Estimating small area health-related characteristics of populations: a methodological review

Azizur Rahman

School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, Australia

Abstract

Estimation of health-related characteristics at a fine local geographic level is vital for effective health promotion programmes, provision of better health services and population-specific health planning and management. Lack of a micro-dataset readily available for attributes of individuals at small areas negatively impacts the ability of local and national agencies to manage serious health issues and related risks in the community. A solution to this challenge would be to develop a method that simulates reliable small-area statistics. This paper provides a significant appraisal of the methodologies for estimating health-related characteristics of populations at geographical limited areas. Findings reveal that a range of methodologies are in use, which can be classified as three distinct set of approaches: i) indirect standardisation and individual level modelling; ii) multilevel statistical modelling; and iii) micro-simulation modelling. Although each approach has its own strengths and weaknesses, it appears that micro-simulation-based spatial models have significant robustness over the other methods and also represent a more precise means of estimating health-related population characteristics over small areas.

Correspondence: Azizur Rahman, School of Computing and Mathematics, Charles Sturt University, Locked Bag 588, Wagga Wagga, NSW 2678 Australia.
Tel: +61.2.6933.4744.
E-mail: azrahman@csu.edu.au

Key words: Health-related characteristics; Indirect standardisation; Micro-simulation modelling; Multilevel models; Small area estimates.

Acknowledgements: an earlier version of this article was presented at the 23rd Australian Statistical Conference, which was held in conjunction with the 14th Australasian Data Mining Conference (AusDM) and the 9th Australian Conference on Teaching Statistics (OZCOTS) in Canberra. The author would like to thank the anonymous reviewers and the editor for their valuable and insightful comments that were used to shape this final version.

Received for publication: 20 July 2016.
Revision received: 18 November 2016.
Accepted for publication: 21 November 2016.

Geospatial Health 2017; 12:495
doi:10.4081/gh.2017.495

Introduction

Health-related characteristics of a population in a society are significant to health promotion programs and to the provision of better health services. The efforts of feasible health planning generally target limited areas such as the local health region or small area health units, while the population-specific health program planning often requires precise estimates of health behaviour at these levels, for which health-related data are not always available. Even if regional level knowledge of the quantitative dimensions of health-related behaviour can be attained by conducting a costly sample survey, such surveys seldom generate reliable data for small geographic surroundings, such as the statistical local areas (SLAs) in Australia, counties in United States (USA) or wards (electoral divisions) in the United Kingdom (UK). Therefore alternative techniques are necessary to get small-area estimates of health indicators.

Researchers or policy makers sometimes rely on national or state-level datasets to understand the health needs of their communities. The lack of a national dataset for characteristics of individuals at small area levels negatively impacts the ability of local and national agencies to manage serious health issues in the community and their associated risks factors. A solution to this problem would be to develop a model that simulates spatial micro-population datasets at a fine geographic level. This can be achieved by using small area estimation (SAE) techniques – commonly known as the statistical modelling approaches, such as indirect standardisation and individual level modelling (Schaible, 1996; Charlton, 1998; Bajekal *et al*., 2004; Scholes *et al*., 2008), multilevel statistical modelling (Ghosh and Rao, 1994; Pfeffermann, 2002; Rao, 2003; Saei and Chambers, 2003; Rahman *et al*., 2010; Pfeffermann, 2013; Lehtonen and Veijanen, 2015; Rao and Molina, 2015) and micro-simulation modelling (Brown *et al*., 2004; Chin and Harding, 2006; Rahman, 2008; Edwards and Clarke, 2009; Rahman *et al*., 2013; Rahman and Harding, 2014; Rahman and Upadhyay, 2015; Rahman and Harding, 2016).

The SAE procedures can provide robust estimates of the population health behaviour in small geographic areas to support comparisons within and between local areas such as SLA levels and state- or national level estimates. Although methodologies of the indirect standardisation and individual level modelling approaches are simpler than the multilevel modelling approaches, they are not robust in terms of computational processes (Bajekal *et al*., 2004). The latest micro-simulation modelling approaches are conceptually more advanced than the statistical modelling methods, and they are also methodologically and computationally more sophisticated. However, it is yet to be assessed which small area technique produces the most valid, statistically reliable and precise estimates of health-related characteristics.

The basic problem with surveys at the state- or national level

is that they are not designed for efficient estimation of the situation in small areas (Heady *et al*., 2003; Rahman, 2009). For any small area containing respondents to a survey, a conventional estimator of the prevalence of health-related characteristics would be constructed from the survey data. Such conventional estimators mainly have the following limitations (Skinner, 1993): i) prevalence estimates can only be computed for a subset of all areas which contain respondents to the survey; and ii) for those small sampled areas the achieved sample size will usually be very small indeed and the estimator will thus have a low precision.

The low precision will be reflected in rather wide confidence intervals (CI) for the survey estimates that are statistically unreliable. Special statistical models or micro-simulation techniques are therefore required to generate reliable estimates for small areas. For instance, the state- or national level survey data or *confidentialised* unit record files (CURFs) data from government or non-government agencies normally have, at best, only a broad geographical indicator of the state- or the territory level. Small-area data are usually unavailable and must therefore be synthesized/simulated (Levy, 1979; Heady *et al*., 2000; Harding *et al*., 2003; Chin and Harding, 2006). Due to the lack of enough sample information in small geographic areas, there is much interest in creating simulated or synthetic estimators for small areas in any country (Rahman and Harding, 2011). The estimates of small-area health-related characteristics such as the smoking behaviour of youth and adults, characteristics by overweight and obesity *etc*. at small-area levels are not readily available for policy making or evaluation purposes. This article provides a significant appraisal of the methodologies for the estimation of health-related characteristics of populations in geographically limited areas.

A range of methodologies have been used to estimates small-area health-related characteristics (Figure 1). Traditionally there are two types of SAE – direct and indirect model-based estimations. The former is based on survey design and includes three estimators: the Horvitz–Thompson (H-T) estimator, generalized regression (GREG) and the modified direct estimator. Indirect SAE methodologies are divided into statistical and geographic approaches (Rahman, 2008), the first of which are based on different statistical models (*i.e*. implicit and explicit models), while geographic modelling uses micro-simulation.

In the statistical modelling, implicit model-based approaches include three types of estimations – synthetic, composite and demographic estimators, whereas explicit models are categorised as area level, unit level and general linear mixed models. Based on the research interest, each of these explicit models is widely studied to obtained small-area indirect estimates using the (empirical-) best linear unbiased prediction (E-BLUP), empirical Bayes (EB) and hierarchical Bayes (HB) methods (Rahman *et al*., 2010; Pfeffermann, 2013; Rao and Molina, 2015). On the other hand, the geographic modelling approaches are based on spatial micro-simulation models, which essentially create synthetic/simulated micro-population data to produce *simulated estimates* (Rahman and Upadhyay, 2015; Rahman and Harding, 2016). Synthetic reconstruction and reweighting are commonly used in micro-sim-

