 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 (3.8) with respect to w_k and then applying the first order condition, we have

$$\frac{\partial L}{\partial w_k} = \left(\frac{w_k - d_k}{d_k} \right) - \sum_{j=1}^p \lambda_j x_{k,j} = 0 \quad (3.9)$$

for $k \in s \subseteq \Omega$, along with the p^{th} ($j = 1, 2, \dots, p$) constraints conditions in equation (3.1). As earlier it is convenient to write $x'_k \lambda = \sum \lambda_j x_{k,j}$ for a simple representation. Hence the new weights can be formulated as

$$w_k = d_k + d_k x'_k \lambda. \quad (3.10)$$

To obtain values of the Lagrange multipliers equation (3.10) 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

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

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

$$\lambda = \left(\sum_{k \in s} d_k x_k x'_k \right)^{-1} (T_x - \hat{t}_{x,s}); \text{ for } \left| \sum_{k \in s} d_k x_k x'_k \right| \neq 0. \quad (3.11)$$

Hence using the resulting values of Lagrange multipliers, λ , one can easily calculate the new weights w_k from equation in (3.10). Moreover to minimize the truncated Chi-squared distance function in (3.7), an iterative procedure known as the Newton-Raphson method is used in GREGWT program (see, Bell 2000). It adjusts the new weights in such a way that minimises equation (3.7) and produces generalised regression estimates or the synthetic estimates.

The simulated estimates produced by GREGWT macro have their own standard errors. GREGWT calculates these standard errors using a 'group jackknife' approach which is a replication based method. The key idea of the group jackknife method is to divide the survey sample into a numbers of sub-sample replicate groups (practically 30) and calculate the jackknife estimate for each replicate group based on the total sample excluding the replicate group. GREGWT achieves this by computing grossing factors that are also adjusted to meet the benchmarks for each resulting sub-sample. Then the difference between these new estimates and the original sample estimates is used to estimate the standard error. For details about the group jackknife approach see for example, Bell (2000a) and references therein.

3.2.2 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-

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

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,

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

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

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

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

$$A^{-1} = \left(\sum_{k \in s} d_k x_k x'_k \right)^{-1} = \begin{bmatrix} 0.03661582 & -0.00901288 & 0.00228602 & -0.00429437 & -0.00538212 \\ -0.00901288 & 0.03088625 & -0.01214961 & 0.00183273 & 0.00155596 \\ 0.00228602 & -0.01214961 & 0.09100053 & -0.02239201 & -0.01204764 \\ -0.00429437 & 0.00183273 & -0.02239201 & 0.00794951 & 0.00000656 \\ -0.00538212 & 0.00155596 & -0.01204764 & 0.00000656 & 0.02468079 \end{bmatrix}$$

and then by using the results in the relationship (3.11), the Lagrange multipliers should be calculated for this simple particular example as

$$\lambda' = (0.14209475, 0.03501717, 0.18600019, -0.08176176, -0.00426682).$$

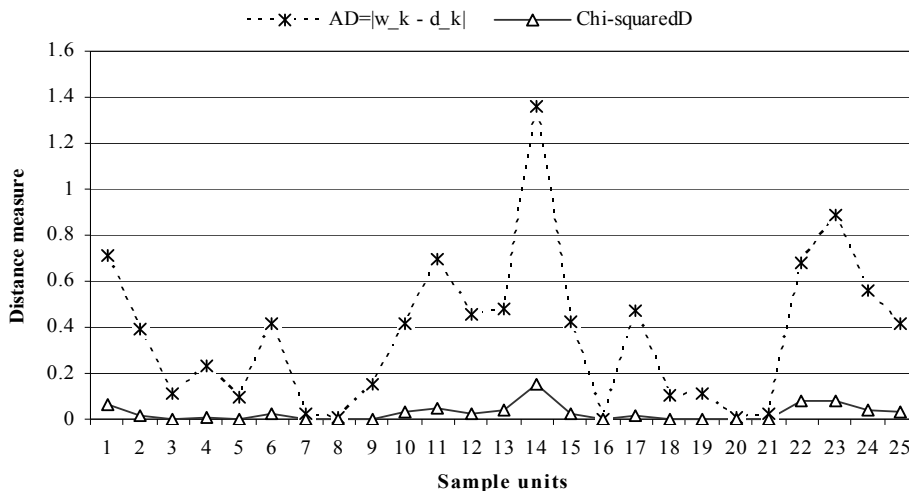
Now using this result in equation (3.10), the new weights or calibrated weights for the Chi-squared distance measure can be easily obtained. The calculated new weights and its distance measures to the sample design weights are given in the following Table (3-2).

