ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

*J. R. Statist. Soc.* A (2018)
**181**, *Part* 4, *pp.* 927–979

# From start to finish: a framework for the production of small area official statistics

Nikos Tzavidis, Li-Chun Zhang and Angela Luna

*University of Southampton, UK*

and Timo Schmid and Natalia Rojas-Perilla

*Freie Universität Berlin, Germany*

[*Read before* The Royal Statistical Society *at a meeting organized by the* Official Statistics Section *on Wednesday, May 9th, 2018*, Mr M. Baxter *in the Chair*]

**Summary.** Small area estimation is a research area in official and survey statistics of great practical relevance for national statistical institutes and related organizations. Despite rapid developments in methodology and software, researchers and users would benefit from having practical guidelines for the process of small area estimation. We propose a general framework for the production of small area statistics that is governed by the principle of parsimony and is based on three broadly defined stages, namely specification, analysis and adaptation, and evaluation. Emphasis is given to the interaction between a user of small area statistics and the statistician in specifying the target geography and parameters in the light of the available data. Model-free and model-dependent methods are described with a focus on model selection and testing, model diagnostics and adaptations such as use of data transformations. Uncertainty measures and the use of model and design-based simulations for method evaluation are also at the centre of the paper. We illustrate the application of the proposed framework by using real data for the estimation of non-linear deprivation indicators. Linear statistics, e.g. averages, are included as special cases of the general framework.

*Keywords*: Census; Design-based methods; Diagnostics; Inequality; Model-based methods

## 1. Introduction

Small area (or domain) estimation has been and still is a very fertile area of theoretical and applied research in official statistics. Although the term domain is more general as it may include non-geographic dimensions, the term small area estimation (SAE) is the established term. We shall follow the custom in this paper and use the terms area and domain interchangeably. In recent decades an increasing number of national statistical institutes (NSIs) and other organizations across the world have recognized the potential of producing small area statistics and their use for informing policy decisions. Some small area estimates have gained accreditation as national official statistics. Two examples in the UK are the annual set of unemployment estimates for unitary authorities and local authority districts by gender and age groups, and the estimates of average income for electoral wards. Other organizations and research groups have promoted the use of SAE techniques via the development of new methodologies and computational tools that are available for public use. An excellent example is the work by the World Bank and the use of its software `PovMap` (World Bank, 2013). In collaboration with country teams, the World

*Address for correspondence*: Timo Schmid, Institut für Statistik und Ökonometrie, Freie Universität Berlin, Garystrasse 21, Berlin 14195, Germany.
E-mail: Timo.Schmid@fu-berlin.de

Bank has used SAE techniques for producing poverty maps in more than 20 developing countries. This is perhaps the most widespread application of SAE to date. Case-studies can be found in Bedi *et al*. (2007).

Over time users' needs have surpassed the limits of what can be achieved with traditional SAE methods. Nowadays in addition to simple linear statistics such as averages and proportions, users request the estimation of more complex indicators, e.g. measures of deprivation and inequality. Meeting the increasing complexity of users' needs requires specialized methodology and software, beyond conventional survey operations within NSIs. This has created opportunities for closer collaboration between researchers and NSIs and for transferring research into practice. Given the fast development of SAE methods and software, researchers (or analysts) and users of small area statistics can benefit from having practical guidelines for the SAE process. This can help to improve the understanding of what is achievable and to ensure that the methods that are adopted or developed are appropriate for the actual users' needs. In this paper we propose a framework based on three broadly defined stages, namely

(a) specification,
(b) analysis and adaptation and
(c) evaluation,

which are summarized in Fig. 1. A description of user needs, the available data and existing SAE methods are the most important inputs to the first, specification, stage. With the help of the analyst, the user defines a set of possible target geographies and indicators and identifies potential existing small area methods that are applicable given the available data. These are the necessary inputs for the second stage.

The second stage, analysis and adaptation, is where the estimators are developed. In our view it is helpful if this process is governed by the principle of parsimony, i.e. one should be looking to use the simplest possible method that achieves acceptable precision. Parsimony may be defined in terms of a hierarchy of estimation methods in increasing order of complexity. It is always possible to start by producing initial estimates that are easy to compute as part of the usual survey process within an NSI without involving explicit modelling or additional data sources. This can include direct, synthetic and composite estimators (see Section 3.1). Typically, these estimators can be improved by the use of standard unit or area level models (see Section 3.2). Clearly this is a more complex step as it involves model building and diagnostics. Finally, elaborations of the model may include use of transformations, correlated random effects over time and space, non-normal random effects and robust estimators, and semiparametric or nonparametric model specifications. The principle of parsimony dictates that such endeavour should only be introduced to overcome specific shortcomings which have been identified in the more basic methods, and the potential improvement must be weighed against the extra complexity and possible drawbacks. Although such a definition of parsimony is not exact, we believe that it provides a useful framework for guiding the process of producing small area estimates.

The aim of the third stage, evaluation, is to evaluate the multiple sets of estimates that are produced at the previous stage. This involves both uncertainty assessment and method evaluation (see Sections 4.1 and 4.2). Hopefully, the SAE process is finalized provided that at least one set of estimates is considered of acceptable precision. It is common practice for NSIs to have guidelines about precision thresholds for publishing estimates. Such thresholds can be used to define the basis of what is acceptable. However, what constitutes acceptable precision should also be defined relatively by comparing a range of methods in terms of gains in precision, sensitivity to underlying model assumptions, additional investment in resources for implementing the methods and subsequent operational costs and risks. If after following these steps no set of acceptable

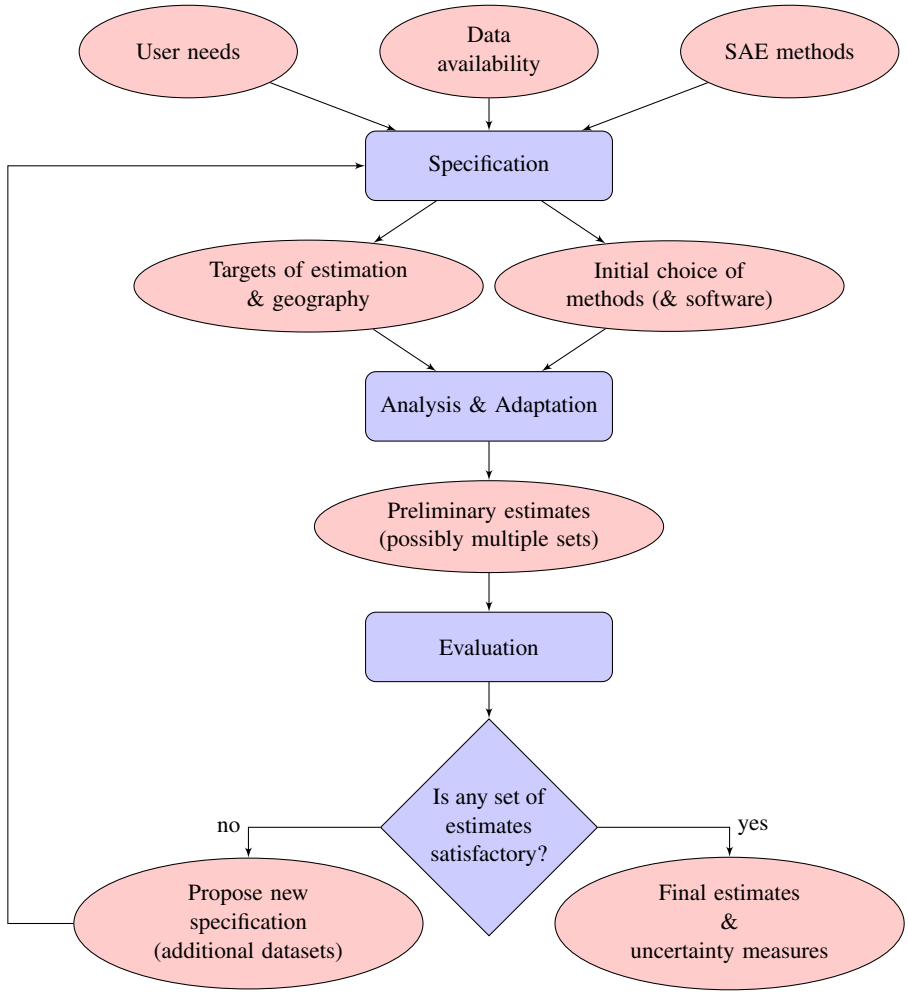



**Fig. 1.** Framework for the production of small area statistics (text in parentheses indicates optional items): ▭, stages of the project; ⬭, inputs and outputs of each stage; ◇, decisions to be made; ⟶, direction of the relationship

small area estimates has been found, the process may need to return to the specification stage for defining alternative geographies, target indicators and/or sources of data.

To keep a practical focus it is important to illustrate the application of the proposed framework by using real data. The data that we use in this paper come from Mexico. Although it has one of the largest economies in Latin America, according to the World Bank Mexico is also among the most unequal countries in the world. Developing policies against deprivation therefore requires a detailed description of the spatial distribution of income deprivation and inequality. The National Council for the Evaluation of Social Development Policy (the Consejo Nacional de Evaluación de la Política de Desarrollo Social) (CONEVAL) is responsible for estimating measures of poverty, social deprivation and inequality in Mexico. Furthermore, the General Law for Social Development (the *Ley General de Desarrollo Social*) requires measures at the national and state levels to be obtained every 2 years and measures at the municipal level every 5 years. For the empirical analysis in this paper we use a sample from the household income and

expenditure survey (the *Encuesta Nacional de Ingresos y Gastos de los Hogares* (ENIGH)) and a large sample of census microdata. Both data sets are produced by the National Institute of Statistics and Geography (the Instituto Nacional de Estadística y Geografía) and were provided to the authors by CONEVAL. In the present paper we shall illustrate the SAE process for estimating linear and non-linear indicators on the basis of continuous outcomes, recognizing that in practice discrete and categorical variables may also be of interest.

The paper is structured as follows. Sections 2–4 describe the three stages of the SAE process: one for each stage. Section 5 provides a review of open source software for SAE. In Section 6 we conclude the paper with some final remarks and comments on open areas for research.

The programs that were used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`

## 2. Specification

In this section we describe the elements of the first stage in our framework. This includes specifying the user needs, the targets of estimation, the target geography and reviewing the data sources that are available and their geographical coverage.

### *2.1. Specify user needs: targets of estimation and target geography*

Sample surveys are designed to provide estimates with acceptable precision at national and specific subnational levels but usually have insufficient sizes to allow precise estimation at lower levels of aggregation. An important task at this stage is the specification of the target level of geography and the targets of estimation, which will impact on all the subsequent SAE processes. It is very tempting for the user to target a geography that is unrealistically low. As we shall see later, doing so will affect the methods and the assumptions that are required for computing the estimates and evaluating their precision. It is also becoming increasingly common that the user is interested in more than simple linear indicators such as averages and proportions, and aims for more complex, non-linear indicators, e.g. estimating the percentiles of the income distribution locally. As will be explained in the next section, increasing the complexity of the targets of estimation increases the granularity of the data that one needs to have access to. Hence, the recommended approach is to start from a relatively high level of geographical aggregation, at which direct estimation with acceptable precision is supported by the survey data, and to move on to more disaggregated levels of geography after assessing the feasibility of producing small area estimates at each level in turn. It would be ideal if a level can be chosen which both serves the user's needs and is well supported by the data that are available. Sometimes, however, the user may have a non-negotiable target level of geography—as is the case in Mexico—dictated by specific policy needs or predetermined by law. Even in this case, it is still the responsibility of the statistician to explain to the user the consequences of the various choices and the extent to which the results will depend on finding a sufficiently good predictive model for the level of interest.

Besides the target level of geography and the targets of estimation, the most important properties of the estimation method also need to be clarified. For instance, whether the user is more interested in cross-sectional estimates or estimates of change over time will affect both the data that are required and the models that are used. For purposes such as fund allocation, policy evaluation and monitoring, it may be important to pay attention to the various ensemble characteristics of the estimates such as the range, the rank and order statistics. The standard approach to deriving model-based small area estimates is to minimize the squared prediction error for

each given area subject to unbiased prediction. This is intuitive for area-specific cross-sectional estimation but is generally not optimal if there are other properties that are more important to the actual use of the small area estimates. A clear understanding of the most desirable properties of the estimates is therefore necessary to ensure that the user's needs are served in the best possible way.

### *2.2. Data availability and geographical coverage*

Identifying what data are needed affects not only the estimation results but also the workload of staff at NSIs and similar organizations. SAE is a prediction problem and typically relies on the use of survey data and data from the census, or administrative or register data sources. The census data contain auxiliary information that is potentially correlated with the target variable and can be used to improve the estimation. Access to census and administrative data sources is usually challenging because of confidentiality constraints. Commonly, access to census aggregate (area or domain) level data is possible but access to census microdata may not be possible. The question is how the type of census data that are available affects SAE. If the user is interested in estimating linear statistics, e.g. small area averages, access to area level census or administrative data will be sufficient for SAE. To illustrate this, suppose that we have data on an outcome variable $y_{ik}$ and a set of covariates $\mathbf{x}_{ik}$ for individuals $i$ in domains $k$. The target of estimation is the domain average and for now let us assume that estimation is assisted by a regression model with model parameters $\boldsymbol{\beta}$. An estimator of the small area average is defined as

$$\hat{\bar{\theta}}_k = N_k^{-1}\left(\sum_{i=1}^{n_k} y_{ik} + \sum_{i=n_k+1}^{N_k} \mathbf{x}_{ik}^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right), \tag{1}$$

where $n_k$ and $N_k$ denote respectively the sample and population size in domain $k$ and $\mathbf{x}_{ik}^{\mathrm{T}}$ is the transpose of the vector $\mathbf{x}_{ik}$. The first summation in equation (1) is computed by using the survey data in domain $k$, assuming that sample data are available in the domain, whereas the second summation represents the out-of-sample model predictions. It is easy to see that, to compute equation (1), there is no need to have access to covariate microdata. Instead, access to domain level totals $\sum_{i=1}^{N_k}\mathbf{x}_{ik}$ will be sufficient. If the interest is, however, in estimating non-linear indicators, then access to census or administrative microdata may be needed. Access to such data is very challenging and has implications for staff resources, in for example ensuring appropriate use of the data and respect for confidentiality constraints. Hence, the complexity of the targets of estimation determines the data requirements for SAE. Although the illustration of methods in this paper assumes the availability of census or administrative microdata for covariates, it is important to discuss briefly what options are available when such data are not available. One possibility is to assume a model for the observed covariates and to impute the missing values from that model (e.g. Sverchkov and Pfeffermann (2004)). With many covariates this might be too cumbersome and Pfeffermann and Sikov (2011) developed a simple non-parametric alternative that is shown to work well. An alternative approach would be to use area level models. Fabrizi and Trivisano (2016) considered hierarchical Bayes approaches to fitting area level models for estimating non-linear indicators. Schmid *et al.* (2017) presented a first attempt to use sources of 'big data', in particular mobile phone data, as covariate information in area level models. We believe that researchers should invest more effort on developing methodologies and software that can be used when population microdata for the covariates are not available or are available for only a sample from the target population.

It is also necessary to examine the data coverage at the specified level of geography. The analyst should explore whether sample observations are available for every small area and also

check the distribution of the sample size across areas. For example, if many of the target areas have no sample data (out-of-sample areas), the user must realize that SAE will heavily rely on model assumptions. Even when data are available for every domain one may still decide to use models in an attempt to improve the precision of direct estimation. Deciding whether to use models and which model to use is a complex process which is governed by a trade-off between improved efficiency and dependence on model assumptions. Our recommendation is for users to be open to alternative methodologies and for researchers to place emphasis on diagnostic analysis for evaluating small area estimates. The process of model building will be illustrated later in the paper.

### *2.3. Illustration using the Encuesta Nacional de Ingresos y Gastos de los Hogares data*

In the ENIGH data case the targets of estimation and the required geography are specified by the General Law for Social Development (see Section 1). The Mexican Government is interested in estimates of proportions and totals of social and economic deprivation, as well as more complex, non-linear, indicators such as estimates of the Gini coefficient (Gini, 1912; Ceriani and Verme, 2012) and income ratio. Methodologists in CONEVAL have access to microdata from the most recent census and survey data from the ENIGH. Hence, the estimation of the target indicators that are specified by the General Law for Social Development is feasible at least in principle.

Let us now look in more detail at the data that are available and their geographic coverage. Mexico is divided into 32 federal entities (states). The State of Mexico (*Estado de México* ('EDOMEX')) has the highest population density and is also regarded by the United Nations Development Programme (UNDP) as being one of the states that most contribute to inequality in Mexico. EDOMEX is made up of 125 municipalities, which by their geographical and demographic characteristics are further grouped into 16 districts. The pilot data that we have available were provided by CONEVAL and come from the 2010 ENIGH survey and the 2010 census in EDOMEX. The ENIGH survey data comprise 2748 households in 58 out of 125 municipalities. The census microdata cover all EDOMEX municipalities. The survey and census data sources include a large number of sociodemographic variables, many of which are common and are measured in similar ways in both data sets. Total equivalized household income is an example of a variable that is available in the ENIGH survey but not in the census.

For the ENIGH survey more than 50% of municipalities are out of sample, making direct estimation for these municipalities impossible. For in-sample municipalities, the median sample size is 21 households and the mean is 47.4 households. The case here illustrates the situation where the user has a non-negotiable target geography predetermined by legal requirements, which clearly poses challenges for estimation. On the one hand, the use of SAE methods can be justified if

(a) they can produce municipal estimates that are more efficient than direct estimates and
(b) they can produce acceptable estimates for non-sampled municipalities.

On the other hand, it is important that the analyst carefully communicates the potential effect of model assumptions and appropriately evaluates the methods and the estimates.

## 3. Analysis and adaptation

The second stage in SAE involves the analysis of the data and the adaptation of the models. As explained earlier, in our view the process should be governed by the principle of parsimony. Section 3.1 presents a triplet of small area estimates described in Eurostat (2012). As we shall

explain, these estimators can always be obtained as by-products of the original sample survey estimation set-up without any additional modelling effort. Ideally this triplet of estimates should be provided by the user to the analyst as an input to the analysis and adaptation stage but this is hardly ever so. The analyst will most probably need to extend the triplet of estimates, by developing suitable models for SAE, both to improve the method of estimation and to be able to handle more complicated target parameters. Sections 3.2 and 3.3 use the ENIGH data to describe and illustrate the core activities of analysis and adaptation including the relevant issues of how to use a model for prediction, model building, model testing, diagnostic analysis, and finally adaptations of the model that are informed by the diagnostic analysis.