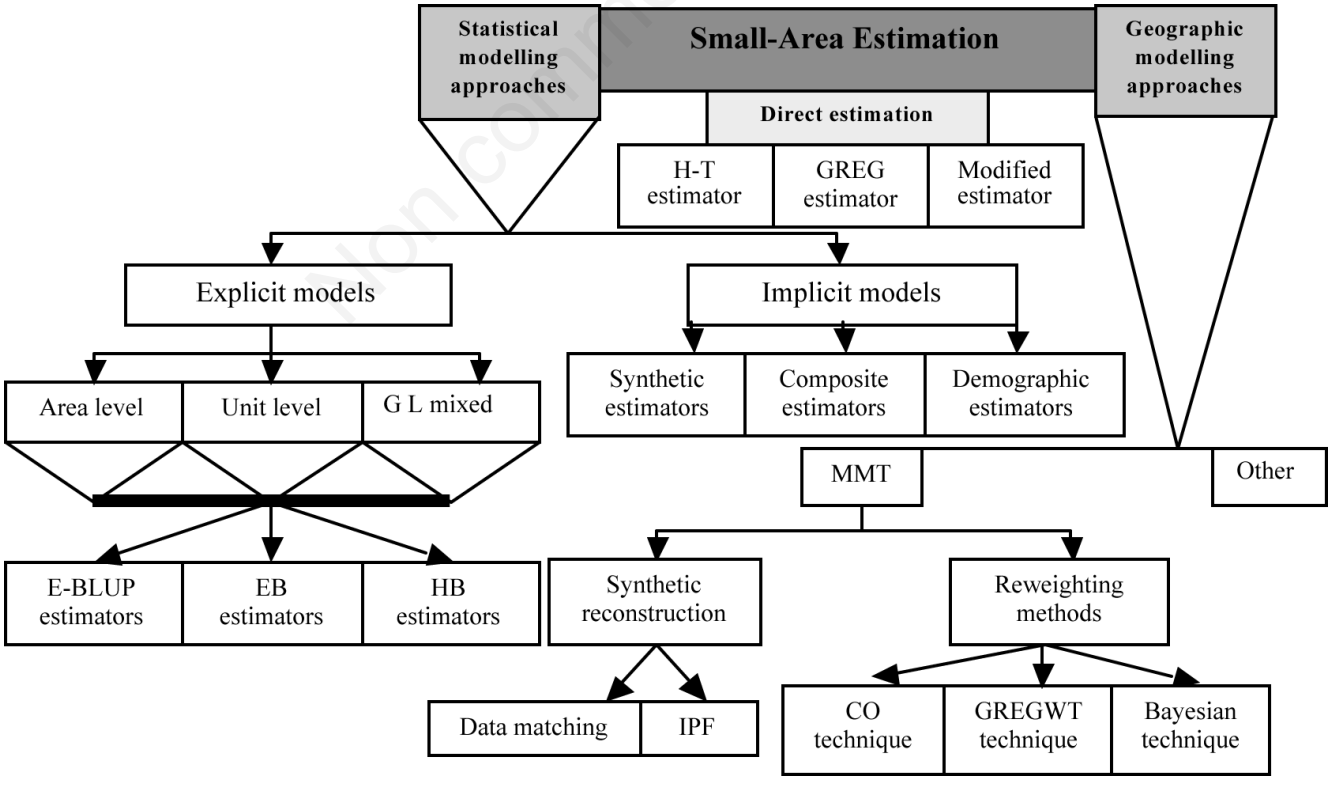

Figure 1. A summary of different techniques and estimators for small area estimation.

ulation, and each is stimulated by different techniques such as combinatorial optimization (CO), generalized regression weighting (GREGWT) and Bayesian reweighting algorithms to produce simulated estimators (Figure 1).

All of these SAE methodologies have not been applied for health modelling, especially not for estimating health-related characteristics. There are several reasons for this, *e.g*. the lacuna of initial data requirements for some methods and the distribution of predictors at the small area level is unknown. In addition, the appropriateness of each method depends on research objectives and settings. Hence, SAE methodologies can be classified as three sets of distinct modelling approaches, *i.e*., i) indirect standardisation and individual level modelling; ii) multilevel statistical modelling; and iii) micro-simulation modelling technology.

Indirect standardisation and individual level modelling

This approach is easy and straightforward. It usually follows a simple indirect standardisation procedure or models that are based on individualistic fallacy (Alker, 1969), *i.e*. individual level covariates from the national- or state level datasets such as CURFs or from Census data.

Simple indirect standardisation procedures involve applying national-level estimates derived from survey data to small-area level population counts to generate small-area estimates. Suppose a researcher is interested in an indirect estimate of the proportion of youth (aged 18-24 years) smoking in a particular SLA in Australia. At first, the researcher can estimate the national/state level proportion of youth smoking from the national health survey data. Applying these national estimates to the census counts of youth within the same age group for the particular SLA would give an estimate of the number of youth that smoke in that SLA, from which the proportion could be estimated simply by dividing the number of youth smoker by the total census count of youths in that particular SLA. Essentially, therefore, the national prevalence rates of smoking for youth is weighted by the proportion of youth in that pre-specified age group. Gibson and Asthana (2001) used this method to calculate the prevalence of heart disease, and Pickering *et al*. (2004) used it to generate the estimates for smoking at the small-area level in England. Kang *et al*. (2016) used such a Bayesian spatial modelling framework to estimate the health outcome measures of the standardised incidence ratio (SIR) and relative excess risk (RER) for the cancer disease at small areas across Queensland in Australia.

The main advantages of this procedure are: i) it is easy and inexpensive to apply since the cell proportions at the local level are available from the Census, and the national estimates for demographic classes are easily obtainable from national surveys such as, for example, the National Health Survey of Australia; ii) the approach is sufficiently flexible to make estimates at the national level - a possible option is to adjust the method by calculating rates for different types of areas using some form of area classification (such as urban and rural, quintiles of deprivation or income *etc*.), and use them to the constituent small areas in each type; and iii) the estimates produced by this method for each small area within a larger area can be ratio-adjusted so that a weighted average of the adjusted small area estimates equals the direct estimate for the larger area. On the other hand the main weakness of this approach is that it considers the notion that the national level prevalence rates for each subgroup apply uniformly everywhere. That means it assumes that the differences in health behaviour measures between areas are solely due to differences in their socio-demographic composition. However, research has shown that individual health-related behaviour, even within the same social group, varies by contextual factors operating at the small-area level (Macintyre *et al*., 2002). To deal with such small-area differences in health-related behaviour, a more complex model is needed to effectively capture the variation between areas that exists over and above that due to differences in their demographic and social composition (Bajekal *et al*., 2004).

An extension of the indirect standardisation method is known as the individual level modelling approach. This type of modelling uses the modelled relationship between individual health behaviour measures obtained from obtainable data against a set of covariates for the same individuals recorded in the survey. In general, the covariates chosen for the model are those that are available as counts for all small areas (for example, covariates from the Census or CURFs data). The individual level modelling estimates the probability of the health behaviour of a person given a set of specific known characteristics of that person such as age, sex, education, marital status, economic class, *etc*. Regression models – in particular logistic regression, linear regression *etc*. – can produce such probabilities and a selection of more and functionally better covariates may greatly improve the model fit. The model-based probabilities are then converted into estimated proportions in each subgroup of the individuals (defined by the covariates) who fall into the relevant health category. These proportions are then applied to the covariate counts available from the Census to derive an overall estimate for the small area in much the same way as for simple indirect standardisation procedure. This modelling approach has been used by Flowers (2003), who applied a logistic regression model to calculate the probabilities of coronary heart disease for agesexsocial classethnicity groups.

Logistic regression can model how the probability of an event may be affected by one or more predictor variables or covariates, which means that it can detect changes in measurements brought about by addition of a new predictor to the regression equation. A remarkable feature of this model is that it makes no assumption about the distribution of the predictor variables. They do not have to be normally distributed, linearly related or of equal variance within each group. A mathematical expression of logistic equation is as follows:

$$\log it\{p(x)\} = \log\left\{\frac{p(x)}{1-p(x)}\right\} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \qquad \text{Eq. 1}$$

where β_0 is the intercept of the regression equation, β_i the regression coefficient of the predictor variable x_i (i=1,2,…k) and p the logistic function.

Note that, even though the inclusion of more and better selection of survey covariates is likely to greatly improve the fit of such individual level models, researchers are restricted in the choice of covariates for estimation by the requirement to have equivalent covariate information for all areas. Hence the main drawback of individual level modelling concerns its data requirements. This form of small-area health behaviours estimation requires an exact correspondence between the covariates used in the model and the data available from the Census or other administrative data sources. However the limited number of cross classifications of

socio-demographic information such as age, sex, ethnicity, socioeconomic class available from the census restricts the choice of covariates in these models.

Multilevel statistical modelling

A more complex set of models in use for small-area health-related characteristics estimation is known as multilevel statistical modelling. The approach extends more traditional statistical techniques by explicitly modelling social context. For example, it can explain the variability in human behaviours and attitudes, as well as how their behaviours and attitudes are modified and constrained by shared social contexts (such as the family composition, community, and residential location, *etc*). In other words, multilevel statistical models can assess the contributions of individual and small-area level factors to both between-individual and between-area variability, showing how individual level and area level factors can contribute to the variability at both levels. This type of analysis also allows for the possibility that different factors contribute to within small area and between small area variability, and permits estimation of area level effects after accounting for compositional differences across small areas (Twigg and Moon, 2002). Using the multilevel modelling technique, statistical models can be applied to survey data that simultaneously account for both individual and small-area level effects or small-area random effects on health-related behaviour. Hence the models are also known as statistical mixed models or random effects models.