Table 3-2: **New weights and its distance measures to sampling design weights**

| d_k | w_k | $w_k - d_k$ | $\chi^2_{G_k}$ |
|-------|-------------------|-------------------------|-----------------------|
| 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 |

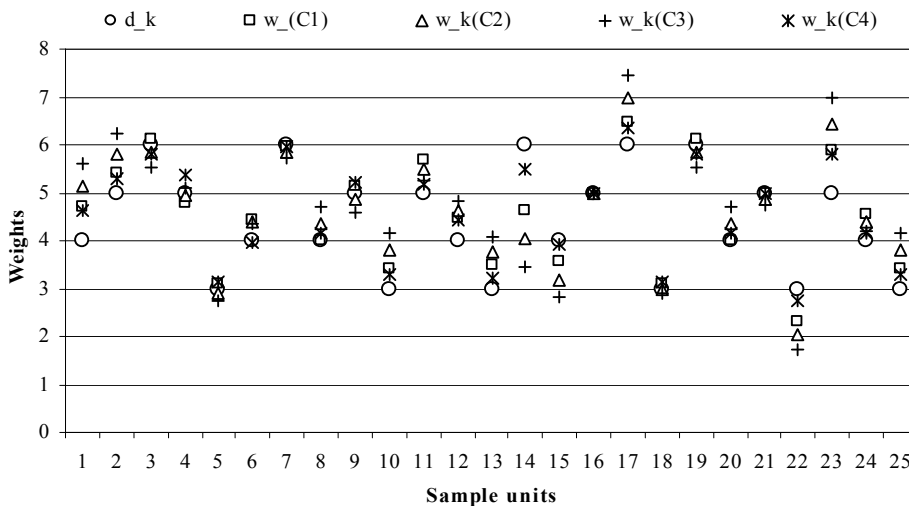
For the 16th unit of our hypothetical data, the new weight remains unchanged to the sampling design weight due to the fact that all entries for this unit are zero. However it is very rare in GREGWT reweighting. In addition, the total absolute distance (TAD) indicates higher quantity. While *absolute distance* has higher value the corresponding *Chi-squared distance* measure also indicates higher value. However the fluctuations within absolute distances are remarkable compare to Chi-squared distance measures (see, Figure 3-1). 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.

Figure 3-1: A comparison of absolute distance and Chi-squared distance measures



Moreover it is interesting to note that the values of a set of *new weights* vary a lots with the changing values of vector for difference between $\hat{t}_{x,s}$ and T_x . Four alternative cases of difference vectors $C1 = [4,3,1,-6,1]'$, $C2 = [8,3,1,-6,1]'$, $C3 = [12,3,1,-6,1]'$ and $C4 = [4,3,1,2,1]'$ (where $C_j = [T_x - \hat{t}_{x,s}]$ for $j = 1,2,3,4$) have been considered for a analysis and the resulting sets of new weights are plotted in Figure (3-2). Results show that the case $C4$ generates more consistent set of new weights compare to 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 3-2: 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 should be 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 it is achieving an appropriate combination of households that best fits known small area benchmarks (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:

- collect a sample survey microdata file (such as CURFs in Australia) and small area benchmark constraints (for example, from census or administrative records);
- select a set of households randomly from the survey sample which will act as an initial combination of households from a small area;
- tabulate selected households and calculate total absolute difference from the known small area constraints;
- 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
- 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 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 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 detail 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.

3.3.1 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 (see Kirkpatrick et al. 1983; van Laarhoven and Aarts 1987; Williamson et al. 1998; Ballas 2001).

The simulated annealing algorithm used in CO reweighting approach is 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):

“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.”

Note that to search an appropriate combination of households from a survey dataset that best fits to 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 CO reweighting by simulated annealing algorithm, a combination of households assume 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