### *3.1. Initial triplet of estimates*

The initial triplet of estimates for the small area parameter $\theta_k$ are the direct, synthetic and composite estimates. The direct estimator, which is denoted by $\hat{\theta}_k^{\text{Direct}}$, uses the data from area $k$ only, so it is available only for in-sample areas. For areas with small sample sizes we expect that the direct estimator will have low precision. The synthetic estimator, which is denoted by $\hat{\theta}_k^{\text{Synthetic}}$, uses the data from a broader area that includes area $k$ and so it can be derived for any out-of-sample area as well. Use of a synthetic estimator reduces uncertainty but at the cost of possibly introducing bias. Let us make things more specific and distinguish between two situations of standard design-based sample survey estimation. The first is when no auxiliary data are available and the estimation is based on the design weights directly. For example, let $\bar{\theta}_k$ be the area population mean. The Hajek–Brewer ratio estimator is defined by

$$\hat{\bar{\theta}}_k^{\text{Direct}} = \frac{\sum_{i=1}^{n_k} y_{ik}/\pi_{ik}}{\sum_{i=1}^{n_k} 1/\pi_{ik}}, \tag{2}$$

where $\pi_{ik}$ is the corresponding sample inclusion probability (Hajek, 1958; Brewer, 1963). A synthetic estimator of the mean $\hat{\bar{\theta}}_k^{\text{Synthetic}}$ is given similarly, based on the subsample from a broad area including area $k$, denoted by $\hat{\theta}_k^{\text{Synthetic}} = \hat{\theta}$, where $\hat{\theta}$ is a broad area estimate. The second situation is when auxiliary data are available, in which case the estimation may be based on model-assisted weights (Särndal *et al.*, 1992), denoted by $w_{ik}$, for unit $i$ in area $k$. In this case the direct estimator of the area population mean is given by

$$\hat{\bar{\theta}}_{k,\text{GREG}}^{\text{Direct}} = \frac{1}{N_k} \sum_{i=1}^{n_k} w_{ik} y_{ik},$$

where $w_{ik} = g_{ik}/\pi_{ik}$, and

$$g_{ik} = 1 + \left( \mathbf{X} - \sum_k \sum_{i=1}^{n_k} \mathbf{x}_{ik}/\pi_{ik} \right)^{\mathrm{T}} \left( \sum_k \sum_{i=1}^{n_k} \mathbf{x}_{ik}\mathbf{x}_{ik}^{\mathrm{T}}/\pi_{ik} \right)^{-1} \mathbf{x}_{ik},$$

and $\mathbf{X}$ is the population total of $\mathbf{x}_{ik}$. A synthetic estimator $\hat{\theta}_k^{\text{Synthetic}} = \bar{\mathbf{x}}_k^{\mathrm{T}} \hat{\boldsymbol{\beta}}$ is obtained by the linear model $E(y_{ik}|\mathbf{x}_{ik}) = \mathbf{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta}$, with

$$\hat{\boldsymbol{\beta}} = \left( \sum_k \sum_{i=1}^{n_k} \mathbf{x}_{ik}\mathbf{x}_{ik}^{\mathrm{T}}/\pi_{ik} \right)^{-1} \left( \sum_k \sum_{i=1}^{n_k} \mathbf{x}_{ik} y_{ik}/\pi_{ik} \right),$$

and $\bar{\mathbf{x}}_k = N_k^{-1} \Sigma_{i=1}^{N_k} \mathbf{x}_{ik}$. One approach to reconciling the possibly large bias of a synthetic estimator and the possibly large variance of a direct estimator is to define a composite estimator,

which is a linear combination of the two. This defines the last estimator in the triplet of initial estimators:

$$\hat{\theta}_k^{\text{Composite}} = \alpha_k \hat{\theta}_k^{\text{Direct}} + (1 - \alpha_k)\hat{\theta}_k^{\text{Synthetic}}, \tag{3}$$

for some chosen coefficient $\alpha_k \in [0, 1]$, where by definition $\alpha_k = 0$ for any out-of-sample area.

There are several choices of $\alpha_k$ for the composite estimator (3), including the James–Stein estimator that uses a common $\alpha$ in all areas, and the area-specific minimizer of the mean-squared error (MSE). The latter is not very practical and Rao and Molina (2015) discussed various approaches for selecting $\alpha_k$. One alternative approach is to define $\alpha_k$ as a function of the domain sample size such that for domains with larger sample size a higher weight is given to the direct estimator. It is worth noting that the composite estimator appears more intuitive for target parameters that are linear statistics of the $\{y_{ik}\}$, like domain averages. However, estimators of more complex statistics, e.g. percentiles of the domain-specific distribution function and non-linear indicators, have recently attracted some interest in the small area literature (Tzavidis *et al.*, 2010; Alfons and Templ, 2013). Regardless of how the initial triplet of estimates is produced, it provides useful input to the analysis and adaptation stages and possibly to the specification stage also.

The initial triplet estimates would certainly be more useful if some appropriate measure of the associated uncertainty can be produced in addition. However, it can be challenging to obtain a stable estimate of the potential bias of the synthetic and composite estimator, as we shall discuss in Section 4. At the very minimum, the direct estimates need to be analysed and their uncertainty quantified as these will offer an indication of the improvement that is required for producing small area estimates. It is common that the analyst will subsequently consider the use of more complex model-dependent SAE methods. In this case juxtaposing the direct, synthetic and composite estimates provides a tangible appreciation of the between-area variation of the target parameter, i.e. the heterogeneity across the areas, as well as possibly the predictive power of the auxiliary variables that are already in use.

### 3.2. Use of models for small area estimation

SAE is one of the areas in survey sampling where the use of models is widely accepted as necessary. Model-based methods assume a model for the population and sample data and construct optimal predictors of the target parameters under the model. The term predictor instead of estimator is conventionally used as, under the model, the target parameters are assumed to be random. Here we describe how to use a model to estimate both linear and non-linear small area parameters of interest. In Section 3.3 we describe model building, diagnostic analysis and model adaptations in more detail.

Users of small area statistics in Mexico are interested in the estimation of key income-related indicators such as the head count ratio (HCR) and the Gini coefficient. To this set we add average income, which is also of interest for NSIs. The most widely used approaches for estimating non-linear indicators require the use of unit level survey data for the outcome variable and the covariates, and unit level census microdata for the covariates. Area level models for non-linear indicators have been proposed in the literature (Fabrizi and Trivisano, 2016) but these models lie outside the scope of the present paper.

Two predominant approaches for estimating non-linear indicators are the World Bank method (Elbers *et al.*, 2003) and the empirical best predictor (EBP) method (Molina and Rao, 2010). To start with, both methods make use of a unit level nested error regression model (Battese *et al.*, 1988). The response variable is a welfare variable that is available only in the survey, e.g. income or consumption. The explanatory variables, which are used for modelling the welfare variable,

are available both in the survey and in the census data sets. After the model has been fitted by using the survey data, the estimated model parameters are combined with census microdata to form unit level synthetic census predictions of the welfare variable. The synthetic values of the welfare variable along with a defined poverty line are then used for estimating non-linear indicators, e.g. the HCR or the Gini coefficient. Linear statistics such as average income can also be estimated by using the same synthetically generated values.

We first describe the EBP approach, before we provide a brief discussion of the similarities to and differences from the World Bank method. Under the EBP approach census predictions of the welfare outcome are generated by using the conditional predictive distribution of the out-of-sample data given the sample data. The starting point is the unit level nested error regression model

$$y_{ik} = \mathbf{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + u_k + \epsilon_{ik}, \qquad u_k \sim N(0, \sigma_u^2), \quad \epsilon_{ik} \sim N(0, \sigma_\epsilon^2), \tag{4}$$

where $u_k$ denotes the domain random effect. A random effect is necessary when the covariates that we include in the model do not fully explain the between-domain variability. Assuming normality for the unit level error and the domain random effects, the conditional distribution of the out-of-sample data given the sample data is also normal. The synthetic values of the welfare variable for the entire area population (of size $N_k$) are then generated from the model

$$y_{ik}^* = \mathbf{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + \tilde{u}_k + u_k^* + \epsilon_{ik}^*, \qquad u_k^* \sim N\{0, \sigma_u^2(1-\gamma_k)\}, \quad \epsilon_{ik}^* \sim N(0, \sigma_\epsilon^2), \quad \gamma_k = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2/n_k}, \tag{5}$$

where $\tilde{u}_k = E(u_k|y_s)$ is the conditional expectation of $u_k$ given the sample data $y_s$. In model (5), $\mathbf{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + \tilde{u}_k$ is the conditional mean of $y_{ik}$ in the population given the sample data, whereas $u_k^* + \epsilon_{ik}^*$ are simulated from the conditional normal distribution of $y_{ik}$ for the units outside the sample. Implementation of model (5) requires replacing the unknown quantities $\boldsymbol{\beta}$, $\sigma_u^2$ and $\sigma_\epsilon^2$ with estimates and simulating $L$ synthetic populations of the welfare outcome, $\mathbf{y}^*$. Linear and non-linear indicators are computed in each domain $k$ for each replication and the estimates are averaged over $L$. A moderate number of Monte Carlo simulations, $L = 50$ or $L = 100$, is used in practice. MSE estimation for model-based SAE will be discussed in Section 4.1. For now we note that evaluation of the uncertainty both for in-sample and for out-of-sample domains is usually performed by using a parametric bootstrap under models (4) and (5). Alternatively, protection against model misspecification can be offered by the wild bootstrap. In this case the bootstrap for the unit level error term uses the empirical distribution of scaled residuals instead of a normal distribution.

We now briefly compare the World Bank and EBP methods. Although both methods use a nested error regression model, one key difference in practice is that in the World Bank method it is common to specify the random effect at a much finer geography (cluster) level (indexed by $l$) whereas in the EBP method the random effect is specified at the domain level. A second key difference is that the EBP method simulates population realizations of the outcome from the estimated conditional distribution (5) whereas the World Bank method simulates from the marginal distribution

$$y_{il}^* = \mathbf{x}_{il}^{\mathrm{T}}\boldsymbol{\beta} + u_l^* + \epsilon_{il}^*, \qquad u_l^* \sim N(0, \sigma_u^2), \quad \epsilon_{il}^* \sim N(0, \sigma_\epsilon^2), \tag{6}$$

with all parameters replaced by their estimates. We now distinguish two cases. When clusters coincide with the target domains, Molina and Rao (2010) demonstrated the superior performance of the EBP method for in-sample domains. For out-of-sample domains the predicted

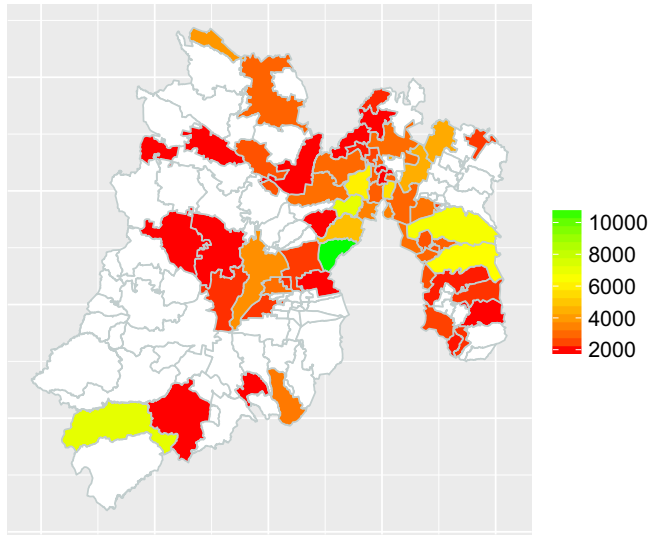
random effect $u_k$ and the shrinkage factor $\gamma_k$ in model (5) are both 0 by default so model (5) reduces to model (6) and the two methods yield the same estimates. Next, consider the more common case where clusters and target domains do not coincide. Since in most applications the between-domain variation tends to be small compared with the between-household variation, the conditional distribution (5) may not differ much from the unconditional distribution, as long as the variance of $\tilde{u}_k$ is small compared with the total variance of $y_{ik} - \mathbf{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta}$. Meanwhile, since the World Bank method is applied at the cluster level, it is possible to capture much of the variability beyond the between-household variability at the cluster level, provided that relevant cluster level covariates are included in the fixed part of model (4). Moreover, the use of the conditional distribution (5) may be impossible in most of the clusters because of the absence of sample units. The World Bank method is then well suited in practice, despite the use of the marginal distribution (6). Having said this, Marhuenda *et al.* (2017) recently proposed EBP methodology that allows for a twofold nested error regression model that can accommodate both cluster and domain random effects.

### 3.3. Model building, residual diagnostics and transformations in practice

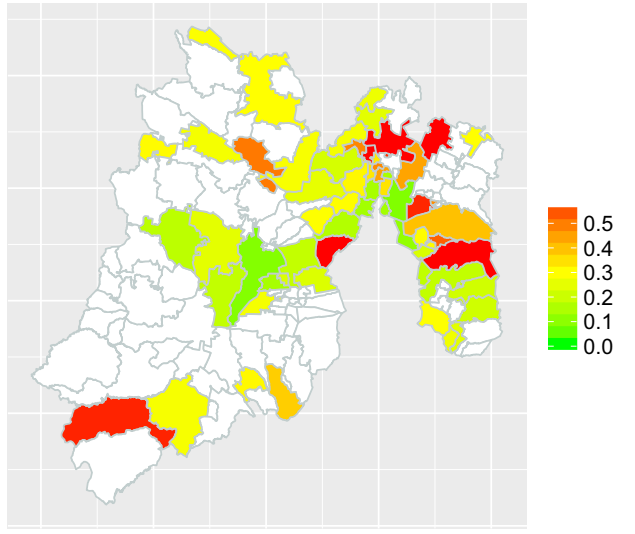
Before considering model-based estimation, an assessment of initial estimates that are produced with the ENIGH data is necessary for motivating the use of more complex methods. The data provider did not supply the initial triplet of estimates that were described in Section 3.1. Producing appropriate sets of initial estimates and their corresponding coefficients of variation (CVs) would require access to data about the sampling design that are beyond our reach. The analysis below, obtained by using the function `direct` of the `sae` package in R (Molina and Marhuenda, 2015), attempts to replicate such initial estimates in a way that can inform the subsequent stages of the process. Fig. 2(a) presents point estimates of average equivalized household income at the municipality level calculated from the ENIGH survey data by using the final weights supplied. Fig. 2(b) shows estimated CVs, obtained under the assumption of a single-stage Poisson sampling of households in each municipality, with first-order inclusion probabilities given by the inverse of the final weights. The assumption of single-stage Poisson sampling is made for convenience. We expect the CVs that are estimated under this assumption will be overly optimistic considering that the actual sampling design of the ENIGH includes stratification and two stages of selection, and has a design effect around 3.3 for the income variable (Encuesta Nacional de Ingresos y Gastos de los Hogares, 2010). However, even under this optimistic scenario it can be seen that, with the exception of a few municipalities, the CVs are clearly above usual publication thresholds of 20–25%. Note also that direct estimates cannot be produced for the out-of-sample municipalities (the white areas). Hence, to satisfy the current user needs we should explore the use of model-based methods.

The use of models aims to improve the precision of small area estimates by making optimal use of the data that are available. Hence, model building, model diagnostics, sensitivity analysis and validation play a central role in model-based SAE. There is no single approach to model building. Here we describe some best practice guidelines that one could follow and illustrate these guidelines for estimating income-related indicators with the ENIGH data.

Model-based estimation requires the use of a model that usually includes area random effects. However, before discussing the use of random effects, the most important step in building the model remains the specification of the fixed effects part. Ideally, one should aim to explain as much between-domain variation as possible by using the available covariates so that random effects can potentially be avoided in the spirit of parsimony. A reasonable starting point for building the model is therefore to use a standard regression model with uncorrelated errors.




(a)




(b)

**Fig. 2.** (a) Direct estimates of average household equivalized income and (b) CVs for EDOMEX municipalities

Alternatively, if one suspects that despite the inclusion of covariates there is unexplained between-domain variability which can affect inference for the regression parameters, the analyst can consider a regression model with correlated errors, e.g. an exchangeable correlation structure in the simplest case. To decide whether to include a covariate in the fixed part of the model one can use simple $t$-statistics—computed by using the correct variance under the model—or information criteria, e.g. the Akaike or the Bayesian information criteria AIC or BIC computed under the standard linear model with uncorrelated errors. In the case of the ENIGH data and following the recommendation by the data provider (CONEVAL), $y$ is defined as the total household

*per capita* income (*ictpc*) measured in Mexican pesos, which is the current monetary and non-monetary income of households adjusted by equivalent scales and economies of scales. Using AIC and a standard linear regression model the following covariates that are available both in the survey and in census data have been identified as good predictors of ictpc:

(a) the percentage of employees who are older than 14 years in the household;
(b) the highest degree of education completed by the head of household;
(c) the social class of the household;
(d) the percentage of income earners and employees in the household;
(e) the total number of communication assets in the household;
(f) the total number of goods in the household.