In multilevel statistical modelling, a random effects specification at the small-area level is necessary. This specification generally assumes that significant systematic variation between small areas remains after taking account of the effects of covariates in the model (Rahman and Harding, 2010). Such unexplained variation is modelled through the addition of small-area specific random parameters to the fixed effects (Saei and Chambers, 2003). Therefore multilevel models have extended the ability to incorporate unexplained variability between small areas into the health-related behaviour estimation procedures.

Using mathematical notation a simple expression of a two-level model can be written as:

$$SS_{ij} = \alpha_j + \beta_j Income + \varepsilon_{ij} \quad \text{Eq. 2}$$

where j(=1,2, … ,m) refers to level-2 units (*i.e*., small areas or domains) and i(=1,2, … ,n) to level-1 units (*i.e*., individuals or objects), while SS_{ij} represents the smoking status of the i^{th} individual within the j^{th} small area, α_j and β_j represent the intercept and slope parameters, respectively, which are random, *i.e*. they vary from one small area to another, and ε_{ij} is the error term for the model. If we replace the random parameters α_j by $\beta_{0j}=\beta_0+e_{0j}$ *and* β_j *by* $\beta_{1j}=\beta_1+e_{1j}$ and by where e_{0j} and e_{1j} are random elements with parameters, then the above two-level model can be expressed as

$$SS_{ij} = \beta_0 + \beta_{1j} Income + (e_{0j} + e_{1j} Income) + \varepsilon_{ij} \quad \text{Eq. 3}$$

There has been considerable work on more complex structures in multilevel modelling (for example REALCOM: Developing multilevel models for REAListically COMplex social science data project at the University of Bristol in the UK; Browne *et al*., 2001; Goldstein, 2003; Steele *et al*., 2007; Rasbash *et al*., 2009; CMM, 2015). In general, populations of interest to economic and social researchers have a hierarchical or nested structure. This type of populations can be thought of as a pyramid with different levels. For example, individuals in households located in a geographically defined small area can be seen as living at three levels: individuals as the base level, households as the intermediate level and small-area communities as the highest level. Furthermore, some populations have a cross-classified structure. For example, patients can be defined by their family doctor and by the hospital they attend. So far the developments in multilevel statistical modelling have mostly been concerned with analysing data with a nested structure. Nevertheless, although there is no limit to the number of levels of a hierarchy within populations in theory, in practice researchers are rarely in the position to carry out analyses with more than four levels of nesting due to computational constraints.

According to Bajekal *et al*. (2004), two main features of multilevel statistical models make them suitable for producing synthetic estimates for small areas. Firstly, multilevel models are suited to the nature of social surveys where individuals are clustered within households which in turn are clustered within suburbs or postcode sectors. Cluster information provides more accurate standard errors (SEs) and CIs that are generally more *conservative* than the traditional estimates obtained by ignoring the presence of clustering in the data (Goldstein, 2003). Secondly, by allowing the use of covariates measured at various levels of the hierarchy, a multilevel model enables researchers to explore the extent to which any difference between geographical small areas (such as wards or SLAs) are associated with individual, household and small-area level characteristics.

Twigg *et al*. (2000) outlined a multilevel statistical modelling approach in small-area health-related behaviour estimation in an innovative way. They used both individual and area-level covariates to obtain prevalence estimates of smoking and problem-drinking for each ward in England by combining data from a health survey of England with small-area census data. The approach was an advance towards capture of small-area effects on health-related behaviour compared to simple standardisation and individual level modelling. The proposal by Twigg and colleagues involved three stages. Firstly, the small areas covered by the health survey, a multilevel model of individual smoking behaviour using both individual (sex, age and marital status) and area level predictors (for example, the survey estimate of the percentage of private rented households in the postcode sector) was fitted to the survey data. In the second stage, the model parameters of individual and area effects, as well as their interactive effects, were combined to estimate the proportion of smokers in each combination of age, sex and marital status based on ward residents with varying proportions of private renters and car owners. At the last stage, these estimates were applied to the corresponding census counts to provide a synthetic estimate of smoking prevalence for all wards in England.

The Office of National Statistics in UK used a more restricted multilevel statistical model, in which health-related behaviour, such as the smoking status of individuals living in the survey areas, were predicted based on area level covariates only (Heady *et al*., 2003; Pickering *et al*., 2004). This approach results in a set of regression coefficient estimations that relate to area variations. These coefficient estimates are then attached to the known area means, or proportions of the covariates for all small areas taken from the Census and other administrative data sources, to obtain

synthetic estimates of health-related behaviour. This restricted approach claims that controlling for differences in a small area profile is all that is needed for predicting area level differences in health-related behaviour. However concerns about the issue of disaggregation of estimates represent a potential limitation of this method.

There are some significant advantages of multilevel statistical modelling. Firstly, the approach offers a more explanatory model of health-related behaviour than methods that conduct analyses at the individual level. In particular, with respect to people's health-related behaviour, the multilevel statistical models take into account both the effects of individual circumstances and the social and physical environment in which they live. The inclusion of individual level covariates such as age, sex and social status in the model in combination with the corresponding census counts permits the further generation of separate estimates for relevant demographic groups within each small area as well. Secondly, the multilevel statistical modelling approach cannot only generate small area estimates of health-related characteristics, but it can also calculate the statistical reliability measures such as CIs for those estimates. Thirdly, the populations of interest to social and economic researchers have a nested structure or a cross-classified structure and multilevel statistical models analyse the levels of these structures simultaneously. As a result, questions about the appropriate level of analysis are redundant. Fourthly, this modelling technique can fruitfully be applied to repeatedly measured data and to multivariate data, and is especially valuable in situations when data are missing. Finally, multilevel models can easily be fitted by different softwares, *i.e.* multilevel windows (MLwiN), hierarchical linear modelling (HLM) and statistical analysis system (SAS), *etc*.

A number of potential limitations should be borne in mind when applying the multilevel statistical modelling approach in practice. Inclusion of covariates at the individual level in the model imposes quite stringent data requirements, as there must be exact correspondence between these covariates and the Census counts. The limitations on the number of cross tabulations available for small areas such as wards/SLAs from the Census restrict the choice of covariates. Therefore important individual-level predictors of health-related characteristics may be eliminated from the model simply because their distribution at the small-area level is unknown. Additionally, estimating the SEs for the synthetic estimates based on a multilevel statistical model based both on individual and area-level covariates is considerably more complex than the approaches discussed previously. Another disadvantage of area-level models is the ecological fallacy in which relationships between characteristics of individuals are often wrongly inferred from data about groups at the area level, *e.g.*, SLAs. The term ecological fallacy describes the false assumption that relationships revealed from aggregated data (*e.g.*, proportion of a population in a given area) can be used to make predictions about the individuals used to devise these aggregated data, and vice versa (Robinson, 1950; Selvin, 1958), a mistake that can lead to errors in the estimation of relationship magnitude as well as direction when dealing with spatial units. Changes in relationship magnitude and direction can also emerge by simply changing the size of the spatial unit used when analyzing data collected over large areas (Hamil *et al*., 2016). This statistical anomaly is referred to as the modifiable areal unit problem (MAUP) (Dark and Bram, 2007) and is of particular concern given the increasing use of remotely-sensed data or synthetic data for applied spatial research (Pettorelli *et al*., 2014). Another aspect of MAUP is the placement of spatial units, *i.e.* the *zone effect*, which can also affect the magnitude and directionality of relationships (Gotway and Young, 2002, 2007).