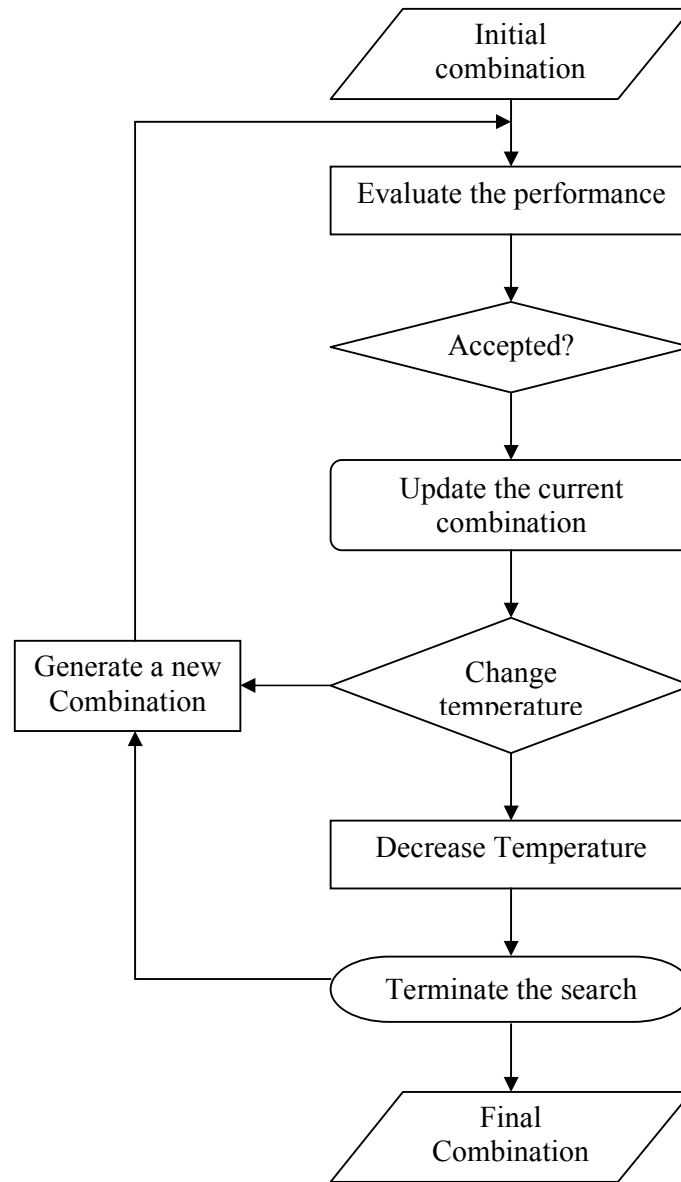
control for the maximum level of performance degradation (% of reduction) acceptable for the change of one element in a combination from sample data. In this case the control parameter is then lowered in steps, with the system being allowed to approach equilibrium for each step by generating a sequence of combinations by obeying the Metropolis criterion. Besides 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 3-3. 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. It is worth mentioning that 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 simulated annealing algorithm in the CO reweighting approach can generate real individuals living in actual households in a sense that individuals are from modelled outputs and not synthetically reconstructed (Ballas 2001).

Figure 3-3: A flowchart of the simulated annealing algorithm



(after Pham and Karaboga 2000)

3.3.2 An illustration of CO process for hypothetical data

A simplified combinatorial optimisation process is depicted in Figure 3-4. Note that in this process when the total absolute difference (aforementioned TAD in subsection 3.2.2 on page 27) 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.

Figure 3-4: A simplified combinatorial optimisation process

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

| <i>Survey Sample Microdata</i> | | | | <i>Known small area constraints</i> | | | |
|--------------------------------|-----------------|-------|----------|-------------------------------------|-----------|---------------------|-----------|
| Household | Characteristics | | | 1. Household size | | 2. Age of occupants | |
| | size | adult | children | Household size | Frequency | Type of person | Frequency |
| a | 2 | 2 | 0 | 1 | 1 | adult | 3 |
| b | 2 | 1 | 1 | 2 | 0 | child | 2 |
| c | 4 | 2 | 2 | 3 | 0 | | |
| d | 1 | 1 | 0 | 4 | 1 | | |
| e | 3 | 2 | 1 | 5+ | 0 | | |
| | | | | Total | 2 | | |

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.

| <i>Household size</i> | Estimated frequency | Observed frequency | Absolute difference | <i>Age</i> | Estimated frequency | Observed frequency | Absolute difference |
|-----------------------|---------------------|--------------------|---------------------|------------|--|--------------------|---------------------|
| | (1) | (2) | (1)-(2) | | (1) | (2) | (1)-(2) |
| 1 | 0 | 1 | 1 | adult | 4 | 3 | 1 |
| 2 | 1 | 0 | 1 | child | 1 | 2 | 1 |
| 3 | 1 | 0 | 1 | | | | |
| 4 | 0 | 1 | 1 | | | | |
| 5+ | 0 | 0 | 0 | | | | |
| | Sub-total: | | 4 | | Sub-total: 2 | | |
| | | | | | Total absolute difference = 4+2 = 6 | | |