To investigate whether the use of a mixed effects model is necessary, we estimated a linear model with an exchangeable correlation structure by using generalized least squares (GLS) (Pinheiro and Bates, 2000). The model is estimated in R with function `gls` within the `nlme` package (Pinheiro *et al.*, 2016). The class of GLS models contains the standard linear model that assumes independence as a special case. Therefore, given the fixed effects, the standard linear model is nested within the model with exchangeable correlation structure and a likelihood ratio test or other information criteria can be used to decide whether the model with exchangeable correlation structure fits the data better. First, we compared the GLS with an exchangeable correlation structure against a standard linear model where both models included only an intercept term. This enables us to quantify how much of the between-municipality variability is explained by the model covariates. We conclude that the model with exchangeable correlation structure fits the data better than the standard linear model (AIC for GLS with an exchangeable correlation structure, 54239, *versus* AIC for the standard linear model, 54275). One could also use a likelihood ratio test for comparing the two models which produces a $p$-value for testing. The value of this test statistic is 37.52, with a $p$-value $4.521 \times 10^{-10}$, which provides evidence of significant unobserved heterogeneity between municipalities. Care must be taken with using a likelihood ratio test when a parameter like a random-effects variance is on the boundary of the parameter space (see for example Snijders and Bosker (2012)). In the second step the GLS and standard linear regression models with the set of six covariates that were identified above were compared against each other. The AIC and the likelihood ratio test ($p$-value 0.029) suggest that the model with the exchangeable correlation structure fits the data marginally better (AIC for GLS with an exchangeable correlation structure, 53 077, *versus* AIC for the standard linear model, 53 079). The difference between these AIC-values is very small, indicating that the covariates that we included in the model explain a substantial part of the between-municipalities variability. In particular, the intracluster correlation ICC for the empty GLS model is 0.054 and for the GLS model that includes the six significant predictors it reduces to 0.015. In light of the marginally better fit of the GLS model, the benefits of a random-effects model are likely to be small. We discuss this in Section 4 where we compare indirect and regression synthetic estimates. Although not used in the case-study, model selection and testing procedures under the random-effects model have been proposed in the literature. Here we refer to the use of a conditional AIC-criterion (Vaida and Blanchard, 2005) that accounts for the prediction of random effects in selecting covariates to be included in the model. We further refer to a test for the inclusion of random effects that was proposed by Datta *et al.* (2011), who showed that, if random effects are not needed and are removed from the model, the precision of point and interval estimators is improved. Additional testing procedures have been proposed by El-Horbaty (2015) and reviewed by Pfeffermann (2013).

After the best possible set of covariates has been identified, the inclusion or not of random effects has been decided and the model has been fitted, the next step in model selection uses

residual diagnostics and assessment of the predictive power of the model. Despite the inclusion of some significant covariates, the model may have low predictive power. The user must remember that SAE is concerned with prediction and not with discovering associations and causal mechanisms between the explanatory variables and the outcome. Hence, assessing the overall predictive power of the model is important. One can use simple measures such as the coefficient of determination, $R^2$, of the model without random effects. Alternative, computer-intensive methods such as cross-validation can be used. Cross-validation was mentioned by Pfeffermann (2013) and consists of leaving some areas out of the model fitting process and comparing model-based predictors for these areas with corresponding design-based estimates. For example, we may use as a validation benchmark design-based estimates for larger areas which can be trusted. For residual diagnostics we propose the use of graphical diagnostics such as normal $Q$–$Q$-plots of the residuals (unit level and domain level) for checking the model assumptions, and plots of standardized residuals against fitted values for testing the assumptions of constant variance. If residual diagnostics indicate that the model assumptions hold, the analyst can proceed to the production of point and MSE estimates. However, in most applications some adaptations of the model will be needed.

To illustrate the use of diagnostic analysis and model adaptation we focus on the EBP method that we described in Section 3.2 which relies on the normality of the residual terms. Fig. 3 shows normal $Q$–$Q$-plots of household level and municipal level residuals (random effects) that were obtained by fitting model (4) to the ictpc-variable, using the six covariates that we identified before and including municipality-specific random effects. There are notable departures from normality. This can be seen both from the shape of the normal $Q$–$Q$-plots and from Table 1 where the skewness and kurtosis of the two sets of residuals are clearly different from what is expected for normal data.

When residual diagnostics indicate that there are departures from normality, the analyst has several options. The first option is to use alternative parametric specifications that are more realistic. In the case of income data two possible distributions are the Pareto distribution or the generalized beta distribution of the second kind. The complication with using alternative distributions is that the analyst may need to develop new estimation and inference theory for each new application. Alternative semiparametric approaches to model-based SAE have also been proposed (Weidenhammer *et al.*, 2014). Use of semiparametric methods also requires new theory and additional training for the users. There is also a large body of literature on extensions of the nested error regression model to handle real data challenges better. Examples include outlier robust estimation (Datta and Lahiri, 1995; Ghosh *et al.*, 2008; Sinha and Rao, 2009; Chambers *et al.*, 2014; Fabrizi *et al.*, 2014), models with non-parametric instead of linear signal

**Table 1.** Coefficients of determination, skewness and kurtosis for household level residuals and municipal level residuals of the working models for EBP with and without transformations

| *Transformation* | *Household level residuals* | | *Municipal level residuals* | | $R^2$ |
|---|---|---|---|---|---|
| | *Skewness* | *Kurtosis* | *Skewness* | *Kurtosis* | |
| Without | 10.10 | 177.00 | 2.09 | 9.87 | 0.31 |
| Logarithmic | −2.71 | 26.50 | −0.60 | 3.52 | 0.43 |
| Log-shift | 0.00 | 4.91 | −0.24 | 3.03 | 0.51 |
| Box–Cox | −0.24 | 7.95 | −0.12 | 3.00 | 0.49 |

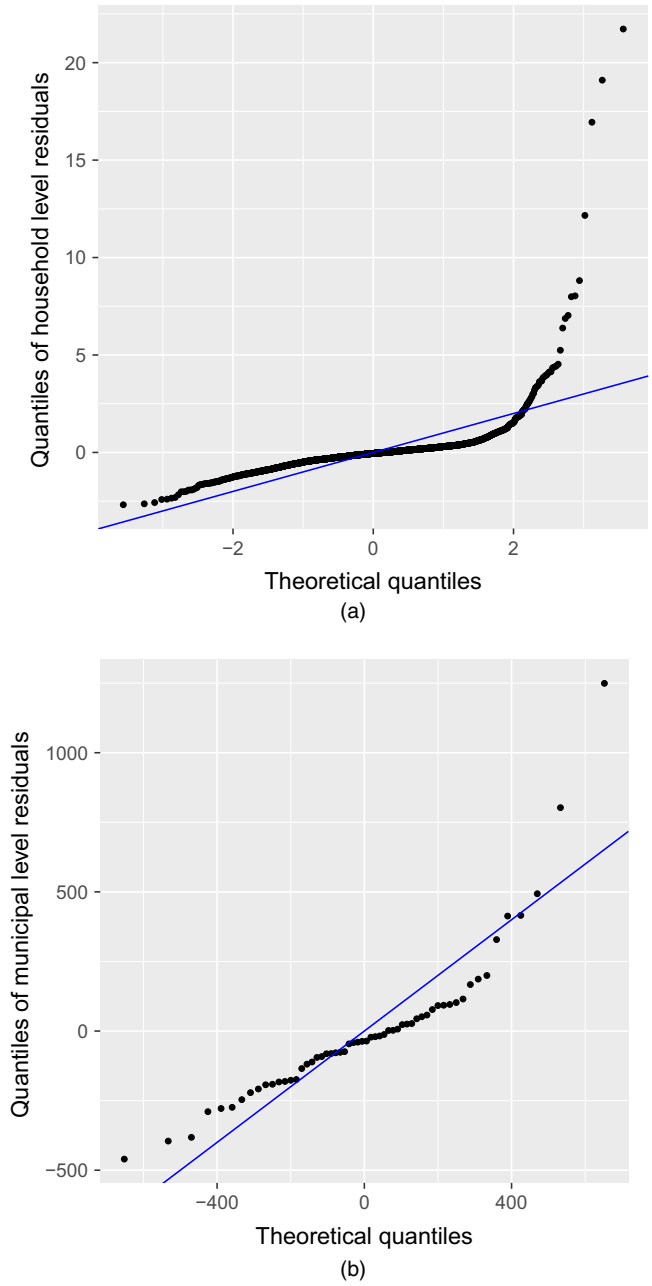



**Fig. 3.** Normal *Q*–*Q*-plots for (a) household level residuals and (b) municipal level residuals obtained from the model that uses raw income as the response variable

specification (Opsomer *et al.*, 2008; Ugarte *et al.*, 2009) and models that extend the covariance structure of the model by allowing for spatially correlated domain random effects (Pratesi and Salvati, 2009; Schmid *et al.*, 2016) or for complex variance structures (Jiang and Nguyen, 2012). An option—when diagnostic analysis shows departures from the model assumptions—and one that is based on the principle of parsimony is to find a transformation of the data such that the

normality assumptions of the EBP are met. Doing so means that the analyst can keep using standard estimation tools and software for SAE. The challenge in this case is in finding the most appropriate transformation. This adds another layer of complexity to the model building process. We now discuss the use of transformations in detail as an example of adapting the model. This is something that we encourage prospective users to explore before deciding to use more complex models.

Elbers *et al*. (2003) and Molina and Rao (2010) considered the use of a logarithmic or a logarithmic shift transformation, which are popular for income data. A better approach is to use data-driven transformations with optimally chosen parameters. Data-driven transformations may offer better predictive power and hence small area estimates with improved precision. For an illustration using the ENIGH data we consider the log-shift—with an optimally chosen shift—and the Box–Cox transformation (Box and Cox, 1964; Gurka *et al*., 2006). One key difference between the logarithmic and these additional transformations is that in the latter case the choice of transformation is adaptive, i.e. driven by the data. This is achieved by a transformation parameter, which is denoted by $\lambda$, which must be estimated. The logarithmic transformation is then a special case of this family of transformations when $\lambda = 0$. Denoting by $T_\lambda(y_{ik})$ the transformed outcome, the log-shift transformation is defined by

$$T_\lambda(y_{ik}) = \log(y_{ik} + \lambda). \tag{7}$$

The Box–Cox transformation is defined by

$$T_\lambda(y_{ik}) = \begin{cases} \dfrac{(y_{ik}+c)^\lambda - 1}{\kappa^{\lambda-1}\lambda}, & \lambda \neq 0, \\ \kappa \log(y_{ik}+c), & \lambda = 0, \end{cases} \tag{8}$$

for $y_{ik} > -c$, where $c$ is a fixed parameter, which makes the data positive to enable the use of the Box–Cox transformation, and $\kappa$ is the geometric mean of $y_{ik}$ (Box and Cox, 1964; Gurka *et al*., 2006). This is an example of a scaled transformation. Conditional on $\kappa$, the Jacobian of the transformation is 1. Using the scaling by the geometric mean enables the use of the likelihood function under the nested error regression model and, as a result, standard software for fitting this model with the transformed data can be used. This is consistent with the principle of parsimony. Different approaches have been proposed in the literature for estimating the optimal transformation parameter in linear models. These methods are mainly based on maximum likelihood theory. However, little attention has been paid to the use of these techniques with linear mixed models. Gurka *et al*. (2006) used Box–Cox transformations based on restricted maximum likelihood theory for the estimation of the power transformation parameter in linear mixed models. In addition, the minimization of a measure of the asymmetry such as the skewness of the residuals for the log-shift transformation has been discussed by Feng *et al*. (2016). An empirical approach for choosing $\lambda$ in transformation (7) is to define a grid of values for $\lambda$, to fit the nested error regression model by using each of the transformed outcomes $T_\lambda(y_{ik})$ and to select the transformation that makes the distribution of the residuals as close as possible to normal. Note, however, that here we deal with two sets of residuals and to our knowledge there is no formal approach to defining the distance from normality. Recent work by Rojas-Perilla *et al*. (2017) studied the use of various scaled transformations and estimation methods for $\lambda$ in SAE. A general algorithm for implementing the EBP method with power transformations is as follows.

*Step 1*: define a parameter interval for $\lambda$.
*Step 2*: set $\lambda$ to a value that is inside the interval.

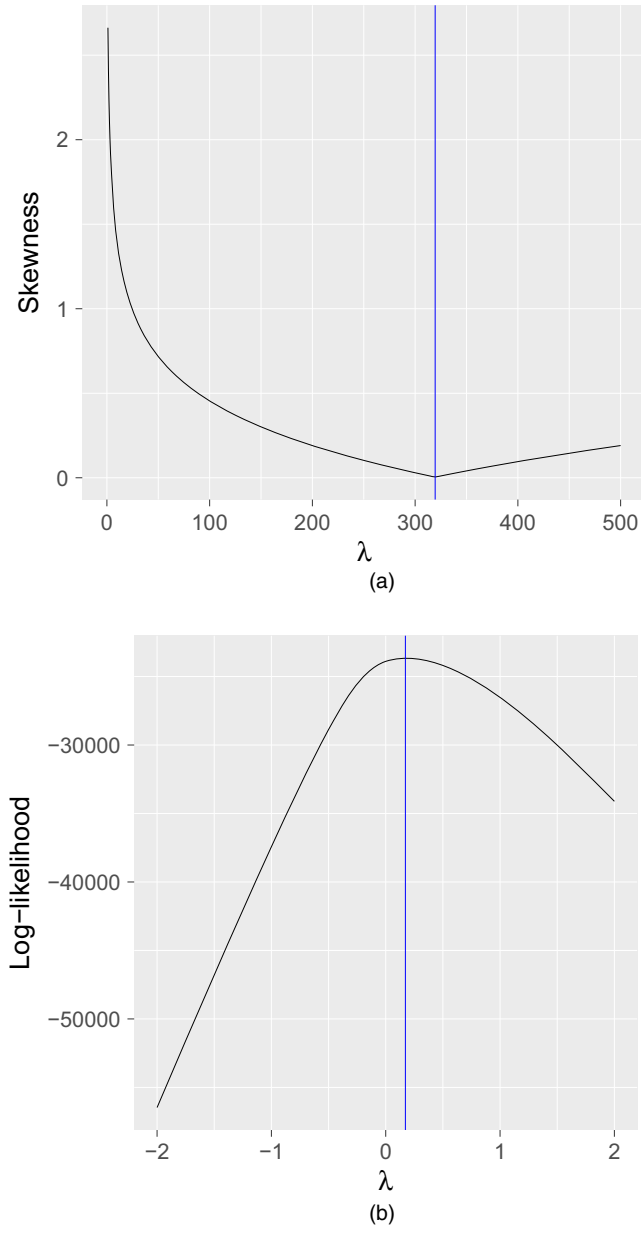



**Fig. 4.** Shift parameter for (a) the log-shift transformation and (b) optimal $\lambda$ for the Box–Cox transformation

*Step 3*: maximize the restricted log-likelihood function with respect to the vector of model parameters conditional on the fixed value of $\lambda$.
*Step 4*: repeat steps 3 and 4 until the value of $\lambda$ that maximizes the likelihood is found.
*Step 5*: apply the EBP method with the chosen value of $\lambda$.

Using the ENIGH data we apply the EBP method with three transformations for the outcome, namely the log-, log-shift and scaled Box–Cox. Fig. 4(b) shows the graphical representation of the maximization of the restricted maximal log-likelihood on a grid $\lambda \in [-2; 2]$ in the case of the

Box–Cox transformation. In this case the optimal $\lambda$ is approximately equal to 0.17. A similar graph in Fig. 4(a) shows the shift parameter that minimizes the skewness of the household level error term. The resulting parameter is equal to 319.52. The question is whether the use of the transformations that were identified above improves the diagnostic analysis and the predictive power of the model. We start with comments on the normal $Q$–$Q$-plots (Fig. 5) and the distribution of the residuals in Table 1. For municipality random effects, all three transformations offer a good approximation to normality (see also Table 1). The picture is different for household level. In particular, the household level residuals under the log-model show severe departures from normality. The situation is clearly improved when using the log-shift and power transformations (see also Table 1) with the log-shift transformation leading to less extreme and more symmetrical tails than do the other transformations.

To assess the assumption of homoscedasticity, we produce plots of the fitted values (the $x$-axis) against the standardized residuals (the $y$-axis) obtained by fitting model (4) by using the raw income data (Fig. 6(a)) and the Box–Cox power transformation (Fig. 6(b)). It can be observed that using transformations helps to stabilize the variance of the residuals. The corresponding plots for the log- and the log-shift transformations are similar.

The proportion of variability explained under each model is quantified by the coefficients of determination $R^2$ that are summarized in Table 1. Note that, as $R^2$ is computed on the basis of the transformed outcomes, the $R^2$-values are not directly comparable. As pointed out before, using the raw values of income in the EBP nested error regression model produces clearly unsatisfactory normal $Q$–$Q$-plots and $R^2 = 31\%$. The use of transformations improves the predictive power of the model for the transformed variables.

On the basis of the results from the diagnostic analysis we conclude that two transformations, namely the log-shift with shift parameter $\lambda = 319.52$ and the Box–Cox transformation with $\lambda = 0.17$, provide a better approximation to normality than the logarithmic transformation or the no-transformation cases, albeit not perfect. In particular, the symmetry of the distribution of the residuals is improved but the tails of this distribution remain heavier than those of the standard normal distribution. The following questions are raised at this stage. How important is the choice of transformation in SAE? Does the improvement in the predictive power of the model with transformation and less severe departures from the model assumptions translate to more precise small area estimates on the original scale? Is the choice of transformation equally important for parameters that are associated with the centre of the distribution and parameters that are associated with tails of the distribution? We attempt to address these questions in Section 4, which presents an evaluation framework for SAE. For now, we comment on Fig. 7 which shows maps of point estimates of average income, Gini coefficients and HCR for municipalities in EDOMEX produced by the EBP approach using different transformations.