The number of published papers shows that the multilevel statistical modelling approach is becoming more popular in different fields of social and economic research. Roux (2008) summarised past work that has used multilevel statistical models to investigate the multilevel determinants of health. Although multilevel modelling is applicable to the study of a broad range of socio-demographic groups or socioeconomic contexts, the vast majority of applications in the health field have focused on geographically defined contexts, such as countries (Chung and Muntaner, 2007), states (Kim *et al*., 2006; Kim and Kawachi, 2007), counties (Muntaner *et al*., 2006; Jonker *et al*., 2013) and most commonly neighbourhoods defined in various ways when dealing with smaller administrative areas (Chaix *et al*., 2007; Rundle *et al*., 2007; Yu *et al*., 2008; Roux and Mair, 2010). The types of group or area level constructs that have been studied in different research include income inequality (Subramanian and Kawachi, 2006), socioeconomic position (Meyer *et al*., 2014), gender differences (Burke *et al*., 2009; Kim *et al*., 2013), social capital (Lindstrom *et al*., 2003; Mohan *et al*., 2005; Kim *et al*., 2006), residential segregation, women's status, and neighbourhood characteristics (*e.g.*, neighbourhood disadvantage or other similar measures, such as social and physical environments) (Moon *et al*., 2007; Roux, 2008; Metcalfe *et al*., 2011; Chum and O'Campo, 2013; West *et al*., 2014; Geronimus *et al*., 2014; O'Campo *et al*., 2015). Most of these studies have used multilevel statistical modelling to isolate associations of group or area level factors with individual level health outcomes after accounting for individual level confounders (*e.g.*, individual level variables associated with health outcomes and with group or community membership and, therefore, with group or community characteristics).

Micro-simulation modelling technology

This approach is currently receiving attention by health researchers (*e.g.* Burden and Steel, 2016, and references therein) for its robustness to use geographical information at small-area levels and examination of small-area impacts of policy changes (Rahman and Harding, 2016). A growing literature indicates that micro-simulation models are becoming increasingly popular and a powerful tool within health research to estimate current health-related behaviour, future prevalence rates, cost of treatment, provision of care needs, and the potential outcomes of policy intervention at small-area levels (for example see, Brown *et al*., 2004; Brown and Harding, 2005; Smith *et al*., 2007; Procter *et al*., 2008; Rahman and Harding, 2011, 2014). This is a promising technique for developing detailed synthetic or simulated micro-data describing household characteristics at the small-area level by combining aggregate census data and more detailed individual record files or households survey datasets (Ballas *et al*., 2003; Chin and Harding, 2006; Rahman, 2009, 2012; Rahman *et al*., 2013; Rahman and Upadhyay, 2015; Rahman and Harding, 2016). However the creation of reliable synthetic micro-data at the small-area level is often challenging for some regions. For example, spatially reliable, disaggregated data are not readily available in the real world. Even if such data would be available in some form, they typically suffer from severe limitations, either due to lack of characteristics or lack of geographical detail.

Micro-simulation modelling can be conducted by reweighting

a generally national level sample so as to estimate the detailed socioeconomic and demographic characteristics of populations and households at the small-area level. An effective reweighting technique combines individual or household micro-data, currently available only for large geographical areas, with spatially disaggregated data to generate synthetic micro-populations for small areas. Thus, the presence of geographical information and detailed household characteristics which both have impact on health-related behaviour in the synthetic spatial micro-population indicates the applicability of a micro-simulation modelling. The features of micro-simulation modelling technology and the associated theories, tools and techniques behind this approach are provided in a number of studies (for instance see, Ballas *et al*., 2006; Chin and Harding, 2006; Chin *et al*., 2006; Rahman *et al*., 2013; Rahman and Harding, 2016) and hence a very brief appraisal of them is provided.

Two types of methodologies for creating simulated micro-population datasets are in use: i) synthetic reconstruction; and ii) reweighting. The former approach includes data matching or fusion (Moriarity and Scheuren, 2001; Tranmer *et al*., 2001) and iterative proportional fitting (Williamson, 1992; Norman, 1999), while the latter utilizes GREGWT (Bell, 2000; Chin and Harding, 2006; Rahman, 2012) and combinatorial optimisation (CO) (Huang and Williamson, 2001; Ballas *et al*., 2003; Rahman, 2008). As reweighting techniques are currently seems to be more popular for creating spatial micro-data than the synthetic reconstruction techniques (Rahman and Harding, 2016), a brief account is given here.

GREGWT is an iterative generalised regression algorithm written in SAS macros to calibrate survey estimates to benchmarks. Calibration can either be looked at as a way of improving estimates or as a way of making the estimates add up to benchmarks (Bell, 2000; Rahman, 2009). This means that 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 algorithm used in GREGWT is based on a constrained distance function known as the truncated Chi-square distance function that is minimised subject to the calibration equations for each small area (for details about the calibration equations, see Rahman, 2008). The method is also known as linear truncated or restricted modified Chi-square (Singh and Mohl, 1996; Rahman, 2009) or the truncated linear regression method (Rahman *et al*., 2010). The truncated Chi square distance function used in the GREGWT algorithm is as follows:

$$D_{\chi^2} = \sum_{k \in S} \frac{(W_k - D_k)^2}{D_k}; \text{ for } L_k \le \frac{W_k}{D_k} \le U_k \qquad \text{Eq. 4}$$

where D_k is the given sampling design weights, W_k the new weights, and L_k and U_k the pre-specified lower and upper bounds, respectively, for each unit k in sample S.

The basic advantage of this method over linear regression is that the new weights must lie within a pre-specified boundary condition for each small area unit. The upper and lower limits of the boundary interval could be constant across sample units or proportional to the original sampling weights. The GREGWT algorithm uses the Newton-Raphson method of iteration (http://mathworld.wolfram.com/NewtonsMethod.html) to minimize this distance function. It adjusts the new weights in such a way that it minimises above distance equation and produces the simulated estimates. The synthetic estimates produced by GREGWT technique have their own SEs. GREGWT calculates these SEs using a *group jackknif*e approach, which is a replication-based method, which uses replicate weights in micro-simulation modelling, and the problem is basically computational, not statistical. By this method one would end up with 30 weights for each small area. For details about the group jackknife approach see, for example, Rahman (2008) and references therein.

Although the MMT approach is generally robust, it suffers from the disadvantage that there has to be a fairly large number of observations in each sample selection stratum (Rahman and Harding, 2016). In practice, it is rare in a survey sample to achieve this number of observations, especially at small-area levels or for micro level data. As a result, when there are too few observations in a sample stratum, the resulting SE estimates should be statistically unreliable. Note that about 30 observations per stratum is a good minimum working number and we may produce this number of observations by suitable combination of classes. However a problem for micro-simulation modelling is the size of the final file: *e.g.*, 1,300 columns (SLA)×30 weights=39,000 columns; then 39,000 columns × 12,000 households = 468 million cells in the final file.

Another method to simulate spatial micro-population datasets is the CO algorithm. This process involves selecting an appropriate combination of household records from available survey micro-data offering the best fit for known benchmark constraints in the selected small areas. In the CO algorithm, 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 area. 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 the fit improves. The iterative process continues until an appropriate combination of households that best fits known small-area benchmarks has been achieved (Voas and Williamson, 2000; Tanton *et al*., 2007; Rahman and Harding, 2014). A simplified CO process is given elsewhere (for instance, see Huang and Williamson, 2001; Rahman, 2008). The overall CO process involves the following five steps. First, collect a sample survey micro-data file (such as CURFs in Australia) and small-area benchmark constraints (for example, from Census or administrative records). Second, select a set of households randomly from the survey sample, which will act as an initial combination of households from a small area. Third, tabulate selected households and calculate total absolute difference from the known small area constraints. Fourth, 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. Fifth, repeat step 4 until no further reduction in total absolute difference is possible.