Step 4: Randomly select one of 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.

| <i>Household size</i> | Estimated frequency (1) | Observed frequency (2) | Absolute difference $ (1)-(2) $ | <i>Age</i> | Estimated frequency (1) | Observed frequency (2) | Absolute difference $ (1)-(2) $ |
|-----------------------|----------------------------|---------------------------|------------------------------------|------------|--|---------------------------|------------------------------------|
| 1 | 1 | 1 | 0 | adult | 3 | 3 | 0 |
| 2 | 0 | 0 | 0 | child | 1 | 2 | 1 |
| 3 | 1 | 0 | 1 | | | | |
| 4 | 0 | 1 | 1 | | | | |
| 5+ | 0 | 0 | 0 | | | | |
| | Sub-total: | | 2 | | 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 households combination best fits the small area benchmarks):

| <i>Household size</i> | Estimated frequency (1) | Observed frequency (2) | Absolute difference $ (1)-(2) $ | <i>Age</i> | Estimated frequency (1) | Observed frequency (2) | Absolute difference $ (1)-(2) $ |
|-----------------------|----------------------------|---------------------------|------------------------------------|------------|--|---------------------------|------------------------------------|
| 1 | 1 | 1 | 0 | adult | 3 | 3 | 0 |
| 2 | 0 | 0 | 0 | child | 2 | 2 | 0 |
| 3 | 0 | 0 | 0 | | | | |
| 4 | 1 | 1 | 0 | | | | |
| 5+ | 0 | 0 | 0 | | | | |
| | Sub-total: | | 0 | | Sub-total: 0 | | |
| | | | | | Total absolute difference = 0+0 = 0 | | |

(after Huang and Williamson 2001)

3.4 A COMPARISON BETWEEN GREGWT AND CO

A comparison of GREGWT and CO methodologies is provided in this subsection. Although both the approaches are 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; while combinatorial optimisation is based on 'intelligent searching' techniques and attempts to select a combination of appropriate households from a sample that best fits the small area benchmarks.

Tanton et al. (2007) provides 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. Besides the GREGWT takes more time to run compared to combinatorial optimisation, and it is still unchecked 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 3-3. The focus is here mostly on methodological issues. However some entries are consistent with Tanton et al. (2007).

Table 3-3: A comparison of the GREGWT and CO reweighting methodologies

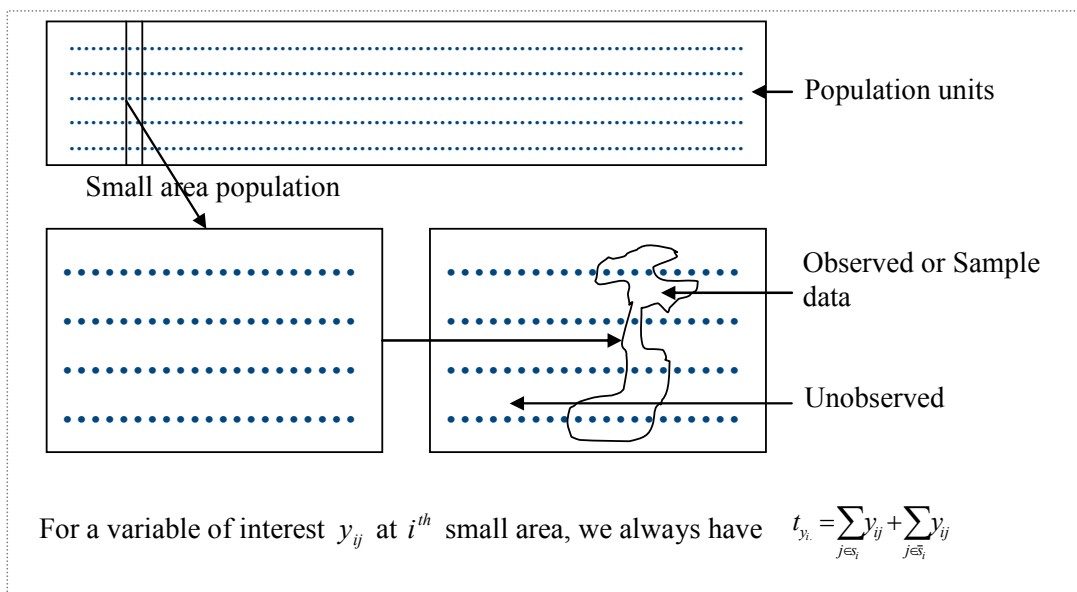
| GREGWT | CO |
|---|--|
| An iterative process | An iterative process |
| Use the Newton-Raphson method of iteration | Use a stochastic approach of iteration |
| Based on a distance function | Based on a combination of households |
| Attempt to minimize the distance function subject to the known benchmarks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