The maps for average income, Gini coefficient and HCR clearly indicate regional differences. As mentioned before, EDOMEX has 125 municipalities which by their geographic and demographic characteristics are grouped into 16 districts. The maps of the estimated income-based indicators for all transformations suggest intraregional differences of poverty and inequality within and between the districts. Estimates of average income and HCR show that some of the wealthiest districts are concentrated in the central–east and northern zones of EDOMEX. The most unequal municipalities are in the central and south-west parts of EDOMEX. There are, however, some differences in the maps of point estimates that are produced with different transformations. Estimates of average income appear not to be affected by the choice of transformation. The same holds true to a large extent for estimates of HCR. In contrast, estimates of the Gini coefficient appear to be more sensitive to the choice of transformation. These results suggest that the user should be very careful with the choice of transformation as this can have

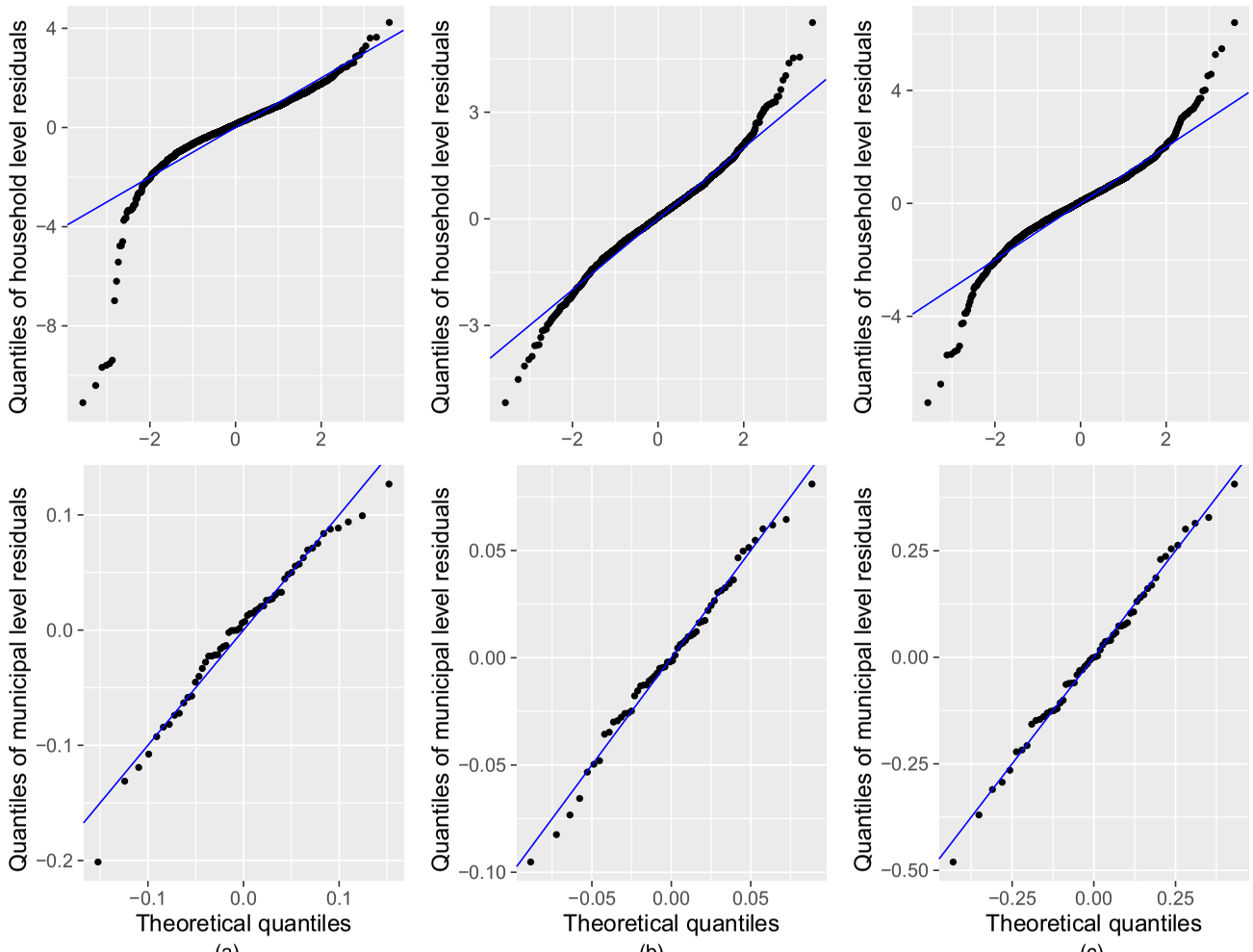



**Fig. 5.** Normal *Q*–*Q*-plots for household level residuals and municipal level residuals under the three transformations for income (a) log-transformation, (b) log-shift transformation and (c) Box–Cox transformation

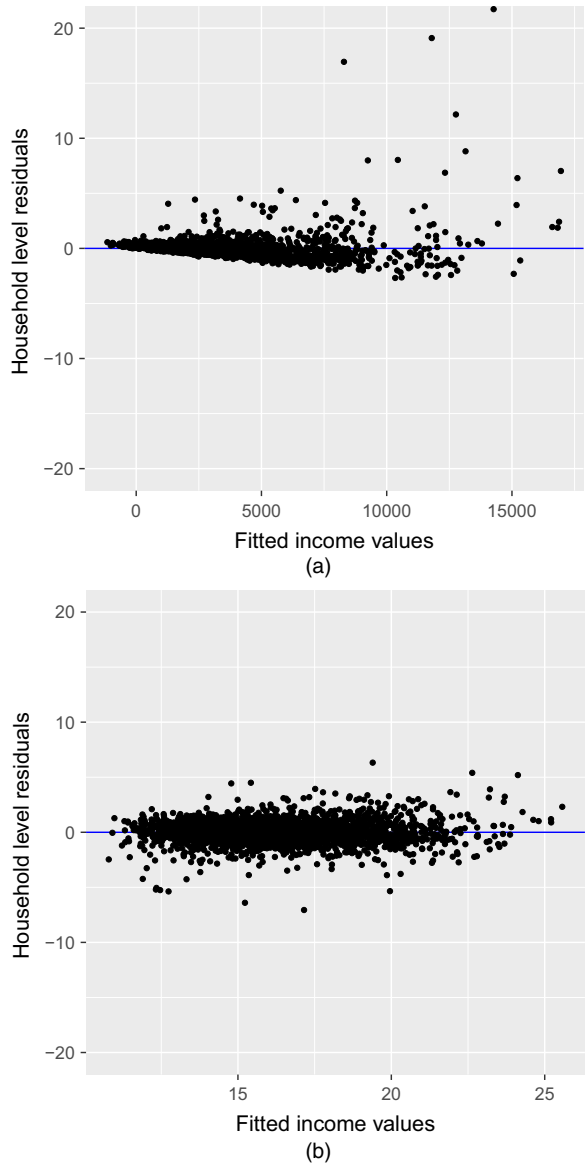



**Fig. 6.** Standardized household level residuals against fitted values (a) without and (b) with Box–Cox transformation for income

an influence on point estimation especially when interest is in non-linear indicators that depend on the entire distribution. We shall return to this discussion at the end of Section 4.

## 4. Evaluation

The small area estimates are a set of numbers of identical definition and simultaneous interest. Evaluating the small area estimates is a relevant question for which there are hardly any definitive answers. For example, whether to measure the uncertainty by using a design or a model-based

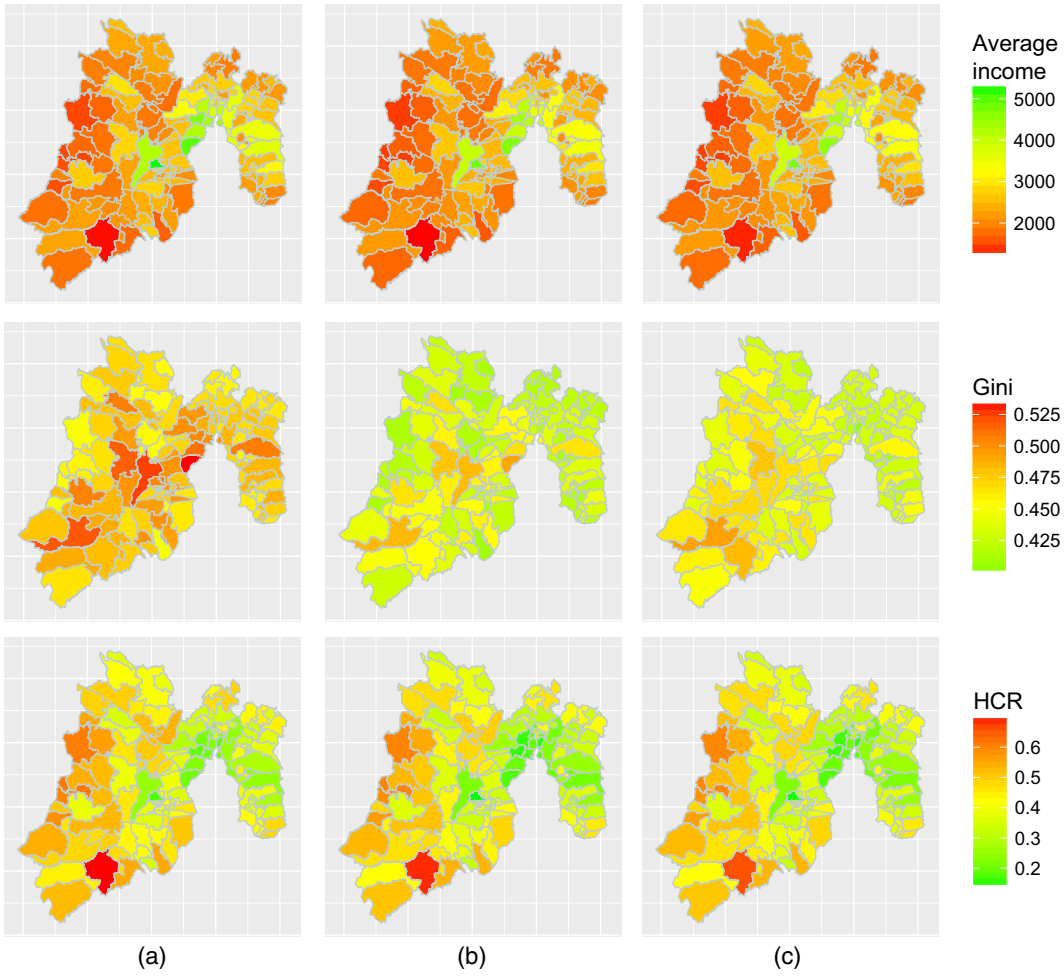



**Fig. 7.** Map of municipal estimates of average income, Gini coefficients and HCR in EDOMEX by using the EBP method under the (a) log-, (b) log-shift and (c) Box–Cox transformations

MSE causes lively debates among researchers and practitioners. Comparing sets of optimal small area estimates that are produced under alternative models and deciding whether one set is better than another can be a challenging task also. Assessing ensemble properties of small area estimates such as the range or ranks of the estimates is a relevant topic which has been largely overlooked. A detailed discussion on evaluation is beyond the scope of this paper. Our approach below is to describe some aspects of evaluation, which we believe should be taken into consideration in any application. In particular, we highlight the distinction between uncertainty assessment and method evaluation, which in our experience is a matter that is often either misunderstood or overlooked. The purposes of each and the most common uses in SAE are described in Section 4.1 and 4.2 respectively. Some illustrations with the ENIGH data are given in Section 4.3.

### *4.1. Uncertainty assessment*

Let $\theta_k$ be the target parameter of area $k$, for $k = 1, \ldots, m$. Let $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$ be the collection

of them. Let $\hat{\theta}_k$ be the estimator of $\theta_k$ and $\hat{\boldsymbol{\theta}}$ the collection of them. We assume that we are equally interested in all elements of $\boldsymbol{\theta}$ and cannot fix only on one particular $\theta_k$, or a few of them, and disregard how estimators perform in the rest of the areas.

The first question for uncertainty assessment is 'what is the target of estimation?', which refers back to the specification of the problem. Generally speaking, in SAE we may distinguish between the area-specific and ensemble targets of $\boldsymbol{\theta}$. An ensemble characteristic of $\boldsymbol{\theta}$ is defined by using all $\theta_k$s. For example, let $\bar{\theta}_w = \Sigma_{k=1}^m N_k \theta_k / N$ be the population mean, where $N_k$ is the population size in area $k$ and $N = \Sigma_{k=1}^m N_k$, or let $G = \Sigma_{k=1}^m (\theta_k - \bar{\theta})^2/(m-1)$ be the dispersion (i.e. population variance) of $\boldsymbol{\theta}$, where $\bar{\theta} = \Sigma_{k=1}^m \theta_k / m$. Other examples include the range, the order statistics and the ranks of $\boldsymbol{\theta}$. Although the various ensemble target parameters may be very important for purposes such as benchmarking, subgroup analysis, fund allocation, evaluation and monitoring (see for example Ghosh (1992) and Shen and Louis (1998)), area-specific prediction seems to have been the focus in the majority of applications. The most common uncertainty measure for area-specific prediction is the MSE. Below we explain the three types of MSE in use after which interval estimation will be briefly described.

Let $y_k$ denote generically all the observed data in area $k$, for $k = 1, \ldots, m$. Let $\mathbf{y} = \{y_1, \ldots, y_m\}$ be the collection of them. Given a population model for $\boldsymbol{\theta}$, the (unconditional) MSE is given by $E[(\hat{\theta}_k - \theta_k)^2]$, where the expectation is over both $\boldsymbol{\theta}$ and $\mathbf{y}$. Prasad and Rao (1990) developed a second-order accurate analytic MSE estimator under the linear mixed model, which corrects the bias of the direct plug-in MSE estimator. Jackknife methods have been developed for the same purpose under a wider range of models (Jiang *et al.*, 2002). The bootstrap (most commonly the parametric boostrap) is more generally applicable, especially if either the target parameter or the performance measure is non-differentiable (Hall and Maiti, 2006; Pfeffermann and Correa, 2012), such as when the target parameter is a population quantile.

Using the bootstrap is particularly relevant for uncertainty estimation of indicators such as the Gini coefficient and the HCR. For example, for the EBP method that was described in Section 3.2, simple unconditional MSE estimation uses the following parametric bootstrap, where the unknown model parameters are replaced by their estimates and treated as fixed. Generate $B$ bootstrap populations by using the fitted marginal model (4). Compute the population value of the target parameter from each bootstrap population, denoted by $\theta_k^*$. From each bootstrap population select a bootstrap sample and compute bootstrap estimates of the target parameter, $\hat{\theta}_k^*$, by using the same method as used with the original sample. Finally, compute the average of the $B$ squared bootstrap errors—defined as the difference between $\hat{\theta}_k^*$ and $\theta_k^*$—as an estimate of the unconditional MSE. Note that the procedure here is not second-order accurate, unlike the more sophisticated, but more computer-intensive, bootstrap methods that were cited above. In the case of using a transformation, the bootstrap populations are generated by using the model fitted to the transformed data but MSE estimates are computed at the end by back-transforming to the original scale. Estimation of the transformation parameter $\lambda$ should be implemented for each bootstrap sample, hence capturing the variability due to its estimation.

According to Booth and Hobert (1998), the conditional MSE of prediction (CMSEP) is given by $E[(\hat{\theta}_k - \theta_k)^2 | y_k]$, where the corresponding within-area $y_k$ is held fixed, and the pairs $(u_j, y_j)$ are independent across the areas, for $j = 1, \ldots, m$. They argued particularly for its use under generalized linear mixed models and elaborated their approach in terms of the linear predictor. When the model parameters are known, denoted by $\boldsymbol{\psi}$, the best predictor is $\tilde{\theta}_k = E[\theta_k | y_k; \boldsymbol{\psi}]$, and the only natural measure of its uncertainty is the CMSEP that reduces to the variance $V(\theta_k | y_k; \boldsymbol{\psi})$. When the model parameters are estimated, denoted by $\hat{\boldsymbol{\psi}}$, the CMSEP is decomposed into two terms $V(\theta_k | y_k; \boldsymbol{\psi})$ and $E[(\hat{\theta}_k - \tilde{\theta}_k)^2 | y_k; \boldsymbol{\psi}]$, where $\hat{\theta}_k = E[\theta_k | y_k; \hat{\boldsymbol{\psi}}]$. The first term is evaluated with respect to $u_k$ given $y_k$, and the second with respect to $\hat{\boldsymbol{\psi}}$ which varies with only the rest of the $y_j$s,

for $j \neq k$, given $y_k$, where $u_k$ and $\hat{\psi}$ are conditionally independent (Booth and Hobert, 1998). Lohr and Rao (2009) proposed a second-order accurate jackknife estimator of the conditional MSE. For a practical example, Zhang (2009) applied the CMSEP to estimates of small area compositions subjected to informative missing data.

The third type of MSE that we describe is given by $E[(\hat{\theta}_k - \theta_k)^2|\boldsymbol{\theta}]$, where only the observed data $\mathbf{y}$ are allowed to vary but the values of $\boldsymbol{\theta}$ are treated as fixed. The key difference from the two types of MSE above is that the set of small area parameters $\boldsymbol{\theta}$ is now held fixed, and for this reason we may refer to this MSE as the finite population MSE (FPMSE). There are several variations of the FPMSE in practice, where $\boldsymbol{\theta}$ may either be the actual population values or the theoretical values under a model, and the MSE may be evaluated with respect to the sampling design or a model for $\mathbf{y}|\boldsymbol{\theta}$. The FPMSE becomes the well-known design-based MSE, when $\boldsymbol{\theta}$ are population quantities such as the area means and $\mathbf{y}$ vary according to the sampling design (Rivest and Belmonte (2000), for example). Often, however, simplifying assumptions are adopted, e.g. by assuming area-stratified simple random sampling with the observed area sample sizes treated as fixed, because we may not have access to the details that are required to implement the sampling design. Chambers *et al*. (2011) calculated the FPMSE under the model for $\mathbf{y}|\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ are the theoretical area means rather than the population area means. Note that Chambers *et al*. (2011) used the term 'conditional' MSE, where it is the $\theta_k$s that are treated as fixed not $y_k$ as under the CMSEP. Finally, because the FPMSE is a small area parameter itself, unbiased estimation is unstable whether it is with respect to the sampling design or model. Hence, we need to treat the estimation of FPMSE as an SAE problem in its own right.

Deciding which MSE to use is important. Tukey's remark on this matter is that one should 'focus on the questions, not models' (in the discussion of Nelder (1977)). There are times when the target parameter $\theta_k$ is of a theoretical nature. It is then quite appropriate to consider the $u_k$s as random variables, and to use the unconditional MSE or the CMSEP as the uncertainty measure. For instance, in life expectancy calculation one would first smooth the actual known death rates, which could only make sense if one considers the actual population death rate as an estimate of some unknown hypothetical parameter called the mortality rate. But there are also many other situations, such as when $\theta_k$ is the area unemployment rate, where it is clearly defined as a descriptive statistic of the given population. We can still treat $u_k$ as a random effect to achieve a sensible bias–variance trade-off, e.g. by using model (4) to motivate a choice of $\alpha_k$ in the composite estimator (3). Without introducing the random-effects model, we would have to resort to other means for deriving $\alpha_k$. However, we believe that although it is inferentially consistent to report the model-based MSE here, which treats $\boldsymbol{\theta}$ as random, we are entitled to question its relevance when $\theta_k$ is a descriptive statistic and the assumption $E[u_k] = 0$ may be doubtful for a given $k$. In such a case, the FPMSE is attractive for many survey practitioners. However, as explained above, the estimation of the FPMSE needs to be treated as a small area problem in its own right.