In the CO algorithm, the fit of a combination of individuals to known small area benchmark constraints is evaluated by the *total absolute error* (TAE), which is the sum of the absolute differences between estimated and observed counts. By simple notations TAE is defined as:

$$TAE = \sum_{ij} | O_{ij} - E_{ij} | \rightarrow 0 \qquad \text{Eq. 5}$$

where, O_{ij} and E_{ij} are the observed and expected counts, respectively, for the i^{th} row in the j^{th} column. Unlike the distance function in GREGWT, the TAE should be minimizing to zero here. Ideally, an optimal solution (the selection of a combination of households that best fits the known benchmarks) would have a TAE of 0, which means there is no difference between the observed and the estimated counts, in another words a *perfect fit*. However the measures of SE are not available yet for the simulated estimates produced by the CO technique. In theory, it may be possible to obtain all possible combinations of households from a finite dataset and the set of combination that best fits the small area benchmarks. However, in practice, it is almost unachievable due to computing constraints for an extremely large number of all possible solutions. To overcome this problem, the CO approach uses several ways of performing *intelligent searching*, effectively reducing the number of possible solutions. Williamson *et al*. (1998) described this problem in more detail and explored various techniques of intelligent searching for the CO process, including the 'hill climbing' approach, the *generic algorithm* approach, and the *simulated annealing* approach. The authors found that modified simulated annealing stood out as the best solution. To improve the accuracy and consistency of outputs, Voas and Williamson (2000) developed a *sequential fitting procedure*, which satisfies a level of minimum acceptable fits for every table used to constrain the selection of households from the survey sample data. Significant features of the GREGWT and CO are summarised in Table 1. The focus here is on methodological issues of these micro-data simulation methods.

One of the newest techniques to MMT is the Bayesian prediction-based reweighting technique for generating spatial micro-population dataset. This method considers the complete scenario of micro-population data units at the small-area level and produces the statistical reliability measure of small-area estimates (Rahman and Upadhyay, 2015).

Suppose Ω represents a finite population and Ω_i the subpopulation of the small-area *i*. 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} represent a variable of interest for the j^{th} observation in the population at the i^{th} small area. Thus, we always have the estimate of population total at i^{th} small area as:

$$t_{Y_i} = \sum_{j \in S_i} Y_{ij} + \sum_{j \in \bar{S}_i} Y_{ij} \qquad \text{Eq. 6}$$

The basic steps involved with this process of spatial micro-data simulation are as follows.

First, obtain a suitable joint prior distribution of the event under research E_i, say smoking status in the population at the i^{th} small area, *i.e.* $p(E_i)$ for $\forall_i$.

Second, find the conditional distribution of unobserved sampling units given the observed data, *i.e.* $p(Y_{ij} : j \in \bar{S}_i \mid Y_{ij} : j \in S_i)$ for $\forall_i$.

Third, derive the posterior distribution using Bayes' theorem, *i.e.* $p(\theta|S,X)$; $E_i \subseteq \theta$, where θ is the vector of model parameters and X an auxiliary information vector.

Fourth, get simulated copies of the entire population from this posterior distribution by the Markov chain Monte Carlo (MCMC) simulation technique.

The method is based on a joint posterior density of parame-

Table 1. A comparison of the generalised regression weighting, combinatorial optimisation, and Bayesian reweighting methodologies.

GREGWT	CO	Bayesian
Use the Newton-Raphson iteration based on a distance function	Use a stochastic approach of iteration based on a combination of households	Use MCMC simulation method based on a Bayesian prediction function
Attempt to minimize the distance function subject to the known benchmarks	Attempt to select a set of combination that best fits the known benchmarks	Attempt to simulates complete scenarios of the whole population at a small area
Use the Lagrange multipliers tools to minimise the distance function	Use a range of intelligent search tools in optimizing combinations of households	Use the Bayesian methodology to obtain the joint posterior density of parameters
Weights are in fractions and a boundary condition is applied to new weights to achieve a solution where the benchmark constraints are fixed for the algorithm	Weights are in integers and there is no boundary condition on new weights. The benchmark constraints at small areas are not fixed for the algorithm	Weights are in fractions and there is no boundary condition on new weights. The benchmarks at small areas are variables for the algorithm
Typically focus on simulating micro-data at small-area levels and aggregation is possible at larger domains	Offers a flexibility and collective coherence of micro-data, so analysis is possible at any level of aggregation	Typically focus on simulating complete micro-data at small areas and aggregation is possible at larger domains
Estimates have their own standard errors obtained by a group jackknife approach	There is no information about standard errors measure in literature and nothing is practicable yet	Estimates have their own standard errors obtained by the Bayesian approach
In some cases, convergence does not occur and requires adjusting the boundary limits or a proxy indicator for non-convergence	There are no convergence issues. But the final selected combination may still fail to fit specified benchmark constraints	In some cases, convergence does not occur and requires adjusting the prior density and/or linking model.
Sensitive to disagreements between target benchmarks, and the iteration procedure can be unstable near a horizontal asymptote or at local *extremum*	Insensitive to disagreements between target benchmarks, and the iteration algorithm can be fairly stable at local *extremum* within the solution	Insensitive to disagreements between benchmarks, and the simulation algorithm can be stable at local extremum within the solution
There is a standard index to check the statistical reliability of the estimates	There is no standard index to check the statistical reliability of the estimates	Able to create the statistical reliability measures *i.e.* Bayes' credible interval

GREGWT, generalized regression weighting; CO, combinatorial optimization.

ters for the observed sample units Y_i and unobserved population units $\bar{Y}_i$ at the i^{th} small area (Rahman and Harding, 2016):

$$f(\beta,\Sigma \mid Y_i,\bar{Y}_i) \cong |\Sigma|^{-\frac{N_i+p+1}{2}} \left|I_p + \Sigma^{-1} Q\right|^{-\frac{\nu+p+N_i-1}{2}} f\left(\bar{Y}_i \mid Y_i\right) \quad \text{Eq. 7}$$

where β is the coefficient parameters of the (k-1) predictor variables regression model under the p –dimensional matrix-T errors assumptions with the location parameters zero, scale parameters Σ and shape parameter ν (*i.e.* degrees of freedom),

$$Q = (Y_i - X_i\beta)'(Y_i - X_i\beta) + (\bar{Y}_i - \bar{X}_i\beta)'(\bar{Y}_i - \bar{X}_i\beta) \quad ,$$

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

for

$$C(Y_i,H) = \frac{(\pi)^{-\frac{(N_i-n_i)p}{2}} \Gamma_p\left(\frac{n_i-k}{2}\right)|H|^{-\frac{p}{2}}}{\Gamma_p\left(\frac{N_i-k}{2}\right)|S_Y|^{\frac{n_i-k}{2}}} \quad \text{Eq. 8}$$

is the prediction distribution with the location matrix $\bar{X}_i(X_i'X_i)^{-1}X_i'Y_i$ and the covariance matrix $\frac{(n_i-p-k+1)}{(n_i-p-k-1)}[S_Y \otimes H]$,

(here the symbol $\otimes$ refers to the Kronecker product of the matrices S_Y and H), H), $\hat{\beta} = (X_i'X_i)^{-1}X_i'Y_i$, $S_Y = (Y_i - X_i\hat{\beta})'(Y_i - X_i\hat{\beta})$, $H = I - \bar{X}_i(X_i'X_i + \bar{X}_i'\bar{X}_i)^{-1}\bar{X}_i'$, as well as n_i and N_i, are the sample units and the total population units at i^{th} area.

The key feature of this new method is that it simulates complete scenarios of the whole micro-population in a small area, which means it can produce more reliable small-area estimates and their variance estimation. It also enables the creation of the statistical reliability measures, such as the Bayes' credible interval (BCI) of the small-area estimates from micro-simulation models, which are still a challenging issue for other reweighting techniques. This is a probabilistic approach, which is quite different from the deterministic approach used in GREGWT (Chin and Harding, 2006) and the intelligent searching tool *simulated annealing* used in CO (Williamson *et al*., 1998; Huang and Williamson, 2001). Nevertheless, the new approach can adopt the generalised regression model operating in the GREGWT algorithm to link observed units in the sample and unobserved units in the population. In contrast, from the viewpoint of the CO technique, this new system uses the MCMC simulation with a posterior density-based iterative algorithm. As the Bayesian joint posterior probabilities of the parameters are estimated through the MCMC method, this spatial micro-data simulation methodology is somewhat linked with a MCMC sampling. However, it is rather different from the multiple imputation technique advanced by others (Rubin, 1987). The basic computation process of this new approach is predominantly associated with a prediction distribution of unobserved population units given the sample units.