4 SOME NEW POSIBILITIES IN THE SMM METHODOLOGIES

4.1 REWEIGHTING VIA THE BAYESIAN PREDICTION THEORY: A PROSPECTIVE TOOL

A new approach to generating synthetic spatial microdata is offered this subsection. Note that after the sample survey a finite population usually has two parts which are observed sample units called *data* and unobserved sample units (see Figure 4-1). 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 $s_i \cup \bar{s}_i = \Omega_i \subseteq \Omega$ for $\forall i$, where \bar{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 $t_{y_i} = \sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} y_{ij}$.

Figure 4-1: A diagram of prospective tool for generating spatial microdata



The main challenge in this process will be 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 2008c) can deal with such a prediction problem. A detailed description of the Bayesian prediction theory is given by Rahman (2008b). The main steps involve with this reweighting process be as follows:

1. Obtain a suitable joint prior distribution of the interested event 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 $p(y_{ij} : j \in \bar{s}_i | y_{ij} : j \in s_i)$ for $\forall i$;
3. Derive the posterior distribution using Bayes theorem, that is $p(\theta | s, X); E_i \subseteq \theta$, 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 features of this process will be it should produce more reliable small area estimates and the variance estimation of the estimates. It will also be able to create the statistical reliability measures (for example, the Bayes credible region or confidence interval) of SMMs estimates which are still unavailable in literature. To find a suitable prior distribution of our interested event, as well as the appropriate link between observed data and unobserved sampling units, will be difficult in practice.

4.2 STATISTICAL RELIABILITY MEASURES OF SMM ESTIMATES

Validation is an important issue in the area of spatial microsimulation modelling. In microsimulation modelling, a synthetic spatial microdata is generated using different reweighting techniques to simulate data that typically does not exist at small areas. Thus it is quite difficult to validate synthetic small area microdata and to check the statistical reliability of small area estimates. In practice, different researchers use their own ways to validate the model outputs. However, there is no well accepted statistical means to deal with this issue.

One way of validating SMM outputs is to reaggregate estimated data sets to levels at which observed data exists and then compare the estimated distributions with the observed (Ballas 2001; Ballas et al. 2005). In this case if over estimation and under estimation exist at small area levels, that may have less effect in such an aggregated levels comparison. Additionally Taylor et al. (2004) use three types of validation processes for spatial microsimulation estimates, where they compare their model estimates against subpopulation counts data, Australian Bureau of Statistics data, and actual area specific records from the experts knowledge.

In addition, a hierarchical validation method is used in Sweden for validating the SVERIGE model (see, Holm et al. 2001). At first the authors validate the most aggregated level model results with the entire population, and then they compare estimates of demographic variables with the corresponding observables data for last ten years. Note that although the hierarchical validation process can detect source of errors at different modules of the model, it requires data for the entire population, and in reality that is not available for all

countries. It is obvious that there is a need to overcome the weakness in validation of SMMs. Hence we need the prospective methodological advancement in this area. Some new ideas are introduced here in the following subsections.

4.2.1 *Test statistic: A way to test statistical significance of SMMs estimates*

Small area estimates by spatial microsimulation modelling are based on simulated or synthetic spatial microdata, where the microdata are created by using reweighting techniques. Here it is more likely that two types of biased are associated with the SMMs estimates – which are *sampling error bias* and *model bias*. It is still unknown whether these biases of SMMs estimates are obtainable or not due to the complexity in theory and its real world practices. However there is a possibility of testing the statistical significance of SMMs estimates compare to a real value in the small area population.

Suppose $\hat{y}_{i\cdot}$ represents the SMM estimate of j^{th} characteristic of an interest variable y_{ij} at the i^{th} small area. Now if a true value of the estimate is available, say Y_{i0} , in the i^{th} small area population, then we can setup the following hypotheses:

$$H_{i0} : \hat{y}_{i\cdot} = Y_{i0}$$

$$H_{iA} : \hat{y}_{i\cdot} \neq Y_{i0}$$

where H_{i0} and H_{iA} represent the *null hypothesis* and an *alternative hypothesis* respectively.