Finally, interval estimation may be considered in addition to MSE estimation. Let $C_k = (\hat{\theta}_{k\mathrm{L}}, \hat{\theta}_{k\mathrm{U}})$ be an interval estimator of $\theta_k$, where $\hat{\theta}_{k\mathrm{L}} < \hat{\theta}_{k\mathrm{U}}$. The simplest procedure is to set the bounds such that $\hat{\theta}_k \pm 1.96\,\widehat{\mathrm{MSE}}(\hat{\theta}_k)^{1/2}$, aimed at the 95% nominal confidence level. See Pfeffermann (2013), section 6.2, for a review of interval estimation methods. Let $\delta_k$ equal 1 if $\theta_k \in C_k$ and 0 otherwise. Analogously to the unconditional MSE, the unconditional coverage of $C_k$ is given by $\varsigma_k = E[\delta_k] = P(\theta_k \in C_k)$, where both $\boldsymbol{\theta}$ and $\mathbf{y}$ are allowed to vary. Similarly, we can speak about conditional coverage of $C_k$ given by $E[\delta_k|y_k]$, and finite population coverage given by $E[\delta_k|\boldsymbol{\theta}]$. Note that any model-based $C_k$ that treats $\theta_k$ as random can have rather erratic area-specific finite population coverage compared with the nominal level of confidence. Zhang (2007) defined $\varsigma = \Sigma_{k=1}^{m} E[\delta_k|\boldsymbol{\theta}]/m$ to be the finite population simultaneous coverage of

all $C_k$, each aimed at the same nominal confidence level. For the population from which the sample is selected, this gives the proportion of area parameters that are expected to be covered by their interval estimates without specifying which areas these are. It is shown that, as $m$ increases, $\varsigma$ converges to the nominal level, provided that the underlying population model of $\boldsymbol{\theta}$ is correct.

### *4.2. Method evaluation*

In the previous section we described various measures of uncertainty. In addition to measuring the uncertainty that is associated with $\hat{\boldsymbol{\theta}}$ under the model assumed, an analyst may be interested in method evaluation. This might include comparing different point estimators, assessing how an MSE estimator performs in reality when approximations are used in its derivation or assessing how a small area estimator behaves under departures from the underlying model assumptions. Method evaluation is generally a different matter from uncertainty assessment.

As we describe below, broadly speaking method evaluation can be design based or model based. It is also possible to combine both sources of uncertainty, where the distribution of $\boldsymbol{\theta}$ follows from a population model and the distribution of $\mathbf{y}$ from the sampling design. The evaluation can be performed analytically provided that the required closed form expressions can be derived. More often, both design-based and model-based simulation studies are used for method evaluation.

Conducting a design-based simulation study is very common in practice. Indeed, it is difficult to imagine that an NSI will produce any small area statistics regularly without validating the design-based performance of the adopted method under realistic conditions. Typically, a census or similar population data set is fixed as the population from which samples are repeatedly taken. When such population data are unavailable, there are various proposals in the literature on how one can generate a pseudopopulation for the in-sample areas from the sample data at hand (e.g. Sverchkov and Pfeffermann (2004)). However, a model will be necessary to generate a pseudopopulation for the out-of-sample areas. For each simulated sample, a given estimation method is applied to obtain a replicate set of small area estimates. Within a design-based simulation study different estimation methods or models can be directly compared with each other in terms of their design-based performances. We consider this to be a suitable approach for method evaluation, which establishes how a method is expected to perform over repeated sampling from a finite population, regardless of whether the underlying model is correct or not. Using the ENIGH data in Section 4.3.2 we provide a detailed description of how one can design and implement a design-based simulation that mimics the design and characteristics of the survey data.

Unlike in a design-based simulation study, where the different estimation methods are subjected to the same sampling variation and the population may be based on real data, model-based method evaluation generally requires the use of a model for generating the population. This is common when researchers develop new methods and they are interested in evaluating the properties of estimators. The design of model-based studies requires careful thinking about the choice of the evaluation model that is used for generating the population. A general question is whether it is meaningful to compare directly the MSE of an estimator $\hat{\theta}_{k\mathrm{A}}$ of $\theta_k$ derived under model $M_{\mathrm{A}}$ with that of another estimator $\hat{\theta}_{k\mathrm{B}}$ of $\theta_k$ under model $M_{\mathrm{B}}$, which may involve different random effects or correlation structure. It is always possible to evaluate the MSE of $\hat{\theta}_{k\mathrm{A}}$ under model $M_{\mathrm{B}}$ even though the estimator is motivated and computed under model $M_{\mathrm{A}}$ and vice versa. Since the MSE of $\hat{\theta}_{k\mathrm{A}}$ will differ according to whether the evaluation model is $M_{\mathrm{A}}$ or $M_{\mathrm{B}}$, there is a need to level the ground to avoid misleading comparisons. We may, for example, carry out simulation of both $\hat{\theta}_{k\mathrm{A}}$ and $\hat{\theta}_{k\mathrm{B}}$ under model $M_{\mathrm{B}}$ if $M_{\mathrm{A}}$ is nested in $M_{\mathrm{B}}$. When $M_{\mathrm{A}}$ and $M_{\mathrm{B}}$ are not nested in each other but are from the same class of models, we may

use for the evaluation a model $M_C$ which encompasses both. But it may not be obvious how to find an encompassing model when $M_A$ and $M_B$ belong to different classes of models.

It should be mentioned that, in addition to the methods that were described above, there are several informal evaluation approaches that are of relevance to practitioners, such as compatibility with external data, evaluation by subject-matter experts, bias and goodness-of-fit diagnostics, as described in Brown *et al*. (2001). Finally, a set of small area estimates is expected to be numerically consistent and more efficient than unbiased direct estimates. One can compare the aggregated area estimates with the corresponding direct estimates for the same purpose. If aggregated model-based (indirect) estimates do not agree with the corresponding direct estimates, an analyst can use benchmarking techniques to achieve consistency. Benchmarked small area estimates offer an attractive property for NSIs (see Ghosh and Steorts (2013), Pfeffermann (2013) and Pfeffermann *et al*. (2014) for a discussion on benchmarking methods). A more challenging issue is benchmarking of aggregated ensemble properties, such as the population quantiles, which can be derived from the collection of within-area quantiles.

### *4.3. Illustrating aspects of small area estimation evaluation using the Encuesta Nacional de Ingresos y Gastos de los Hogares data*

In this section we illustrate some of the aspects of SAE evaluation that we discussed in Sections 4.1 and 4.2. In particular, using the results of model selection and diagnostics that we described in Section 3.3, we present results for the estimation of average household equivalized income, the HCR and Gini coefficients for municipalities with the original sample in EDOMEX. We then show how the analyst can prepare a design-based simulation study that can be used for method evaluation. We discuss how the design-based simulation results can guide the production of the final set of SAE estimates.

#### *4.3.1 Analysis with the original sample*

Table 2 presents summaries over municipalities of point, root MSE (RMSE) and CV estimates computed by using the original data supplied to us by CONEVAL and estimated MSEs under the model assumed. To start with, direct estimation is not considered because survey data cover only part of the target geography and—as we discussed in Section 3.3—direct estimates have higher-than-acceptable estimated CVs. Results are presented separately for in-sample and out-of-sample areas. For in-sample areas we produce estimates by using four versions of the EBP method, i.e. with untransformed income and three transformations (logarithmic, log-shift and Box–Cox). For out-of-sample areas we use the four above-mentioned versions of the EBP, which in this case corresponds to synthetic estimation. MSE estimates are obtained by using the parametric bootstrap under the unit level mixed models (see Section 4.1) and different transformations. The synthetic estimates are produced under the marginal model (6).

The results in Table 2 show that the EBP log-shift and EBP Box–Cox methods produce small area estimates that are clearly more efficient than the corresponding estimates produced with the untransformed income model and more efficient than the log-income model. Hence, using the methods that were suggested by model building and diagnostic analysis results in estimates with better efficiency. It is also clear that failing to use transformations, when needed, has an effect on point estimation. The effect of transformations on point estimation is less pronounced for indicators that relate to the centre of the income distribution (average income) than for non-linear indicators such as the HCR and the Gini coefficient. However, even for average income, failing to transform has a substantial effect on the efficiency of the estimates. These results illustrate the importance of model diagnostics in SAE. A final comment about these results relates to MSE estimation. MSE estimates are produced by computing the parametric

**Table 2.** One-sample analysis of income data: medians of point estimates, estimated RMSEs and CVs over municipalities in EDOMEX

| *Method* | *Results for 58 in-sample municipalities and the following indicators:* | | | *Results for 67 out-of-sample municipalities and the following indicators:* | | |
|---|---|---|---|---|---|---|
| | *Mean* | *HCR* | *Gini* | *Mean* | *HCR* | *Gini* |
| *Point estimates* | | | | | | |
| EBP | 2730 | 0.380 | 0.949 | 2042 | 0.436 | 1.261 |
| EBP logarithmic | 2699 | 0.363 | 0.477 | 2244 | 0.439 | 0.474 |
| EBP log-shift | 2600 | 0.329 | 0.433 | 2151 | 0.409 | 0.432 |
| EBP Box–Cox | 2617 | 0.336 | 0.435 | 2171 | 0.409 | 0.440 |
| *RMSE* | | | | | | |
| EBP | 449.2 | 0.040 | 0.177 | 523.4 | 0.048 | 0.400 |
| EBP logarithmic | 249.7 | 0.039 | 0.011 | 256.1 | 0.050 | 0.013 |
| EBP log-shift | 202.3 | 0.036 | 0.010 | 209.3 | 0.048 | 0.011 |
| EBP Box–Cox | 185.2 | 0.034 | 0.010 | 188.4 | 0.043 | 0.011 |
| *CV* | | | | | | |
| EBP | 0.163 | 0.104 | 0.187 | 0.251 | 0.114 | 0.313 |
| EBP logarithmic | 0.095 | 0.108 | 0.024 | 0.111 | 0.119 | 0.027 |
| EBP log-shift | 0.080 | 0.112 | 0.022 | 0.095 | 0.122 | 0.025 |
| EBP Box–Cox | 0.071 | 0.103 | 0.022 | 0.085 | 0.110 | 0.025 |

bootstrap estimator with the original sample. The parametric bootstrap relies on the belief that the model assumptions (after transformation) are met. In reality there are always departures from the model assumptions, the risk of which is uncontrollable for the out-of-sample areas in particular. One question is whether departures can have an effect on MSE estimation. Another question is whether the effect of model misspecification on MSE estimation is different for linear and non-linear indicators. The question becomes relevant when looking at the RMSE estimates for the Gini coefficient, which are quite small. Evaluating MSE estimation subject to model misspecification is not easy. Using evaluation methods such as design- or model-based simulations is essential. However, this can be very computer intensive because it requires bootstrap techniques to be embedded within a Monte Carlo simulation framework. We discuss this issue again in the next section.

### *4.3.2. Method evaluation by using design-based simulation*

In Section 4.3.1 above, the MSE was calculated under the model that was estimated on the basis of the ENIGH survey data. Naturally the user might be interested in knowing how the estimates will be affected if the model assumptions do not hold. Using design-based method evaluation that does not depend on the model assumptions can help with investigating this. We now illustrate an approach for setting up a design-based simulation that involves repeated sampling from a fixed population.

In a design-based simulation, the first and possibly the most important step is deciding how to generate the fixed population from which we draw repeated samples. Sverchkov and Pfeffermann (2004) suggested generating a pseudopopulation by using the sample data. In some cases a variable that is highly correlated with the target variable is available in the census. This is so with the census data from Mexico for which we identified variable *inglabpc*—earned *per capita* income from work—as being highly correlated with the variable of interest ictpc, which is

**Table 3.** Summary statistics over municipalities

| | *Minimum* | *1st quartile* | *Median* | *Mean* | *3rd quartile* | *Maximum* |
|---|---|---|---|---|---|---|
| inglabpc (census) | 0 | 1000 | 1700 | 2717 | 3000 | 100000 |
| ictpc (survey) | 0 | 1310 | 2142 | 3243 | 3518 | 98070 |
| Population size | 394 | 2759 | 6852 | 24820 | 16440 | 349100 |
| Sample size | 3 | 17 | 21 | 47.4 | 42 | 527 |

available only in the survey data. Variable inglabpc does not have the desired income definition and this is why SAE using ictpc is needed. However, for the purposes of method evaluation we are interested in using a variable that has similar distributional characteristics to those of the target variable and inglabpc can play this role. The first reason why we decided not to include inglabpc as a covariate in our small area model is because we wanted to use this variable for evaluation purposes. The second reason is that we wanted to illustrate method evaluation in a situation where the covariates explain a moderate part of the variance. Table 3 presents summary statistics for inglabpc (which was used in the design-based simulation) and ictpc (which was used in the one-sample analysis). The distribution of both variables is similar and the total *per-capita* income ictpc is generally higher compared with *per-capita* income from work inglabpc. In fact, if anything, the census variable inglabpc is even more skewed than the survey variable ictpc, which seems reassuring with respect to the robustness of the evaluation using the census variable. Our design-based simulation will be based on repeated sampling from the Mexican census microdata and modelling of proxy household income inglabpc.

From the fixed population we independently drew $T = 500$ samples. The samples are selected by using a single-stage stratified random sampling with strata defined by the 58 in-sample municipalities in the ENIGH survey. The number of households in each in-sample municipality is the same as the number of households in the ENIGH survey. This leads to a sample size of 2748 households with 58 in-sample municipalities and 67 out-of-sample municipalities as is the case with the ENIGH survey. Summary statistics of the sample and population sizes—over municipalities—are provided in Table 3.

Using each sample selected from the fixed population we compute estimates of average equivalized household income from work, the HCR, and Gini coefficient. For in-sample areas we calculate the direct estimator (2), the EBP based on different transformations and the World Bank estimator (Section 3.2), which is denoted by WB in Table 4. As we mentioned in Section 3.2, for out-of-sample areas and when domains coincide with clusters, the EBP and the World Bank method coincide. All the models use the same six covariates as identified in Section 3.3. The $R^2$ from linear regression models under different transformations (logarithmic, log-shift and Box–Cox) is around 40–50% over the 500 samples, which is consistent with the results that we obtained with the original sample.

The performance of these estimators is evaluated by computing the relative bias RB and RMSE given by

$$\mathrm{RB}(\hat{\theta}_k) = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\theta}_{tk} - \theta_k}{\theta_k},$$

$$\mathrm{RMSE}(\hat{\theta}_k) = \sqrt{\left\{\frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_{tk} - \theta_k)^2\right\}},$$

**Table 4.** Performance of predictors over municipalities in design-based simulations

| *Method* | *Results for 58 in-sample municipalities and the following indicators:* | | | *Results for 67 out-of-sample municipalities and the following indicators:* | | |
|---|---|---|---|---|---|---|
| | *Mean* | *HCR* | *Gini* | *Mean* | *HCR* | *Gini* |
| *RMSE* | | | | | | |
| EBP | 180.2 | 0.095 | 0.497 | 210.6 | 0.073 | 0.846 |
| EBP logarithmic | 187.5 | 0.049 | 0.026 | 216.3 | 0.061 | 0.032 |
| EBP log-shift | 156.6 | 0.038 | 0.022 | 200.7 | 0.062 | 0.031 |
| EBP Box–Cox | 171.7 | 0.045 | 0.025 | 212.6 | 0.060 | 0.032 |
| WB | 188.2 | 0.093 | 0.486 | — | — | — |
| WB logarithmic | 160.7 | 0.054 | 0.026 | — | — | — |
| WB log-shift | 159.4 | 0.041 | 0.022 | — | — | — |
| WB Box–Cox | 168.5 | 0.051 | 0.025 | — | — | — |
| Direct | 543.6 | 0.097 | 0.083 | — | — | — |
| *RB (%)* | | | | | | |
| EBP | 2.39 | 34.77 | 109.6 | 11.28 | −0.69 | 152.6 |
| EBP logarithmic | 2.96 | 12.54 | 3.89 | 12.43 | −5.27 | 2.25 |
| EBP log-shift | 0.93 | 6.49 | 0.08 | 11.19 | −9.86 | −0.21 |
| EBP Box–Cox | 1.98 | 11.18 | 2.32 | 11.91 | −6.60 | 1.09 |
| WB | 2.79 | 34.45 | 110.1 | — | — | — |
| WB logarithmic | 1.84 | 16.65 | 3.89 | — | — | — |
| WB log-shift | 0.80 | 9.59 | 0.10 | — | — | — |
| WB Box–Cox | 1.41 | 14.67 | 2.35 | — | — | — |
| Direct | −0.13 | −0.35 | −7.92 | — | — | — |
| *CV* | | | | | | |
| EBP | 0.082 | 0.262 | 0.534 | 0.109 | 0.179 | 0.693 |
| EBP logarithmic | 0.078 | 0.145 | 0.058 | 0.112 | 0.146 | 0.071 |
| EBP log-shift | 0.073 | 0.123 | 0.048 | 0.107 | 0.166 | 0.068 |
| EBP Box–Cox | 0.076 | 0.137 | 0.056 | 0.110 | 0.154 | 0.071 |
| WB | 0.088 | 0.260 | 0.530 | — | — | — |
| WB logarithmic | 0.072 | 0.174 | 0.058 | — | — | — |
| WB log-shift | 0.078 | 0.144 | 0.049 | — | — | — |
| WB Box–Cox | 0.074 | 0.161 | 0.055 | — | — | — |
| Direct | 0.239 | 0.291 | 0.203 | — | — | — |

where $\hat{\theta}_k$ is generic notation to denote an estimator of the target parameter in municipality $k$, $\theta_k$ denotes the true population parameter in municipality $k$ and $t$ is an index for repeated sampling with $T = 500$ in this case. We further report CV as an additional performance indicator.