Moreover, one of the issues with this reweighting technique is to identify a suitable prior distribution for each event of interest, as well as an appropriate model for linking observed sample data and unobserved units for each small area. These can be difficult in practice as they may vary with unaccounted problems in the real world. It is also quite common that every modelling approach will have to deal with at least a few complex tasks possibly related to suitable model selection and/or computation. Although selection of an appropriate prior and a better linkage model are identified as two challenging tasks for this approach, the Jeffreys's (1961) invariance theory-based prior and Matrix-T errors linking models selected for this methodology have worked decently by producing appropriate results (Rahman and Upadhyay, 2015).

A detailed appraisal of different methodologies to produce small area health characteristics estimates has revealed that there are three diverse sets of modelling approaches utilised by researchers, which are the indirect standardisation and individual level modelling, multilevel statistical modelling; and micro-simulation modelling technology. Although each of these modelling approaches, with its real world applications, has already been discussed above, a further highlight on significant comparative synopsis of these three sets of methodologies is depicted in Table 2.

The findings demonstrate that the indirect standardisation and individual level modelling approach is useful only when the estimation involves applying national level estimates derived from survey data to small-area level population counts to generate small-area estimates. This means that the modelled relationship between individual health-behaviour measures can be obtained from readily available data against a fixed range of covariates for the same individuals recorded in surveys. Although this approach has been used by many researchers for small-area health-behaviour estimation, the data requirement for this method is always of concern. For instance, it requires an exact correspondence between the covariates used in the model and data available from the Census or other administrative data sources. The limited number of cross classifications of socio-demographic information such as age, sex, ethnicity, socioeconomic class available from the Census restricts the choice of covariates in these models.

Discussion

Most of the developed nations are utilising small-area estimation methodologies as essential means to support the process of knowledgeable and effective decision-making and policy analysis for various issues at local or regional levels. The micro-simulation modelling technology-based spatial models are more precise ways in which two or more sources of data can be combined (Rahman, 2016; Whitworth *et al*., 2016; Rahman and Harding, 2016). However they encounter some methodological and computational complexity. The key objective of generating simulated micro-population datasets at the small-area level is to create data that does not currently exist for small areas. Therefore, validation of simulated micro-data is difficult and may be considered one of the drawbacks of MMT. However, the new Bayesian prediction-based model has overcome this drawback by simulating the complete scenario of micro-population units at the small-area level and then producing the statistical reliability measures (the BCI) of the SAEs from MMT-based model (Rahman and Upadhyay, 2015). There are also ways to deal with the validation problem for other MMTs. For instance, one way of validating micro-simulation modelling outputs is to re-aggregate estimated datasets to larger levels at which observed datasets exist and compare the estimated distributions with the observed. Recently, two new types of validation tech-

niques have been developed and successfully exercised by researchers (for examples, see Rahman and Harding, 2013; Rahman *et al*., 2013; Rahman and Harding, 2016), which are more reliable and scientifically standard means for validating the estimates from micro-simulation modelling technology. Yet scholars are working to improve the validation methods and/or trying to develop a further one for the non-Bayesian reweighting based micro-simulation models (Whitworth *et al*., 2016; Rahman, 2016).

The multilevel statistical modelling approach is frequently used by researchers to explain the variability in human characteristics and how their characteristics are modified and constrained by shared membership of social contexts. The method has an extended ability to incorporate unexplained variability between small areas into the health-related attributes estimation procedures, and can be applied to survey data that simultaneously accounts for either individual and small area level effects or small area random effects on health-related behaviours such as smoking. Although there is no limit to the number of levels of a hierarchy within populations in theory, it is very difficult to carry out analyses with more than four levels of nesting in practice, which is due to computational constraints. As the multilevel model includes the individual level covariates this method also imposes quite stringent data requirements, like individual level modelling, and hence important individual level predictors of health-related characteristics may be dropped from the model due to unknown distributions of these variables at the local level. The estimation procedure of SEs for the estimates is also rather complex in multilevel statistical modelling.

The micro-simulation modelling technology-based spatial models have emerged recently as a very useful alternative for small-area estimation of health-related characteristics. The key challenge for this approach is requirement for reliable synthetic spatial micro-data. Findings have revealed that two reweighting methods, the GREGWT and CO, are commonly used tools to produce small-area micro-data. The former utilises a truncated Chi-squared distance function and generates a set of new weights by minimising the total distance with respect to some constraint functions. In contrast, the CO technique uses an intelligent search algorithm, which selects an appropriate set of households from survey micro-data that best fits the benchmark constraints by minimising the total absolute error/distance. The new weights give the actual household units, which are the best representative combination. Thus, CO is a selection process for reaching 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 reveal that they are using quite different itera-

Table 2. A comparison of the three approaches to small area health-related characteristics estimation.

Property	Indirect standardisation and modelling at the individual level	Multilevel statistical modelling	Micro-simulation modelling
General comment	Models are based on individual-level covariates from surveys or Census data	Models are based on multilevel covariates from surveys or Census data	Models are based on synthetically created micro-population datasets that integrate a set of variables from surveys or Census data
Advantages	Easy and inexpensive to apply and the estimate is unbiased for large samples	Easy to apply and a more explanatory model that can provide CIs of the estimates	Rather sophisticated and state of the art method that can generate measures of statistical reliability of the estimates
	Flexible to calculate estimates at the national level for adjustment use by calculating estimates for different spatial scales	Flexible to allow both effects of an individual's circumstances and at any level of hierarchy of the social system, demographic cluster and physical environment	Robust approach in terms of the choice of further aggregation or disaggregation of the small-area estimates on the basis of different spatial scales or demographic domains
	The estimates produced by this method for each small area within a larger area can be ratio-adjusted so that a weighted average of the adjusted small area estimates equals the direct estimate for the larger area	Able to explore the extent of any differences between the small areas and small-area level characteristics. A range of computing software is also available for multilevel modelling	Able to utilise the small-area level synthetic micro-data file for further analysis and updating. It is also possible to measure small-area effects of any policy change. Other traditional statistical approaches do not have the same robustness
Limitations	This approach considers the notion that the national level prevalence rates for each subgroup apply uniformly across all small areas, in fact this is not viable in many cases.	The method imposes quite stringent data requirements as they demand an exact match between the covariates used in the model and available Census counts.	The approach is extremely computing-intensive with regard to the size of the final file.
	The choice of covariates in the model is restricted by the data requirement to have equivalent covariate information for all of the small areas.	Important individual level predictors may be eliminated from the model simply because their distribution at the small-area level is unknown.	Depends solely on a good micro-data generation technique. Several techniques are in use, but a better and commonly available reweighting method is yet to be developed.
	Borrows strength from the overall data but cannot increase the effective sample.	Estimating the SEs for the estimates, which use both the individual level and area level covariates, is considerably more complex than the individual level modelling technique	When there are too few observations in a sample stratum, the resulting SE estimates become statistically unreliable. Also validations of the small-area estimates are challenging for some models
	The estimates are often unreliable due to misclassification of the models and/or use of inconsistent auxiliary data		
Applications	Mostly used for large sample sizes coming from reliable agencies. Most small-area data files are not large enough in many areas.	Widely used for multivariate and repeatedly measured data, sets with missing observations or with clustering information, *etc.*	Quickly becoming a popular methodology in developed nations and frequently used for small-area estimation and social policy analysis.

tive algorithms and their properties also vary considerably. The CO routine has a tendency to include fewer households but give them higher weight – and, conversely, the GREGWT routine has a tendency to select more households but give them less weight. However, the overall performances are fairly similar for both micro-data simulation techniques from the standpoint of use in micro-simulation modelling.

Ultimately, the findings suggest that although each of these three modelling approaches has its own strengths in relation to generating small-area health-related characteristics estimation. The micro-simulation modelling technology is more robust than other methods in the sense that further aggregation or disaggregation is possible on the basis of the choice of spatial scales or domains. In addition, since the methodology uses a list-based approach to micro-data representation, it is possible to use the micro-data file for further analysis and updating.