To test the above hypotheses researchers require an appropriate test statistic. In such a situation for a study on housing stress estimates in Australia we propose to use a test statistic – called *Z-statistic*. Nonetheless another choice for that study should be the use of a modified *standardized residuals test* for validating our SMMs estimates of housing stress. Further detailed descriptions of those test statistics and the study findings should be available soon in our next forthcoming manuscript.

4.2.2 *Confidence interval estimation for SMMs estimates*

The confidence interval estimation is an important measure of statistical reliability for the estimates of a model. To obtain confidence interval for SMMs estimates, the mean square errors or random biases of the estimates should be required for defining the *margin of error*. The margin of error includes the critical value (which is based on the level of significance) and the standard error of the estimate. However it is not easy to evaluate them. There has been some research to obtain such confidence interval estimates for *static* microsimulation models (see, Pudney and Sutherland 1994). The measure of confidence interval is not available for the SMMs yet. Therefore methodological advancement in this area in SMMs will be novel. Our next manuscript on housing stress estimation in Australia will also address the issues of confidence interval estimation for SMMs.

5 CONCLUSIONS

This paper has briefly reviewed the methods of small area estimation and explicitly described some important methodological issues in the spatial microsimulation modelling arena. In the real world situation, often the area-specific sample data are not large enough for all small areas to provide adequate precision of their estimates. Different model based indirect estimations can handle the problem effectively by borrowing strength from auxiliary data to produce better estimates at small area levels. However, these model-based small area estimation methods – especially the spatial microsimulation modelling approaches – involve complex methodologies, tools and techniques, and they suffer from a lack of well-established validation procedures.

Most of the review articles in the small area estimation literature have highlighted methodologies that are fully based on statistical models and theories. But another type of model-based approaches known as the spatial microsimulation modelling has also been widely used in small area estimation. The spatial microsimulation modelling techniques are robust, in the sense that further aggregation or disaggregation is possible on the basis of choice of spatial scales or domains. In addition, since the spatial microsimulation framework uses a list-based approach to microdata representation, it is feasible for the object-orientated programming tools as well as the SAS programming environment to enable further analysis and updating. Also, by linking spatial microsimulation with static microsimulation models, it is possible to measure small area effects of policy changes.

Simulating a reliable spatial microdata population is the key challenge of spatial microsimulation modelling. A brief review of different methodologies such as *synthetic reconstruction* and *reweighting* demonstrates that *reweighting* methods are commonly used techniques in the SMMs, which are playing a vital role to produce small area microdata. There are two reweighting techniques – GREGWT and combinatorial optimisation. The GREGWT technique is based on a truncated Chi-squared distance function – which produces a set of new weights by minimising the total distance with respect to some constraints functions. A very simple theoretical view of GREGWT reveals that the minimisation tool Lagrange multipliers has been used in this process to minimise the distance function and it is based on the Newton-Raphson iterative process. Results from an explicit numerical solution demonstrate that sets of new weights can vary substantially with changing values of the vector of difference between the benchmarks totals and sample based estimated totals. Moreover the Chi-squared distance measures show more smooth fluctuations than the absolute distance measures.

On the other hand, the combinatorial optimisation technique is based on an intelligent searching algorithm *simulated annealing* – which selects an appropriate set of households' combination from a survey microdata that best fits to the benchmark constraints by minimising the total absolute error/distance with respect to the *Metropolis Criterion*. In this process, change in the total absolute error (the role of energy) becomes potential change in households' combination performance to meet the benchmarks constraints. The new weights give the actual household units from the sample survey microdata which are the best representative combination. Thus CO is a selection process to an appropriate

combination of sample units rather than calibrating the sampling design weights to a set of new weights. A comparison between the GREGWT and CO reveals that they are using quite different iterative algorithms and their properties are also different. However their performances are fairly similar according to the advantages of spatial microsimulation modelling.

Validation is an important issue in the area of spatial microsimulation modelling. In a microsimulation model synthetic spatial microdata is generated using different reweighting techniques to simulate data that typically does not exist at small areas. Thus it is quite difficult to validate synthetic small area microdata and to check the statistical reliability of small area estimates. In practice different researchers use their own ways to validate the model outputs, but there is no well accepted statistical means to deal with this issue. As well the small area estimates by SMMs do not have the measures of statistical reliability.

Finally the study points out some new possibilities in the methodology, which are a new reweighting technique based on the Bayesian prediction theory, a way statistical significance test and confidence interval estimation of the small area estimates produced by the SMMs. In our next manuscript we will provide a clear definition and detailed descriptions of the test statistic(s) exercise for the statistical significance test, as well as confidence interval estimation of small area housing stress estimates in Australia through SMM. Further research may look into these options of methodological advancements in more depth to establish them both in theory and in usual practices to all spatial microsimulation models.