Table 4 reports the results split by the 58 in-sample and the 67 out-of-sample municipalities. Table 4 presents median values of RMSE, relative bias and CV over municipalities. In line with the model diagnostics and the one-sample analysis, the performance of the EBP estimates without transformation is inferior to the EBP estimates with transformations (log-shift and Box–Cox) for all indicators. The design-based simulation results confirm that transformations are necessary for improved SAE. As expected, the direct estimator is less efficient than model-based estimators, which justifies the use of indirect methods in this case. A closer look at the EBP-based results with transformations shows that the EBP log-shift and the EBP Box–Cox methods perform somewhat better compared with the EBP logarithmic method in terms of bias and efficiency for all indicators. This indicates that the log-shift and the Box–Cox transformations adapt better to the shape of the underlying distribution, which appears

to be consistent with the results that we obtained from diagnostic analysis (Section 3.3). Comparing the EBP Box–Cox and the EBP log-shift methods in detail we note that in general neither transformation has superior performance over the other. Additional (model-based) simulation studies are necessary for comparing the performance of the Box–Cox transformation and the log-shift transformation. However, this is beyond the scope of the present paper but we refer to some research in this direction by Rojas-Perilla *et al*. (2017). For in-sample areas we note that the WB-estimates are somewhat less efficient than the EBP estimates. On the one hand, despite the relatively small between-area variability, including random effects is recommended for the in-sample municipalities. This can be seen from the increased biases of synthetic estimation for the out-of-sample areas. On the other hand, the relatively small difference between the WB- and EBP estimates highlights the importance of building a model that has a good fixed effects predictor. Doing so is of course also critical for the out-of-sample areas.

It is important to evaluate the performance of MSE estimators. Formal evaluation requires using a parametric bootstrap with each of the 500 samples, which is very computer intensive and beyond the scope of the present paper. Nevertheless, practitioners must be particularly careful when using parametric MSE estimation methods and, in our view, they should always employ design-based method evaluation.

Finally we give an illustration of informal evaluation. Comparing model-based estimates with corresponding design-based estimates for aggregated geographical levels can provide an indication about the quality of model-based estimates. As the Gini coefficient cannot be split into a weighted sum of subarea Gini coefficients, we focus on average income. The State of Mexico consists of 125 municipalities and 16 districts. The maximum sample size in a district is 749 households, the minimum is 18 households, the mean is 172 households and the median is 150 households per district. As the sample size is still quite small for some districts, we compare model-based estimates with design-based estimates for only 13 districts for which design-based estimates have a CV below 30%. Fig. 8 shows point estimates for district level average household equivalized income by using the direct estimator (the dotted curve) and the EBP estimators with logarithmic (the full curve), log-shift (the lighter chain curve) and Box–Cox (the darker chain curve) transformations. The direct estimates are produced by using the district-specific samples. In contrast, the district-specific model-based estimates are aggregated from the corresponding municipality level estimates. For the aggregation we used weights defined by $N_i/N$, where $N_i$ denotes the municipality population size. On the $x$-axis, districts are ordered by the CVs of the direct estimates (in descending order from left to right). We observe that, for districts where the direct estimates are more unreliable (the left-hand part of the plot), the model-based estimates are further from the direct estimates whereas, for districts where the design-based estimates are more reliable (the right-hand part of the plot), the EBP Box–Cox and EBP log-shift estimates tend to be closer to the direct estimates. The correlation between the direct and the EBP Box–Cox and EBP log-shift estimates is also slightly higher than the correlation between the direct and the EBP logarithmic estimates. We emphasize that this is an informal approach to evaluating the quality of model-based estimates and there is no rule of thumb about what is an acceptable level of correlation between model and design-based estimates. An alternative is to average the direct estimates and the corresponding model-based estimates over the smallest eight districts and the largest eight districts, and to compare the numbers, as an indication of the potential bias. The use of cross-validation, where some areas are left out of fitting the model and model-based estimates for these areas are compared with design-based estimates, offers a more structured approach to evaluation.

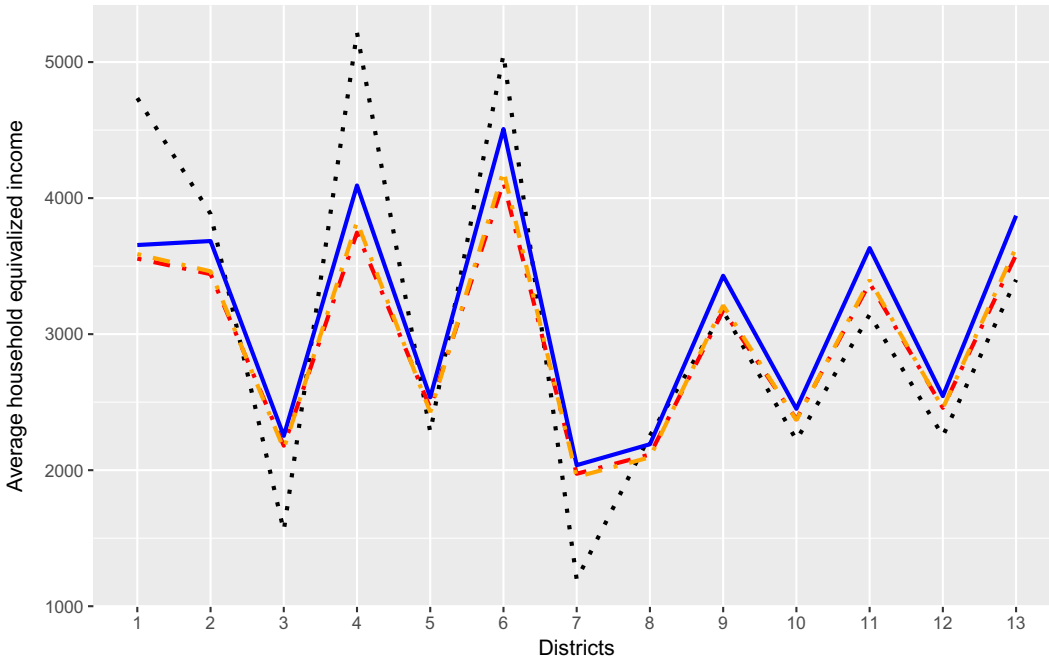



**Fig. 8.** Estimates for average household equivalized income at district level: ·····, direct; –·–·–·, EPB Box–Cox; ———, EBP logarithmic; –··–··–··, EBP log-shift

## 5. An update on small area estimation software

In this section we provide an update on the availability of SAE software. Although from an applied point of view many NSIs have a preference for software such as SAS, most of the recent developments in SAE have been implemented in the open source software R (R Core Team, 2015) via R packages.

A comprehensive review of relevant software is included in the Comprehensive R Archive Network task view on 'Official statistics and survey methodology' (Templ, 2015) with specific categories on complex survey designs, small area estimation and microsimulations. In particular, the section on complex survey designs includes packages, like `survey` (Lumley, 2004) and `sampling` (Tillé and Matei, 2012), that can be used for point and variance estimation of direct estimators of means, totals, ratios and quantiles under complex survey designs. Package `laeken` by Alfons and Templ (2013) provides functions for the estimation of various poverty and inequality indicators such as the at risk of poverty rate, Gini coefficient and quintile share ratio and the corresponding estimates of the variance. The `sae` package by Molina and Marhuenda (2015) can be used for computing synthetic and composite estimators and for implementing SAE with unit level and area (Fay–Herriot) models that allow for complex correlations structures. Code in R for computing EBP estimates that we discussed in Section 3.2 that includes an option for using the transformations discussed in the present paper, visualization and export of the results to Excel is proposed in the package `emdi` by Kreutzmann *et al.* (2018). Collections of R functions for implementing a wide range of SAE methods are available in the documentations of national- and European-funded research projects. Here we refer to the 'Bayesian methods for combining multiple individual and aggregate data sources in observational studies' project (BIAS, 2005) which includes code for the unit level empirical best linear unbiased

prediction and spatial empirical best linear unbiased prediction with correlated random effects (Pratesi and Salvati, 2009). The 'Small area methods for poverty and living condition estimates' project (SAMPLE, 2007) also provides a very wide range of code for implementing parametric, semiparametric and outlier robust SAE and allows for models with spatial and temporal correlations. We refer to Molina *et al*. (2010) for additional details. SAE from a Bayesian perspective is provided in the packages `hbsae` (Boonstra, 2012) and `BayesSAE` (Shi and Zhang, 2013). It is also important to mention two packages, namely `simPop` (Templ *et al*., 2017) and `saeSim` (Warnholz and Schmid, 2016), that support the prospective user in the set-up of design- or model-based simulations that enable method evaluation at the evaluation stage.

In addition to software written in R, alternative SAE software is also available. The World Bank provides open-source software for poverty estimation called `PovMap` (World Bank, 2013). `PovMap` implements the SAE procedure that was developed in Elbers *et al*. (2003) and is a standalone software solution. The European-funded project EURAREA (2001) delivered SAS code for the computation of direct and indirect small area methods. For additional procedures in SAS we refer to McDowell (2011). Finally, all the methods that were discussed in the paper are implemented by computationally efficient algorithms using R. The code is available from the authors on request.

## 6. Concluding remarks

In this paper we propose a general framework for the production of small area statistics and illustrate the SAE process in practice. As part of this framework we have touched on three interrelated topics, namely specification of the problem, analysis of the data and adaptation of the model, and method evaluation. Although much can be said for each of these three areas, it is the interplay between them that provides the key to the successful application of SAE methods. There are no clear-cut ways of trading between them in a formal manner and mastering a balance between these three stages is in many ways the wisdom of applied statistics, which holds true also for SAE. We have illustrated some practical ways of keeping this balance. It is shown that specifying a sensible geography and defining targets of estimation that are supported by the data that are available are the first important steps for successful SAE. Careful model building using the principle of parsimony, model diagnostics and model adaptations are crucial steps for improving estimation without the need for additional sources of data. Finally, obtaining uncertainty measures of good quality and designing method evaluation studies are of paramount importance for reassuring users especially if interest is in using the estimates for official purposes, e.g. in the design of policy interventions. SAE is of course a large research area and hence it is not possible to capture all of its aspects in a single paper. Production of small area statistics with discrete outcomes and use of area level models are not covered although the framework proposed can be applied in most cases.

Nevertheless, there are questions that remain unresolved and which we would like to raise at this stage. Within the context of sample surveys there is currently an apparent contrast between the prevalent preference for design-based approaches to statistics at the higher levels of aggregation and model-based approaches at the lower levels. This seems to imply that at some intermediate level of aggregation the choice between the two approaches may be somewhat blurred. Where are these intermediate levels of aggregation? Is it possible to develop a coherent framework for the different levels in the aggregation hierarchy? Should benchmarking towards aggregate level estimates of acceptable quality actively drive the development of SAE methods or should benchmarking remain a side issue that one only pays attention to at the last stage of estimation, as often it is?

Both area-specific and ensemble properties of a set of small area estimates are undoubtedly of interest. This is a distinctive feature of small area statistics in comparison with the national estimate that is a single number. SAE is a simultaneous rather than a point estimation problem. Multipurpose (multiple-goal) SAE aims to provide a compromise in a theoretical manner. However, the usefulness of such an approach can only be explored together with users if the solution is to have an influence in practice. Can users ever be ready or willing to accept multiple sets of estimates, each optimal for a particular purpose? How can we avoid or limit the misuses of a particular set of estimates in practice? For now we leave these questions open, hoping that they will inform future discussions.

## Acknowledgements

First of all, the authors are indebted to the Editor, Associate Editor and referees for comments that significantly improved the paper. Tzavidis, Zhang, Luna, Schmid and Rojas-Perilla gratefully acknowledge support by grant ES/N011619/1—'Innovations in small area estimation methodologies' from the UK Economic and Social Research Council. The authors are grateful to CONEVAL for providing the data that were used in the empirical work. The views that are set out in this paper are those of the authors and do not reflect the official opinion of CONEVAL. The numerical results are not official estimates and are produced only for illustrating the methods.

## References

Alfons, A. and Templ, M. (2013) Estimation of social exclusion indicators from complex surveys: the R package laeken. *J. Statist. Softwr.*, **54**, 1–25.
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. *J. Am. Statist. Ass.*, **83**, 28–36.
Bedi, T., Coudouel, A. and Simler, K. (2007) More than a pretty picture: using poverty maps to design better policies and interventions. World Bank, New York.
BIAS (2005) Bayesian methods for combining multiple individual and aggregate data sources in observational studies. (Available from `http://www.bias-project.org.uk/`.)
Boonstra, H. J. (2012) hbsae: hierarchical Bayesian small area estimation. *R Package Version 1.0*. (Available from `https://CRAN.R-project.org/package=hbsae`.)
Booth, J. and Hobert, J. (1998) Standard errors of prediction in generalized linear mixed models. *J. Am. Statist. Ass.*, **93**, 262–272.
Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.
Brewer, K. R. W. (1963) Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Aust. J. Statist.*, **5**, 93–105.
Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001) Evaluation of small area estimation methods—an application to unemployment estimates from the UK LFS. In *Proc. Symp. Achieving Data Quality in a Statistical Agency: a Methodological Perspective*. Ottawa: Statistics Canada.
Ceriani, L. and Verme, P. (2012) The origins of the Gini index: extracts from variabilità e mutabilità (1912) by Corrado Gini. *J. Econ. Ineqlty*, **10**, 421–443.
Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014) Outlier robust small area estimation. *J. R. Statist. Soc.* B, **76**, 47–69.
Chambers, R., Chandra, J. and Tzavidis, N. (2011) On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Surv. Methodol.*, **37**, 153–170.
Datta, G. S., Hall, P. and Mandal, A. (2011) Model selection by testing for the presence of small-area effects, and application to area-level data. *J. Am. Statist. Ass.*, **106**, 362–374.
Datta, G. and Lahiri, P. (1995) Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *J. Multiv. Anal.*, **54**, 310–328.
Elbers, C., Lanjouw, J. and Lanjouw, P. (2003) Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355–364.
El-Horbaty, Y. (2015) Model checking techniques for small area estimation. *PhD Thesis*. University of Southampton, Southampton.
Encuesta Nacional de Ingresos y Gastos de los Hogares (2010) Encuesta Nacional de Ingresos y Gastos de los

Hogares 2010. *Enigh Diseño Muestral*. (Available from http://www.beta.inegi.org.mx/app/biblioteca/ficha.html?upc=702825002420.)
EURAREA (2001) Enhancing small area estimation techniques to meet European needs. (Available from http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html.)
Eurostat (2012) Small area estimation. Eurostat, Luxembourg. (Available from http://ec.europa.eu/eurostat/cros/content/sae-finished_en.)
Fabrizi, E., Salvati, N., Pratesi, M. and Tzavidis, N. (2014) Outlier robust model-assisted small area estimation. *Biometr. J.*, **56**, 157–175.
Fabrizi, E. and Trivisano, C. (2016) Small area estimation of the Gini concentration coefficient. *Computnl Statist. Data Anal.*, **99**, 223–234.
Feng, Q., Hannig, J. and Marron, J. S. (2016) A note on automatic data transformation. *Stat*, **5**, 82–87.
Ghosh, M. (1992) Constrained Bayes estimation with applications. *J. Am. Statist. Ass.*, **87**, 533–540.
Ghosh, M., Maiti, T. and Roy, A. (2008) Influence functions and robust Bayes and empirical Bayes small area estimation. *Biometrika*, **95**, 573–585.
Ghosh, M. and Steorts, R. C. (2013) Two-stage benchmarking as applied to small area estimation. *Test*, **22**, 670–687.
Gini, C. (1912) Variabilità e mutabilità: contributo allo studio e delle distribuzioni e relazioni statistiche. Studi Economico-Giuridici della R, Università di Cagliari, Cagliari.
Gurka, M. J., Edwards, L. J., Muller, K. E. and Kupper, L. L. (2006) Extending the Box–Cox transformation to the linear mixed model. *J. R. Statist. Soc.* A, **169**, 273–288.
Hajek, J. (1958) On the theory of ratio estimates. *Apl. Mat.*, **3**, 384–398.
Hall, P. and Maiti, T. (2006) On parametric bootstrap methods for small area prediction. *J. R. Statist. Soc.* B, **68**, 221–238.
Jiang, J., Lahiri, P. and Wan, S. (2002) A unified jackknife theory for empirical best prediction with m-estimation. *Ann. Statist.*, **30**, 1782–1810.
Jiang, J. and Nguyen, T. (2012) Small area estimation via heteroscedastic nested-error regression. *Can. J. Statist.*, **40**, 588–603.
Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Tzavidis, N. and Templ, M. (2018) The R package emdi for estimating and mapping disaggregated indicators. *J. Statist. Softwr.*, to be published.
Lohr, S. and Rao, J. (2009) Jackknife estimation of mean squared error of small area predictors in non-linear mixed models. *Biometrika*, **96**, 457–468.
Lumley, T. (2004) Analysis of complex survey samples. *J. Statist. Softwr.*, **9**, 1–19.
Marhuenda, Y., Molina, I., Morales, D. and Rao, J. N. K. (2017) Poverty mapping in small areas under a twofold nested error regression model. *J. R. Statist. Soc.* A, **180**, 1111–1136.
Molina, I. and Marhuenda, Y. (2015) sae: an R package for small area estimation. *R J.*, **7**, 81–98.
Molina, I., Morales, D., Pratesi, M. and Tzavidis, N. (2010) Final small area estimation developments and simulations results. *Research Project Report Deliverables D12 and D16*, *EU-FP7-SSH-2007-1 SAMPLE*. (Available from http://www.sample-project.eu/.)
Molina, I. and Rao, J. N. K. (2010) Small area estimation of poverty indicators. *Can. J. Statist.*, **38**, 369–385.
Mukhopadhyay, P. and McDowell, A. (2011) Small area estimation for survey data analysis using SAS software. *SAS Global Forum 2011*.
Nelder, J. A. (1977) A reformulation of linear models. *J. R. Statist. Soc.* A, **140**, 48–77.
Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008) Non-parametric small area estimation using penalized spline regression. *J. R. Statist. Soc.* B, **70**, 265–286.
Pfeffermann, D. (2013) New important developments in small area estimation. *Statist. Sci.*, **28**, 40–68.
Pfeffermann, D. and Correa, S. (2012) Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika*, **99**, 457–472.
Pfeffermann, D. and Sikov, A. (2011) Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *J. Off. Statist.*, **27**, 181–209.
Pfeffermann, D., Sikov, A. and Tiller, R. (2014) Single- and two-stage cross-sectional and time series benchmarking procedures for small area estimation. *Test*, **23**, 631–666.
Pinheiro, J. and Bates, D. (2000) *Mixed-effects Models in S and S-Plus*. New York: Springer.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2018) nlme: linear and nonlinear mixed effects models. *R Package Version 3.1–137*. (Available from https://CRAN.R-project.org/package=nlme.)
Prasad, N. G. N. and Rao, J. N. K. (1990) The estimation of the mean squared error of small area estimators. *J. Am. Statist. Ass.*, **85**, 163–171.
Pratesi, M. and Salvati, N. (2009) Small area estimation in the presence of correlated random area effects. *J. Off. Statist.*, **25**, 37–53.
Rao, J. N. K. and Molina, I. (2015) *Small Area Estimation*, 2nd edn. New York: Wiley.
R Core Team (2015) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rivest, L.-P. and Belmonte, E. (2000) A conditional mean squared error of small area estimators. *Surv. Methodol.*, **26**, 67–78.