Conclusions

This paper has reviewed the methodologies for estimating health-related characteristics of populations at small-area levels. Such estimations of various health disparities offer more informative knowledge than other approaches, and it can help to provide direction for developing advance policies to reduce inequities across any population. By linking the spatial models with the static micro-simulation models, it is also possible to assess small-area effects of policy changes. MMT approaches allow *what-if* scenarios in terms of policy changes, which is missing in the other approaches. Finally, from the overall appraisal, it is apparent that MMT is a comparatively precise way to estimate small-area health-related characteristics and evaluate policy changes. Our future research should employ this new approach to produce the estimates of these characteristics – in particular, the estimates of smoking behaviour of adults and/or estimates of the prevalence of overweight and obesity of adults at the small-area level in Australia.

References

Alker HR, 1969. A typology of ecological fallacies. In: M. Dogan and S. Rokkan (eds). Quantitative ecological analysis in the social sciences. Cambridge Press, London, pp. 69-86.

Bajekal M, Scholes S, Pickering K, Purdon S, 2004. Synthetic estimation of healthy lifestyles indicators: stage 1 report. National Centre for Social Research, London, UK.

Ballas D, Clarke G, Dewhurst J, 2006. Modelling the socio-economic impacts of major job loss or gain at the local level: a spatial microsimulation framework. Spat Econ Anal 1:127-46.

Ballas D, Clarke GP, Turton I, 2003. A spatial microsimulation model for social policy evaluation. In: B. Boots and R. Thomas (eds.). Modelling geographical systems. Kluwer, The Netherlands, pp. 143-68.

Bell P, 2000. GREGWT and TABLE macros - user guide (unpublished). Australian Bureau of Statistics, Canberra, Australia.

Brown L, Harding A, 2005. The new frontier of health and aged care: using microsimulation to assess policy options: tools for microeconomic policy analysis. Productivity Commission of Australia, Canberra, Australia.

Brown L, Yap M, Lymer S, Chin SF, Leicester S, Blake M, Harding A, 2004. Spatial microsimulation modelling of care needs, costs and capacity for self-provision: detailed regional projection for older Australians to 2020. Paper presented at the Australian Population Association Conference, Canberra, Australia.

Browne W, Goldstein H, Woodhouse G, Yang M, 2001. An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models. Multilevel Model Newsl 13:4-9.

Burden S, Steel D, 2016. Constraint choice for spatial microsimulation. Popul Space Place 22:568-83.

Burke J, O'Campo P, Salmon C, Walker R, 2009. Pathways connecting neighbourhood influences and mental well-being: socioeconomic position and gender differences. Soc Sci Med 68:1294-304.

Chaix B, Rosvall M, Merlo J, 2007. Neighbourhood socioeconomic deprivation and residential instability: effects on incidence of ischemic heart disease and survival after myocardial infarction. Epidemiology 18:104-11.

Charlton J, 1998. Use of the census sample of anonymised records (SARs) and survey data in combination to obtain estimates at local authority level. Environ Plan Anal 30:775-84.

Chin SF, Harding A, 2006. Regional dimensions: Creating synthetic small-area microdata and spatial microsimulation models. NATSEM, University of Canberra, Australia.

Chin SF, Harding A, Bill A, 2006. Regional dimensions: preparation of 1998-99 HES for reweighting to small-area benchmarks. NATSEM, University of Canberra, Australia.

Chum A, O'Campo P, 2013. Contextual determinants of cardiovascular diseases: overcoming the residential trap by accounting for non-residential context and duration of exposure. Health Place 24:73-9.

Chung H, Muntaner C, 2007. Welfare state matters: a typological multilevel analysis of wealthy countries. Health Policy 80:328-39.

CMM, 2015. REALCOM: developing multilevel models for REAListically COMplex social science data. The Centre for Multilevel Modelling (CMM), University of Bristol, UK.

Dark S, Bram D, 2007. The modifiable areal unit problem (MAUP) in physical geography. Prog Phys Geogr 31:471-9.

Edwards KL, Clarke GP, 2009. The design and validation of a spatial microsimulation model of obesogenic environments for children in Leeds, UK: SimObesity. Soc Sci Med 69:1127-34.

Flowers J, 2003. Development of an indicator of needs-adjusted stain prescribing at PCO level. Eastern Region Public Health Observatory, London, UK.

Geronimus A, Bound J, Ro A, 2014. Residential mobility across local areas in the United States and the geographic distribution of the healthy population. Demography 51:777-809.

Ghosh M, Rao JNK, 1994. Small area estimation - an appraisal. Stat Sci 9:55-76.

Gibson A, Asthana S, 2001. Resource allocation methodologies for the prevention and treatment of specific diseases - a critical review of the attribution and resource-weighting of condition-specific indicative prevalences. Discussion paper to ACRA no. 08. Advisory Committee on Resource Allocation, Leeds, UK.

Goldstein H, 2003. Multilevel statistical models. Halstead Press, New York, NY, USA.

Gotway CA, Young LJ, 2002. Combining incompatible spatial data. J Am Stat Assoc 97:632-48.

Gotway CA, Young LJ, 2007. A geostatistical approach to linking

geographically aggregated data from different sources. J Comput Grap Stat 16:115-35.

Hamil KD, Basil V, Iannone BV, Whitney K, Huang WK, Fei S, Zhang H, 2016. Cross-scale contradictions in ecological relationships. Landscape Ecol 31:7-18.

Harding A, Lloyd R, Bill A, 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. United Nation University Press, Helsinki, pp. 239-61.

Heady P, Clarke P, Brown G, D'Amore A, Mitchell B, 2000. Small area estimates derived from surveys: ONS central research and development programme. Stat Trans 4:635-48.

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

Huang Z, Williamson P, 2001. A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. Population Microdata Unit, Department of Geography, University of Liverpool, Liverpool, UK.

Jeffreys H, 1961. Theory of probability. Oxford University Press, Oxford, UK.

Jonker FM, Congdon PB, Van-Lenthe FJ, Donkers B, Burdorf A, Mackenbach JP, 2013. Small-area health comparisons using health-adjusted life expectancies: a Bayesian random-effects approach. Health Place 23:70-78.

Kang SY, Cramb SM, White NM, Ball SJ, Mengersen KL, 2016. Making the most of spatial information in health: a tutorial in Bayesian disease mapping for areal data. Geospat Health 11:190-8.

Kim D, Kawachi I, 2007. US state-level social capital and health-related quality of life: multilevel evidence of main, mediating, and modifying effects. Ann Epidemiol 17:258-69.

Kim D, Subramanian VS, Gortmaker LS, Kawachi I, 2006. US state- and county-level social capital in relation to obesity and physical inactivity: A multilevel, multivariable analysis. Soc Sci Med 63:1045-59.

Kim J, Richardson V, Park B, Park M, 2013. A multilevel perspective on gender differences in the relationship between poverty status and depression among older adults in the United States. J Women Aging 25:207-26.

Lehtonen R, Veijanen A, 2015. Estimation of poverty rate and quintile share ratio for domains and small areas. In: Alleva G. and Giommi A. (eds.) Topics in theoretical and applied statistics. Springer, New York, NY, USA.

Levy P, 1979. Small area estimation – synthetic and other procedures, 1968-1978. National Center for Drug Abuse Synthetic Estimates for Small Areas, Washington, DC, USA.

Lindstrom M, Moghaddassi M, Bolin K, Lindgren B, Merlo J, 2003. Social participation, social capital and daily tobacco smoking: a population-based multilevel analysis in Malmö, Sweden. Scand J Public Health 31:444-50.

Macintyre S, Ellaway A, Cummins S, 2002. Place effects on health: how can we conceptualise, operationalise and measure them? Soc Sci Med 55:125-39.

Metcalfe A, Lail P, Ghali W, Sauve R, 2011. The association between neighbourhoods and adverse birth outcomes: a systematic review and mea-analysis of multi-level studies. Paediatr Perinat Epidemiol 25:236-45.

Meyer OL, Castro-Schilo L, Aquilar-Gaxiola S, 2014. Determinants of mental health and self-rated health: a model of socioeconomic status, neighborhood safety, and physical activity. Am J Public Health 104:1734-41.

Mohan J, Twigg L, Barnard S, Jones K, 2005. Social capital, geography and health: a small-area analysis for England. Soc Sci Med 60:1267-83.

Moon G, Quarendon G, Barnard S, Twigg L, Blyth B, 2007. Fat nation: deciphering the distinctive geographies of obesity in England. Soc Sci Med 65:20-31.