REFERENCES

- ABS 1999, *Demographic Estimates and Projections: Concepts, Sources and Methods*, Catalogue Number 3228.0, Canberra, Australian Bureau of Statistics.
- ABS 2002, *The 1998-99 Household Expenditure Survey, Australia: Confidentialised Unit Record Files (CURF)*, Technical Manual (2nd ed.), ABS Catalogue 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 Paper presented at the 4th International conference on GeoComputation, Virginia, USA, 25-28 July.
- Ballas, D. 2001, A spatial microsimulation approach to local labour market policy analysis, *Unpublished PhD thesis*, School of Geography, University of Leeds.
- 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 Rowntree 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.
- Battese, G.E., Harter, R.M. and Fuller, W.A. 1988, 'An error component model for prediction of county crop areas using survey and satellite data', *Journal of the American Statistical Association*, vol. 83, pp. 28-36.
- 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.
- Cai, L., Creedy, J. and Kalb, G. 2004, Reweighting the survey of income and housing costs for tax microsimulation modelling, Melbourne, The University of Melbourne.
- 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 The 8th Nordic Seminar on Microsimulation models, Oslo, 7-9 June.
- 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.
- Deville, J.C. and Sarndal, C.E. 1992, 'Calibration estimators in survey sampling', *Journal of the American Statistical Association*, vol. 87, no. 418, pp. 376-382.
- Duley, C.J. 1989, *A Model for Updating Census-Based Population and Household Information for Inter-Censal Years*, School of Geography, University of Leeds, Leeds, England.
- Epstein, J.M. 1999, 'Agent-based computational models and generative social science', *Complexity*, vol. 4, no. 5, pp. 41-60.
- 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.
- EURAREA Consortium 2004, *EURAREA project reference volume*, Office for National Statistics, United Kingdom.
- Evans, S.P. and Kirby, H.R. 1974, 'A three dimensional furness procedure for calibrating gravity models', *Transportation Research*, vol. 8, pp. 105-122.
- Fay, R.E. and Herriot, R.A. 1979, 'Estimation of income from small places: an application of James-Stein procedures to census data', *Journal of the American Statistical Association*, vol. 74, pp. 269-277.
- 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.

- Gonzalez, M.E. 1973, 'Use and Evaluation of synthetic Estimates ', Proceedings of the Social Statistics Section, American Statistical Association, USA pp. 33-36.
- 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.
- 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.
- Holm, E., Holme, K., Makila, K., Kauppi, M.M. and Mortvik, G. 2001, *The SVERIGE spatial microsimulation model - content, validation, and example applications*, Spatial Modelling Centre, UMEA University.
- 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.
- 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.
- Levy, P.S. 1979, 'Small area estimation -- synthetic and other procedures, 1968-1978' in J. Steinberg, (ed.) *Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion*. Washington, D.C., National Institute on Drug Abuse.
- 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.
- Moriarty, 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.
- Moriarty, 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.
- National Center for Health Statistics 1968, *Synthetic State Estimates of Disability*, Washington DC, P.H.S. Publications, Government Printing Office.
- Norman, P. 1999, *Putting iterative proportional fitting on the researcher's desk*, WP 99/03, School of Geography, University of Leeds, Leeds.
- 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.
- Pudney, S. and Sutherland, H. 1994, 'How reliable are microsimulation results? An analysis of the role of sampling error in a U.K. tax-benefit model', *Journal of Public Economics*, vol. 53, no. 3, pp. 327-365.
- Rahman, A. 2008a, A review of small area estimation problems and methodological developments, *Online Discussion Paper - DP66*, NATSEM, University of Canberra.
- Rahman, A. 2008b, *Bayesian Predictive Inference for Some Linear Models under Student-t Errors*, Saarbrücken, VDM Verlag.
- Rahman, A. 2008c, 'The possibility of using Bayesian prediction theory in small area estimation', presentation to the ARCRNSISS/ANZRSIAI Annual Conference, Adelaide, November 30 to December 03.
- Rahman, A. 2008d, 'Methodologies, Tools and Techniques in Small Area Estimation: An Overview', presentation to the ARCRNSISS Methodologies, Tools and Techniques workshop, Newcastle, June 5 - 6.
- 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.
- Saarloos, D.J.M. 2006, *A framework for a multi-agent planning support system*, PhD thesis, Eindhoven, Eindhoven University Press.
- Sarndal, C.E., Swensson, B. and Wretman, J. 1992, *Model assisted survey sampling*, New York, Springer - Verlag Inc.
- 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.
- Singh, A.C. and Mohl, C.A. 1996, 'Understanding Calibration Estimators in Survey Sampling', *Survey Methodology*, vol. 22, no. 2, pp. 107-115.
- Smith, K.S., Nogle, J. and Cody, S. 2002, 'A Regression Approach to Estimating the Average Number of Persons per Household', *Demography*, vol. 39, no. 4, pp. 697-712.
- Tanton, R. 2007, 'SPATIALMSM: The Australian spatial microsimulation model', 1st General Conference of the International Microsimulation Association, Vienna, 20-21 August.
- Tanton, R., Williamson, P. and Harding, A. 2007, 'Comparing two methods of reweighting a survey file to small area data - Generalised regression and Combinatorial optimisation', 1st General Conference of the International Microsimulation Association, Vienna, 20-22 August.
- 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. 2001, *Microdata for Small Areas*, Manchester, The Cathie Marsh Centre for Census and Survey Research (CCSR), University of Manchester.
- 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*, New York, 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, Leeds, England.