Rojas-Perilla, R., Pannier, S., Schmid, T. and Tzavidis, N. (2017) Data-driven transformations in small area estimation. *Discussion Paper 30|2017*. School of Business and Economics, Freie Universität Berlin, Berlin.
SAMPLE (2007) Small area methods for poverty and living condition estimates.
Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer.
Schmid, T., Bruckschen, F., Salvati, N. and Zbiranski, T. (2017) Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *J. R. Statist. Soc.* A, **180**, 1163–1190.
Schmid, T., Tzavidis, N., Münnich, R. and Chambers, R. (2016) Outlier robust small area estimation under spatial correlation. *Scand. J. Statist.*, **43**, 806–826.
Shen, W. and Louis, T. A. (1998) Triple-goal estimates in two-stage hierarchical models. *J. R. Statist. Soc.* B, **60**, 455–471.
Shi, C. and Zhang, P. (2013) BayesSAE: Bayesian analysis of small area estimation. *R Package Version 1.0-2*. (Available from `https://CRAN.R-project.org/package=BayesSAE`.)
Sinha, S. K. and Rao, J. N. K. (2009) Robust small area estimation. *Can. J. Statist.*, **37**, 381–399.
Snijders, T. and Bosker, R. (2012) *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.
Sverchkov, M. and Pfeffermann, D. (2004) Prediction of finite population total based on the sample distribution. *Surv. Methodol.*, **30**, 79–92.
Templ, M. (2015) CRAN task view: Official statistics and survey methodology. (Available from `https://cran.r-project.org/web/views/OfficialStatistics.html`.)
Templ, M., Meindl, B., Kowarik, A. and Dupriez, O. (2017) Simulation of synthetic complex data: the R package simPop. *J. Statist. Softwr.*, **79**, no. 10, 1–38.
Tillé, Y. and Matei, A. (2012) sampling: survey sampling. *R Package Version 2.5.*
Tzavidis, N., Marchetti, S. and Chambers, R. (2010) Robust estimation of small area means and quantiles. *Aust. New Zeal. J. Statist.*, **52**, 167–186.
Ugarte, M., Goicoa, T., Militino, A. and Durban, M. (2009) Spline smoothing in small area trend estimation and forecasting. *Computnl Statist. Data Anal.*, **53**, 3616–3629.
Vaida, F. and Blanchard, S. (2005) Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.
Warnholz, S. and Schmid, T. (2016) Simulation tools for small area estimation: introducing the R package saeSim. *Austr. J. Statist.*, **45**, 55–69.
Weidenhammer, B., Tzavidis, N., Schmid, T. and Salvati, N. (2014) Domain prediction for counts using microsimulation via quantiles. In *Proc. Small Area Estimation 2014 Conf., Poznan*.
World Bank (2013) Software for poverty mapping. World Bank, New York. (Available from `http://go.worldbank.org/QG9L6V7P20`.)
Zhang, L.-C. (2007) Finite population small area interval estimation. *J. Off. Statist.*, **23**, 223–237.
Zhang, L.-C. (2009) Estimates for small area compositions subjected to informative missing data. *Surv. Methodol.*, **35**, 191–201.

## Discussion on the paper by Tzavidis, Zhang, Luna, Schmid and Rojas-Perilla

**Ian R. Gordon** (*London School of Economics and Political Science*)

I very much welcome this ambitious and timely paper and appreciate the opportunity to introduce discussion of some of the issues it raises. It sets itself a commendably broad task, in terms of both the holism of its first-to-last perspective and the generality of domains and situations which its guidance should frame. Practically it moves us towards these goals by fulfilling two very different kinds of function.

First and foremost, it documents an exemplary, rigorous and *experimental* exercise in generating small area income estimates in a situation where direct estimates (of varying precision) are available for about half the cases and lacking for the other half, and where an inequality measure (the Gini coefficient) is required as well as one of average conditions. But, crucially, it also opens up key questions about credible best practice approaches in a burgeoning area of activity, emphasizing the importance of the user–statistician interface at stages from recognition of need to actual use of the constructed estimates.

There are evident tensions between these aspects—with more definitive achievement on the former, but plenty of grist for productive debate on the latter (notably about how users and statisticians can effectively collaborate) to be found between the lines of the paper as well as in its direct treatment of the issue. A tight experimental design (including withholding of the best predictor variable from the modelling stage for use in the subsequent evaluation) reflects a research context in which users were scarcely involved. But it encourages conjecture (for this discussant) on some positive differences that user engagement might have made, both to this specific case and to the design of a framework.

Those I would like to raise reflect my own professional position, as an artisanal statistician or semi-

professional user, and disciplinary affiliation, as a geographer. These lead me to a couple of substantial reservations about the purely technical framework guiding the case exercise. The first relates to the very cursory treatment of the choice and definition of independent variables deployed to generate the modelled estimates. As a specific example, I am struck by the fact that, despite repeated emphasis on the non-linear basis of the Gini coefficient as a source of problems, any real difficulties with this lie in a need to ensure predictor variables (or transformations of these) that can effectively target the two tails of the income distribution, which Fig. 5's plots of residuals show not to have been achieved at either end. Any useful practice guidance manual on small area estimation needs to point up the importance of addressing such substantive specification issues for particular situations, as well as offering the more generalized ('textbook'-type) suggestions for suitably robust estimation and evaluation techniques.

The other issue of substance I would raise about the 'small area' estimation (and evaluation) process reported here involves an almost complete lack of attention to geography—give or take some maps suggesting a need for it to be attended to. How far geography actually matters in any case, and how effectively it may already have been picked up by other independent variables, cannot be prejudged. But, in principle, it is clear that local income levels (or other statistics) are liable to be substantially affected by location, spatial dependence, settlement or agglomeration size and density, and ecological influences from the mix of population and economic activities. (A clear local example of the last of these might be the fact that statisticians living in upper middle class districts around the financial district of the City of London are likely to have substantially higher incomes than those in working-class suburbs!) All of these possible factors are passed over in the paper, including the very strong likelihood that those municipalities lacking direct estimates (because no local residents found their way into a national sample) will differ significantly in ways (e.g. settlement size) that require attention to sample selection bias in the small area estimation procedure. The substantial (positive) bias in mean income estimates for out-of-sample municipalities (*averaging* 11–12% according to Table 4) suggests that to be a real factor in this case—presumably because places with fewer residents are liable both to be uncovered and to have lower incomes (*ceteris paribus*).

In spelling out this pair of reservations, my intention is not to downplay the real strengths and technical sophistication of this paper, nor the impetus it gives towards developing a practice-related framework (or 'good practice' manual) for both users and suppliers of small area statistics. Indeed, I think the current burgeoning of work, encouraged by 'big data' sources, geographical information system technology and better geodata referencing—with the potential for Gresham's law effects, as the range of producers and outlets extends—rather urgently requires an initiative of that kind. But the kind of generality it aims for should (in my judgement) prioritize adequacy to context and substance, rather than specifying universally applicable techniques.

In working towards this, we might usefully debate a couple of general issues which the paper raises (for this reader). One is about what the real value is of the procedural '*parsimony*' that the paper advocates. For me it lies very largely in the kinds of transparency that both statisticians and users require to generate and adopt suitable small area measures or indicators—which deserve to be addressed more directly. The other is about the helpfulness or otherwise of a current assumption (by international organizations as well as in this paper) that area estimates should have *microdata foundations*. For me this seems rather a distraction in a task where causality is not an obvious issue—unless there are some reasons to expect that there might be important interaction effects (e.g. perhaps for predicting numbers in the tails of the income distribution, particularly since ecological effects are entirely apposite in this context.

For this timely stimulus, as well as its professionalism, I am very pleased to move the vote of thanks.

**Paul Williamson** (*University of Liverpool*)

Small area estimation (SAE) involves a mix of potentially complex and technical statistical techniques, pragmatic compromise and experimental iteration. In their paper Tzavidis and his colleagues are to be congratulated for capturing and condensing this process into a clearly elucidated framework without shying away from an airing of the techniques, compromises and choices involved. The key contributions of the paper include

(a) the clear framework provided for the SAE process and
(b) an emphasis on parsimony.

Of course, no one paper can cover all of the issues involved in the process of creating small area estimates. In what follows I highlight a few key points of omission or debate that I believe are worthy of attention.

*Problem specification*

The paper focuses on *official* statistics, which are produced by statisticians who have privileged access to

census and administrative microdata. This permits the adoption of SAE approaches that are not available to others. The coverage of the paper rightly reflects this, but thereby introduces a bias in the approach taken to problem specification.

In addition, the estimate precision required by most end-users is far lower than that implemented by most national statistical institutes (NSIs). See, for example, the widespread adoption of the relatively low precision tools of geodemographics and customer segmentation to target marketing and resource allocation (Harris *et al*., 2005; Tsiptsis and Chorianopoulos, 2010; Voas and Williamson, 2001).

The proportion of estimation areas that are out of sample also changes the problem posed. As the percentage of areas out of sample increases, the focus needs to switch from design-based to model-based solutions.

*Model selection*

The paper rightly advocates a parsimonious approach to model selection, favouring the use of simple models where possible. One area left unaddressed is the level at which models should be fitted: at the unit or area level. The unspoken rule of thumb is to model at the level for which the outcome is required. Hence area level outcomes are normally better modelled by using area level covariates (Heady and Ralphs, 2004; Moon *et al*., 2014; Williamson, 2016).

The World Bank and empirical best predictor approaches discussed in the paper appear to breach this 'rule', using unit level models to make individual level predictions which are then aggregated to the required area level outcomes. This is possible because the use of census data permits unit level imputation of the outcome with 100% population coverage. But how low can population coverage fall before these approaches are no longer appropriate? And are there other statistical consequences?

*Evaluation and purpose*

The paper acknowledges that many users will have more interest in 'ensemble' estimates (e.g. relative rankings of areas) than in 'point' estimates (area-specific values), in an echo of Williamson (2007). This is amply demonstrated by the widespread adoption in England of the index of multiple deprivation to target funding, with the focus being on the areas ranked as most deprived, not on areas scored below some absolute index value (Department for Communities and Local Government, 2015). Despite this the focus of the paper, and NSIs more generally, remains almost exclusively on the production and quality of point estimates.

*Evaluation and precision*

The paper advocates a bootstrap approach to assessing estimate uncertainty, allied with design-based simulation where appropriate. Unfortunately this is easier said than done. Bootstrapping is itself subject to controversy, with the superiority of one or other bootstrap approach depending on the circumstances (see Carpenter and Bithell (2000)). In addition, bootstrapping has a high computational overhead: one too high to allow the paper to complete one of the desired elements of evaluation. Here I feel that the paper sets an unrealistically high bar, departing somewhat from the more pragmatic and parsimonious tone adopted elsewhere. A similar charge applies to the discussion of mean-squared error estimators, where three alternatives are considered, ultimately highlighting why only one is commonly used in practice.

*Spatial analysis*

The paper treats any spatial variation not explained by population composition as a simple random area term. This overlooks the growing overlap between the fields of SAE and spatial analysis (see Schabenberger and Gotway (2005) and Rao and Molina (2015)). Spatial structure (spatial auto-correlation) can be addressed through the inclusion of spatial contiguity, spatial lag and non-compositional area characteristics in the modelling process.

*Conclusion*

SAE remains a relatively specialist and ghettoized activity. This is in stark contrast with standard and more advanced regression techniques, including general linear models and multilevel regression, all of which have been made readily accessible to non-specialists via an extensive range of on-line and printed resources (e.g. Crawley (2015) and Field (2018)). In contrast SAE techniques remain highly impenetrable to outsiders (see Longford (2005) and Rao and Molina (2015)).

By reaching out to NSI practitioners this paper provides a welcome, if partial, bridge from SAE academics to that wider audience. To build this bridge further, as the paper clearly identifies, a continued emphasis on parsimony is key, particularly bearing in mind that, basic techniques mastered, most advances in SAE offer only marginal gains.

The vote of thanks was passed by acclamation.

**Thomas King** (*Newcastle upon Tyne*)

Maps implying homogeneous regions are frustrating, so local augmentation with auxiliary data will smooth the estimates but can obscure the information which finer geographical detail demands. The three stages are elaborated, yet Tzavidis and his colleagues specify and evaluate the statistics that users want, rather than the decisions they want to make with them. Whereas the statistical properties of estimates of change over time are considered, the validity of the model assumptions for local uses needs consideration.

Challenging users and asking what they really need locally, rather than providing national data at a local level, requires greater local analytical capability than typical resources. Research about how users can engage with this sort of estimate, including uncertainty etc., and case-studies or tutorials would be of great benefit. These would also highlight some *caveats* to the expected utility which seems not to consider why fine resolution is desired. A shift to uncertain model-based proportions rather than (deterministic) counts means a cultural change, and planning or consulting on data needs in advance. This has been described as 'data maturity', in an organization culture sense (Centre for Data Science and Public Policy, 2016)—something this work can only highlight the need for.

Cycling statistics provide an example from my own experience: transport is inherently geographic, and detailed to the area which includes the journeys. For cycling, this is typically an urban region such as a city, but data about travel behaviours are collected in the English National Travel Survey. From this the government nationally aimed for short trips to be made by bicycle, and this target was adopted locally (Newcastle City Council, 2011). Statistics are published at subregional level, so no official data are available for local authorities who hold the delivery budgets.

For the small area estimates to be used to understand progress and to evaluate local investment, they need to be able to carry through a signal. The plans are for different progress from nationally so the national data alone are insufficient: the auxiliary data used, and the model specification, need to transmit information. Automatic counters can provide information about flow past a point but lack requisite information about trip length let alone purpose or demographics; similarly the coverage of geolocated smartphone data. Information integration (Judson, 2007) would be necessary, and to be facilitated for quality assurance and lack of local capability: in contrast with the richer background picture for decisions described.

**Maria Giovanna Ranalli** (*University of Perugia*)

I congratulate Tzavidis and his colleagues for having engaged in the not-simple task of proposing a framework for the production of small area official statistics. This is particularly valuable because small area estimation (SAE) has advanced along a very different path from the usual design-based standardized production process of official statistics: SAE methods have been developed and tailored to address specific estimation problems. For this reason, the development of a general framework is particularly challenging and I find that the authors have succeeded in providing a fairly general proposal, although the treatment is particularly suited to welfare variables.

Any researcher who has engaged in the production of small area statistics knows that being successful in the first 'specification' stage is usually more complicated than in the other two stages. This stage requires much expertise, flexibility and foresight. The second stage 'analysis and adaptation' is proposed to be governed by the principle of parsimony: this is a fundamental pillar of a successful framework for SAE, where the tempting availability of very complex models must be traded off with the robustness of the results and the opportunity of conveying them to the users. However, I also believe that *everything should be made as simple as possible, but no simpler* (Albert Einstein). In this regard, since SAE is a prediction problem, I appreciate the use of adaptive transformation of the data, i.e. transformation driven by the data. In the same adaptive spirit, it is possible to include covariates non-parametrically for both model building and model evaluation. Using $P$-splines, this elaboration of the model embeds quite naturally in the framework proposed (Opsomer *et al.*, 2008).

The authors do not discuss tools to address the potential bias introduced by the sampling design in model-based SAE. A simple way to assess whether the sampling design is informative and that can be embedded easily in the proposed framework in the model building phase is to test whether including the sampling weights, and their interaction with the covariates, provides a better fitting model (see for example DuMouchel and Duncan (1983) and Fuller (2009), chapter 5). Alternatively, available design variables (like stratification indicators or measures of size) can be inserted in the fixed part of the model. Note that the inclusion of variables related to the sampling design or to higher levels of aggregation (like strata or larger geographical regions) provides a first simple tool towards benchmarking (Wang *et al.*, 2008).

**Paul A. Smith** (*University of Southampton*)
I congratulate Tzavidis and his colleagues on a thought-provoking paper, and I especially like their focus on the user requirement in the specification stage of their three-stage framework, which argues for small area estimation to be considered alongside the other design requirements (a rare occurrence) rather than being wheeled out at the end when direct estimation does not meet user requirements (which is all too frequent). But thereafter there are many practical aspects of the implementation of small area estimation for official statistics which should be included in a framework. I shall use three specific points as illustrations. First, in the Mexican example it is not possible to refine the user requirement, as it is specified by law; but it is possible to adjust the design of the survey components (sample size, clustering and geographical spread) to affect the quality of the final estimates. The authors suggest examining the data availability but do not include *influencing* the data availability in their framework.

Second, the initial triplet of estimates is a very simple and nice idea to help to understand the usefulness of the available data to meet users' requirements. But even here there are choices—the synthetic estimator requires a larger area (than the target) within which the model can be fitted, but there is no guidance on what is a good choice. Presumably this larger area should be as small as possible, consistent with meeting some minimum accuracy criteria, but can this be formalized, even as a rule of thumb?