Moriarity C, Scheuren F, 2001. Statistical matching: a paradigm for assessing the uncertainty in the procedure. J Off Stat 17:407-22.

Muntaner C, Yong Li Y, Xue X, Thompson T, O'Campo P, Chung H, Eaton WW, 2006. County level socioeconomic position, work organization and depression disorder: a repeated measures cross-classified multilevel analysis of low-income nursing home workers. Health Place 12:688-700.

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

O'Campo P, Wheaton B, Nisenbaum R, Glazier RH, Dunn JR, Chambers C, 2015. The Neighbourhood Effects on Health and Well-being (NEHW) study. Health Place 31:65-74.

Pettorelli N, Laurance W, O'Brien T, Wegmann M, Nagendra H, Turner W, 2014. Satellite remote sensing for applied ecologists: opportunities and challenges. J Appl Ecol 51:839-48.

Pfeffermann D, 2002. Small area estimation -new developments and directions. Int Stat Rev 70:125-43.

Pfeffermann D, 2013. New important developments in small area estimations. Stat Sci 28:40-68.

Pickering K, Scholes S, Bajekal M, 2004. Synthetic estimation of healthy lifestyles indicators: Stage 2 report. National Centre for Social Research, London, UK.

Procter KL, Clarke GP, Ransley JK, Cade J, 2008. Micro-level analysis of childhood obesity, diet, physical activity, residential socio-economic and social capital variables: where are the obesogenic environments in Leeds? Area 40:323-40.

Rahman A, 2008. A review of small area estimation problems and methodological developments. University of Canberra, Canberra, Australia.

Rahman A, 2009. Small area estimation through spatial microsimulation models: Some methodological issues. Paper presented at the 2nd General Conference of the International Microsimulation Association in Ottawa, Canada, pp. 1-45 (June 8-10).

Rahman A, 2012. Recent developments in small area estimation: a microsimulation technology. Paper presented at the 7th Asian Business Research Conference in Dhaka, Bangladesh, (December 21-22).

Rahman A, 2016. Small area housing stress estimation in Australia: calculating confidence intervals for a spatial microsimulation model. Commun Stat (in press).

Rahman A, Harding A, 2010. Some health related issues in Australia and methodologies for estimating small area health related characteristics. Available from: http://www.natsem.canberra.edu.au/storage/Azizur_Rahman_WP-2010.pdf

Rahman A, Harding A, 2011. Social and health costs of tobacco smoking in Australia: level, trend and determinants. Int J Stat Syst 6:375-87.

Rahman A, Harding A, 2013. Tenure specific small area estimation

of housing stress in Australia. Paper presented at the 2013 International Indian Statistical Association (IISA) Conference in Chennai, India (January 2-5).
Rahman A, Harding A, 2014. Spatial analysis of housing stress estimation in Australia with statistical validation, Australas J Reg Stud 20:452-86.
Rahman A, Harding A, 2016. Small area estimation and microsimulation modelling. Chapman and Hall/CRC, London, UK.
Rahman A, Harding A, Tanton R, Liu S, 2010. Methodological issues in spatial microsimulation modelling for small area estimation. Int J Microsimul 3:3-22.
Rahman A, Harding A, Tanton R, Liu S, 2013. Simulating the characteristics of populations at the small area level: new validation techniques for a spatial microsimulation model in Australia. Comput Stat Data Anal 57:149-65.
Rahman A, Upadhyay S, 2015. A Bayesian reweighting technique for small area estimation. In: S.K. Upadhyay, U. Singh, and D.K. Dey (eds.). Current trends in bayesian methodology with applications. Chapman and Hall/CRC, London, UK, pp. 503-19.
Rao JNK, 2003. Small area estimation. John Wiley & Sons, Inc., New York, NY, USA.
Rao JNK, Molina I, 2015. Small area estimation. John Wiley & Sons, Inc., New York, NY, USA.
Rasbash J, Steele F, Browne W, Goldstein H, 2009. A user's guide to MLwiN version 2.10. Centre for Multilevel Modelling, Bristol, UK.
Robinson WS, 1950. Ecological correlations and the behavior of individuals. Am Sociol Rev 15:351-7.
Roux AVD, 2008. Next steps in understanding the multilevel determinants of health. J Epidemiol Commun Health 62:957-9.
Roux D, Mair C, 2010. Neighbourhoods and health. Ann New York Acad Sci 1186:125-45.
Rubin DB, 1987. Multiple imputation for nonresponse in surveys. John Wiley & Sons, Inc., New York, NY, USA.
Rundle A, Roux AV, Free LM, Miller D, Neckerman KM, Weiss CC, 2007. The urban built environment and obesity in New York city: a multilevel analysis. Am J Health Prom 21:326-34.
Saei A, Chambers R, 2003. Small area estimation: A review of methods based on the application of mixed models. S3RI Methodology Working Papers, M03/16. Southampton Statistical Sciences Research Institute, Southampton, UK.
Schaible W, 1996. Indirect estimators in U.S. federal programs. Springer, New York, NY, USA.
Scholes S, Pickering K, Deverill C, 2008. Healthy lifestyle behaviours: model based estimates for middle layer super output areas and local authorities in England, 2003-2005: stage 1 report. The NHS Information Centre for Health and Social Care, Leeds, UK.
Selvin HC, 1958. Durkheim's suicide and problems of empirical research. Am J Sociol 63:607-19.
Singh AC, Mohl CA, 1996. Understanding calibration estimators in survey sampling. Survey Methodol 22:107-15.
Skinner C, 1993. The use of synthetic estimation techniques to produce small area estimates OPCS. Office for Population Censuses and Surveys, London, UK.
Smith DM, Harland K, Clarke GP, 2007. SimHealth: estimating small area populations using deterministic spatial microsimulation in Leeds and Bradford. School of Geography, University of Leeds, Leeds, UK.
Steele F, Vignoles A, Jenkins A, 2007. The effect of school resources on pupil attainment: a multilevel simultaneous equation modelling approach. J Roy Stat Soc A 170:801-24.
Subramanian SV, Kawachi I, 2006. Whose health is affected by income inequality? A multilevel interaction analysis of contemporaneous and lagged effects of state income inequality on individual self-rated health in the United States. Health Place 12:141-56.
Tanton R, Williamson P, Harding A, 2007. Comparing two methods of reweighting a survey file to small area data - generalised regression and combinatorial optimisation. Paper presented at the 1st General Conference of the International Microsimulation Association, Vienna (August 20-22).
Tranmer M, Pickles A, Fieldhouse E, Elliot M, Dale A, Brown M, Martin D, Steel D, Gardiner C, 2001. Microdata for small areas. The Cathie Marsh Centre for Census and Survey Research (CCSR), University of Manchester, Manchester, UK.
Twigg L, Moon G, 2002. Predicting small area health-related behaviour: a comparison of multilevel synthetic estimation and local survey data. Soc Sci Med 54:931-7.
Twigg L, Moon G, Jones K, 2000. Predicting small-area health-related behaviour: a comparison of smoking and drinking indicators. Soc Sci Med 50:1109-20.
Voas D, Williamson P, 2000. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. Int J Pop Geogr 6:349-66.
West J, Lawlor DA, Fairley L, Wright J, 2014. Differences in socioeconomic position, lifestyle and health-related pregnancy characteristics between Pakistani and White British women in the Born in Bradford prospective cohort study: the influence of the woman's, her partner's and their parents' place of birth. Brit Med J Open 4:e004805.
Whitworth A, Carter E, Ballas D, Moon G, 2016. Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. Comput Environ Urban Syst (in press).
Williamson P, 1992. Community health care policies for the elderly: a microsimulation approach. School of Geography, University of Leeds, Leeds, UK.
Williamson P, Birkin M, Rees P, 1998. The estimation of population microdata by using data from small area statistics and sample of anonymised records. Environ Planning Anal 30:785-816.
Yu O, Scribner R, Carlin B, Theall K, Simonsen N, Ghosh-Dastidar B, Cohen D, Mason K, 2008. Multilevel spatio-temporal dual changepoint models for relating alcohol outlet destruction and changes in neighbourhood rates of assaultive violence. Geospat Health 2:161-72.