- 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 anonymized records', *Environment and Planning Analysis*, vol. 30, pp. 785-816.
- Williamson, P. 2007, *CO Instruction Manual*, Working Paper 2007/1, Population Microdata Unit, Department of Geography, University of Liverpool, United Kingdom.
- 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. Although it is a very well known iteration method, the basic theory is provided here.

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

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

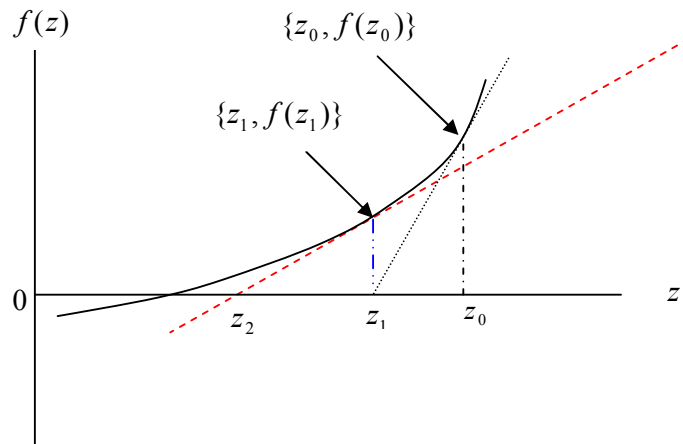
where z_0 is an initial assumed root of $f(z)$, f' represents the first order derivative and ε is a very small arbitrary positive quantity.

Keeping terms only to first order derivative, we have

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

Now (2) is the equation of the 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 (see for example figure a-1).

Figure a-1: Graphical representation of the Newton-Raphson iteration process



The expression in (2) 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 (2), 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.

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

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

Let the process should be repeated until $(r + 1)$ times when a value of the arbitrary quantity, ε_r 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

$$z_{r+1} = z_r - \{f'(z_r)\}^{-1} f(z_r); \text{ for all } r = 1, 2, 3, \dots \quad (\text{a4})$$

It is worth noting that the Newton-Raphson method uses above iterative process to approach one root of a function and a well-chosen initial root value can lead the convergence quickly (see, Fig. a-1). However the procedure can be unstable near a horizontal asymptote or a local extremum. Besides, the Newton-Raphson 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.

Now equation (3.6) in page 26 can be written as a function of the vector λ , which is as follows

$$l_j(\lambda) = C_j - \sum_{k \in S} d_k \{f^{-1}(x'_k \lambda) - 1\} x_{k,j} = 0 \quad (\text{a5})$$

for $j = 1, 2, \dots, p$; where $C = T_x - \hat{t}_{x,s}$ is a known vector, $d_k \{f^{-1}(x'_k \lambda) - 1\}$ 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

$$\lambda_{[r+1]} = \lambda_{[r]} - [l'(\lambda)]_{\lambda_{[r]}}^{-1} [l(\lambda)]_{\lambda_{[r]}}; \text{ for all } r = 1, 2, 3, \dots \quad (\text{a6})$$

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 $[l'(\lambda)]_{\lambda_{[r]}}$ 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 $|\lambda_{[r+1]} - \lambda_{[r]}| < \varepsilon_r = 0.0001$ is satisfied or predefined maximum iteration has reached. However, the ε_r can take any suitable positive arbitrary value and the choice is fully depending on our desired accuracy.