Third, the ensemble properties seem to me to be key to a number of uses—comparisons between areas, ranking of areas, funding decisions, etc.—but there is little on how ensemble properties affect the choices in the framework. If the user requirement is for good ranking properties, does this change the small area model which is fitted, or does it merely affect the interpretation of the output? How do you convey ensemble properties to users? One strategy might be to maximize the accuracy of the worst comparison or element among the outputs, to have some guarantee that everything else is at least as good, but this could be very bad for some estimates. It would be helpful if the authors could suggest some strategies for using ensemble approaches in practice, and how these correspond to the specification phase.

**David Matz** (*Home Office, London*)
Many thanks and congratulations go to Tzavidis and his colleagues for an interesting paper. The following comments and reflections are very much from the perspective of a non-expert in small area estimation methods and are meant to help to suggest ways in which the authors might build on the work presented.

Firstly, it is worth noting that we already have a framework for official statistics in the UK: the 'Code practice for statistics' (`https://www.statisticsauthority.gov.uk/code-of-practice/`). So it would be interesting to consider what is in the proposed small area estimation framework which is not already in the code, and vice versa, as part of explaining why a further framework is needed.

Secondly, from a practical perspective, it would help to consider how the framework proposed could be refined and be implemented. Will this be via a guide for practising statisticians, or a programme of training or how will a framework be developed and used?

Thirdly, in relation to the comments in the paper about avoiding overly complex models, I was reminded of the quotation by George Box: 'All models are wrong but some are useful' (Box, 1976).

Fourthly, there were a couple of points not covered by the paper, which seemed relevant to the discussion and potentially for a framework. The first is the question of considering whether or not to provide data or estimates even when these may be poor quality and potentially misleading, or not well understood by users because of model complexity, or how to address such considerations. The second point is that consideration of statistical disclosure control should be included and may preclude provision of data in multiple geographies (as a comparison of data at a local level from overlapping small areas produced by using different geographies could inadvertently result in disclosure).

Finally I support and strongly agree with comments from other discussants, that a framework should cover not only the technical statistical methods and the associated considerations, but also emphasize engaging users and developing a common understanding of their wants and needs, to avoid errors of the third kind (i.e. the error committed by giving the right answer to the wrong problem (Kimball, 1957)), given diverse user needs and levels of statistical expertise.

**Bernard Baffour** (*Australian National University, Canberra*)
I first commend Tzavidis and his colleagues on a well-written and interesting paper which highlights the need for better discourse on small area estimation in official statistics. Although many policy and planning decisions require consistent and reliable information on subnational patterns and variations (e.g. at the provincial or local government level), in many statistical offices such statistics produced by using small area modelling are deemed 'experimental' and not 'official' statistics. The key reason for this is the general acceptability of such statistics, in my opinion, by policy makers, users and the general public. My question is

in regard to the application of Bayesian methods, which was briefly touched on in the concluding sections. Can the authors say more about how their framework extends to include Bayesian hierarchical modelling? This is particularly important in high dimensional situations with complex correlation structures, such as in age–sex–disease-specific mortality and morbidity small area estimates in public health.

**Peter F. Thall** (*MD Anderson Cancer Center, Houston*)

Tzavidis and his colleagues propose practical guidelines for statisticians to do small area estimation (SAE) in a way that is aimed at users, who may not be statisticians, but who may require such analyses to make policy decisions. The guidelines are motivated by the increasing complexity of users' requests, and by methodological developments in SAE. Major practical issues seem to be the granularity of a specified target level of geography, whether estimates over time are desired and what data actually are obtainable. The discussion of model criticism in Section 3.3 focuses on the use of transformations to obtain approximate normality when goodness-of-fit diagnostics indicate non-normal data. Semiparametric and non-parametric methods are suggested as alternatives, but no mention is made of generalized linear mixed models (Breslow and Clayton, 1993), which provide a large set of practical alternatives for non-normal data. This is surprising but perhaps reveals my ignorance of the practical limitations of SAE. In the *Encuesta Nacional de Ingresos y Gastos de los Hogares* data illustration, the $Q$–$Q$-plots and residual plots suggest a heavy-tailed distribution, or possibly a two- or three-component mixture, which may make sense from a sociological perspective. This might be handled generally by fitting a Bayesian non-parametric (BNP) model with a dependent Dirichlet process prior to accommodate regression structure, using empirical Bayes methods to estimate prior hyperparameters to mitigate criticisms of subjectivity. Inherently, this is a class of mixture models that can closely approximate essentially any distribution, can easily be configured to accommodate the predictive mixed model structure of equations (4)–(6) and does not require development of new theory to validate inferences. Estimation could be done using posterior means and credible intervals, and prediction of observable variables is quite natural in Bayesian inference. This approach also obviates the need for bootstrapping. Boonstra (2012) and Shi and Zhang (2013) seem to be moving in this direction. Although extending existing BNP software to accommodate key SAE structures certainly would be a substantial undertaking, this approach might serve to free analysts to follow Tukey's advice to 'focus on the questions, not models'. In any case, it would be very interesting to see how such a BNP approach compares with conventional SAE methods in a design-based simulation study.

**Danny Pfeffermann** (*Central Bureau of Statistics and Hebrew University of Jerusalem and University of Southampton*)

This paper sets up general guidelines for the production of small area estimates and measures of their precision 'from start to finish'. In many ways, this is a unique paper of its kind because papers on small area estimation (SAE) usually focus on a new methodology for a given goal.

A major part of the paper is devoted to model testing and diagnostics. Although there can be no doubt that model testing and diagnostics are important when relying on models for inference, in SAE the main objective is to produce reliable predictors, and the goodness of fit of the model is of secondary importance. Consequently, the main effort when evaluating a model or when comparing different models should be in terms of their prediction power rather than in terms of residual analysis. This is emphasized in the present paper, applying both model-based and design-based comparisons, although the main comparisons are between different transformations rather than between different models. Interestingly, the Box–Cox transformation seems to produce better predictors than the log-shift transformation for the mean and head count ratio in the one-sample model-based analysis, but the picture is reversed in the design-based simulation. Graphical displays, comparing the root-mean-square error RMSE of the various predictors area by area or for groups of areas, instead of just presenting summary statistics in tables, would be more beneficial for the comparisons, and would probably reveal further insight regarding the performance of the alternative predictors. The authors distinguish between sampled and non-sampled areas in the tables (as expected, the RMSEs in the tables are always larger for the non-sampled areas than for the sampled areas), and the same could be done in the proposed graphical displays.

I commend the authors for this pioneering paper, which I hope will generate other papers of this kind, possibly using the same data. Small area estimates are nowadays a major production of official statistics all over the world, requiring the development of a general system for all the steps involved, 'from start to finish', rather than the use of isolated procedures of statistical software, with different methodologists following different approaches, making internal and international comparisons difficult. For example, the familiar 'Programme for the international assessment of adult competencies' survey, which is carried out

in more than 40 countries, is supposed to provide many diverse estimates for small domains, but so far with no guidelines for how these estimates are to be produced. I believe that the present paper is a first step in this direction.

**Stefan Sperlich** (*University of Geneva*)

The paper by Tzavidis and his colleagues is a very useful combination of a review, a practical guide for users and a critical discussion of methods. There are essentially two aspects which I shall raise which in my opinion deserved more attention.

The first is the use of non-parametric and semiparametric methods. They are briefly mentioned as tools for relaxing distributional assumptions on the random terms, or to extend the linearity of a model. Then data transformation is proposed and intensively discussed. The main motivation is giving preference to parsimonious models. But my understanding of non-parametrics is in the spirit of model-free prediction and inference. The problem with model-based methods is that everything hinges on their correct specification. Even if predicting well, their evaluation (confidence sets and tests) easily fail. In some statistical offices, when urged for area predictions but lacking data for some areas, one borrows information of 'similar' areas and works with (weighted) averages of them. This is actually non-parametric kernel $k$-nearest-neighbour regression; see Lombardía and Sperlich (2008) and González-Manteiga *et al*. (2013). This may look like models too complex in practice, but it is what people intuitively would do; formalizing it in terms of non-parametric regression enables us to make inference, i.e. to evaluate statistically such an intuitive approach. In that sense, non-parametric methods are not more complex models, but even less a model. Complex is only the subsequent inference, but that is the price for being model independent.

The second aspect is the distinction between conditional and unconditional inference. This has been briefly discussed by the authors but without mentioning the negative consequences. My focus is here on subsequent inference in practice. It is clear and generally accepted that the use of mixed effects models leads to conditionally biased predictions (there are not many efforts to estimate those biases). This, however, has implications for their practical use, interpretation and political conclusions. When it comes to prediction intervals and comparative statistics like comparing areas or multiple testing, then the proper statistical way would be to study simultaneous confidence sets and uniform confidence intervals and tests. In contrast, all we offer at present are $1-\alpha$ interval estimates such that for any given data set we expect $100\alpha\%$ of our prediction interval not to contain the 'true' parameter. Recent research shows that these flaws can be resolved without too much practical effort; see Kramlinger *et al*. (2018) and Reluga *et al*. (2018).

**Enrico Fabrizi** (*Università Cattolica del Sacro Cuore, Piacenza*)

Whereas most of the academic literature focuses on methodological issues related to small area estimation, this paper has the merit of bringing the needs and evaluations of users to the forefront. The comments that follow are in this line, I hope that they will be helpful to Tzavidis and his colleagues.

Small area estimation is about the efficient combination of sample survey and auxiliary information, whose quality is essential. Specifically I would give more emphasis to the following points.

(a) Auxiliary variables should be measured consistently in both sources. This is not always so. In my experience with labour force surveys and censuses, different definitions, interview modes and overall quality have created large discrepancies in the measurement of the same variable (i.e. the unemployment rate).
(b) Linking survey and administrative archives can be non-trivial and prone to linkage errors, unless a unique, error-free, identifier is available.
(c) The frequency of updates of the auxiliary information is crucial. The percentage of employees in a household—used in the analysis of the Mexican data—can be very powerful but, if the known totals are not regularly updated, they can lead to biased estimation in non-census years.
(d) The predictive power of auxiliary information plays an important role in improvements in efficiency.

The authors choose not to discuss area level models in detail. I think that the choice between the two alternative approaches to modelling is not only methodological but should be part of the specification stage. Methods based on area level models have pros and cons from the user's perspective. They are less demanding in terms of information requirements, as only summaries of auxiliary variables are needed even for non-linear parameters; design-based properties, such as design consistency, are straightforward to achieve; these methods require less computational resources. There are drawbacks as well: the geography of targeted small areas drives the modelling process, so these methods are prone to the modifiable areal

unit problem type of problems. Moreover the model is often a purely predictive tool, as not specified at the microlevel.

I agree with the authors on the importance they give to ensemble properties. Users always look at a map and compare estimates across areas. Moreover, estimates should evolve reasonably over time. Another issue is multiple comparison: users often want to know which differences in the set of small area estimates are statistically significant. This can be a problem when the areas are many. It is methodologically tricky and has not received due attention in the literature so far.

**Kuldeep Kumar** (*Bond University, Gold Coast*)

I strongly welcome this paper which sets out a framework to produce small area official statistics. Tzavidis and his colleagues have proposed an iterative process on three broadly defined stages which are specification, analysis and adaptation, and evaluation. Keeping in view the fast developments in the area of machine learning I strongly suggest using these tools at the specification stage. Whereas machine learning tools have been successfully used in many areas I am a little surprised that the application of these tools in small area estimation is still at its infancy. Machine learning methods like classification and regression trees, random forests and stochastic gradient boosting ('TreeNet') have several advantages over ordinary least squares such as that they can handle outliers, missing values and model non-linear relationships and local effects which are common problems in small area estimation. These methods are quite efficient in selecting variables and modelling variable interactions. They are also particularly useful for unbalanced data sets. They can be used to select optimal predictors of the target population under the model and can yield parsimonious models.

**Alan H. Dorfman** (*Bethesda*)

Tzavidis and his colleagues are to be congratulated for laying out a framework for small area estimation that includes many elements often neglected or not treated as thoroughly as they should be in producing small area estimates. I disagree though about the completeness of their framework. I do not think they have the right starting point and, as a result, their suggestions for the evaluation of methods do not take in enough.

The authors are addressing the situation where small area estimation is not a mere by-product of the survey under consideration, but a concern from the outset, with typically, as they say, 'a non-negotiable target geography' set out or implied by law (as in the *Encuesta Nacional de Ingresos y Gastos los Hogares* example). In this situation it is proper for the small area analysts to be in on the *planning stage* of the survey, and not to wait till the data have been presented to them to begin considering options. The *small area planner* should have two goals in mind: to allow for greater efficiency of the small area estimates, as has been advocated by Singh *et al*. (1994) and several others since, and to allow for a means of evaluation of the methodology that goes beyond the merely plausible (Dorfman, 2018).

In Dorfman (2018) I argue for the need to have a built-in way to perform an *external* evaluation of the methodology, amounting to the ability to compare small area estimates with corresponding traditionally more trustworthy estimates, and to do so for the set of smaller areas that are typically neglected in the drive for efficient global estimation. In the *Encuesta* example, that would entail comparisons based on the target variable ictpc itself. The variable inglabpc, which the authors use as ictpc's surrogate in their evaluation, would assume its proper role as an auxiliary variable improving estimation efficiency.

The small area estimation world would benefit greatly, I expect, if the authors would extend their attention to addressing the planning stage of surveys intended for small area estimation.

**Stefano Falorsi** (*Italian National Statistical Institute, Rome*)

The paper by Tzavidis and his colleagues defines the flow of the statistical production process in the case of small area estimation (SAE). A standardized production flow is very important for national statistical institutes that must produce quality data for national and local communities on the basis of which policy choices and allocation of funds can be done. For this reason, the Italian National Statistical Institute worked in this field from 2010 to 2012 within the Sae Essnet (Eurostat, 2012: `https://ec.europa.eu/eurostat/cros/content/sae-finished_en`).

To facilitate the effective use of these methods under a standardized process flow (and to try to extend it to internal or external statistical users who are not specialists in SAE) the Italian National Statistical Institute developed an on-line system called the small area estimation tool SMART (`http://smarter:8080/Smart/`) for the production of estimates by using SAE methods. My presentation at the Organisation for Economic Co-operation and Development's workshop in 2015 (`http://www.webcosi.eu/web-cosi-news/2nd-workshop-perspectives-from-official-statistics-and-gove`

rnment-presentations-and-videos-now-available/) describes the characteristics of the second release, namely SMART2.

From a general point of view, SMART enables small area estimates for the Labour Force Survey and Health Multipurpose Survey to be obtained for every possible aggregation of municipalities defined interactively by the users. Furthermore, SMART enables users to provide a system of their own covariates defined at area level. For each variable of interest a different *default model* may be defined. The set of *default auxiliary unit level covariates* includes demographic variables and, if available, small area level means of the target variable coming from a previous census.

SMART2 is based on R code routines and introduces a new estimator based on the logistic linear mixed model other than several different SAE *area* and *unit* level methods based on linear mixed models (also exploiting spatial correlation between municipalities). Flexibility in the model specification is allowed; for example, for each set of target areas defined by the user, the system allows the user to set, for the target variable of interest, the default model specification or an alternative. In addition, a broad choice of diagnostic outputs is provided to select the best estimator in terms of bias and mean-squared error (SAE ESSnet, 2012: https://ec.europa.eu/eurostat/cros/content/sae-finished_en). For each small area estimator and for each small area the Web system provides the estimate and the corresponding mean-squared error and a set of diagnostics.

To address multivariate estimation problems related to SAE of census hypercubes, an extension of the SAE production process and SMART to deal with the census and social surveys integrated system is under study (Falorsi, 2017).

The following contributions were received in writing after the meeting.

**William R. Bell** (*US Census Bureau, Washington DC*) (© US Government work)
(Any opinions and conclusions expressed herein are mine and do not necessarily reflect the views of the US Census Bureau.)

The general framework of Tzavidis and his colleagues seems sufficiently flexible to cover many small area estimation (SAE) problems. For certain specifics I think that area level models, which are not covered in the paper, raise some different considerations.

First, I have concerns about the suggestion in Sections 3.2 and 3.3 that small area model building should include, as standard practice, deciding whether or not the model should include random effects. Although there are circumstances where using a purely fixed effects model makes sense, lacking such circumstances I recommend including random effects in the model as the standard practice. Removing random effects from an area level regression model, where they are the regression error terms, as in the model of Fay and Herriot (1979), has the unreasonable implication that, if the direct survey estimates $y_k$ were replaced by corresponding tabulations from a census of the entire population, then the model would fit the data perfectly. Bell (1999) illustrated problems with small area prediction results that can arise from models with no random effects. These occur if random effects are included but their variance is estimated to be 0, which is an event that generally has positive probability, and is of most concern when the number of areas modelled is not large. These problems have led to the development of strictly positive estimators for variance components, e.g. by Yoshimori and Lahiri (2014). A Bayesian approach also avoids these problems.

Second, although transformation of $y_k$ for area level models can be useful, this needs to be considered carefully since estimation of the sampling variances of transformed survey estimates presents some challenges. In particular, the usual linearization approach has a large sample justification that is not present in SAE.

Third, I suggest as an important tool for area level SAE modelling the use of generalized variance functions to improve direct survey variance estimates. Franco and Bell (2013) have provided an illustration. The general problem is that, if samples are small so SAE is pursued, the direct variance estimates based on the small samples will be unstable.

Finally, I would suggest as additions to the SAE software packages mentioned the WinBUGS (Lunn *et al.*, 2000) and JAGS (Plummer, 2010) programs. For those using a Bayesian approach, these programs provide sufficient flexibility to accommodate a wide range of unit and area level SAE models.

**Geir-Arne Fuglstad and Andrea Riebler** (*Norwegian University of Science and Technology, Trondheim*)
**and Jon Wakefield, Johnny Paige, Katie Wilson and Tracy Dong** (*University of Washington, Seattle*)
We congratulate Tzavidis and his colleagues on their comprehensive treatment of the production of small area estimates. The issue of data sparsity was of particular interest, and we read their suggestions for supplementing direct estimates with synthetic estimates by using data from broader regions and

model-based approaches taking advantage of known covariates with strong interest. However, the high reliance on the availability of relevant covariates seems limiting at finer spatial and temporal scales, and we would like to draw attention to the issues that arise for producing estimates at fine spatial and temporal resolutions.

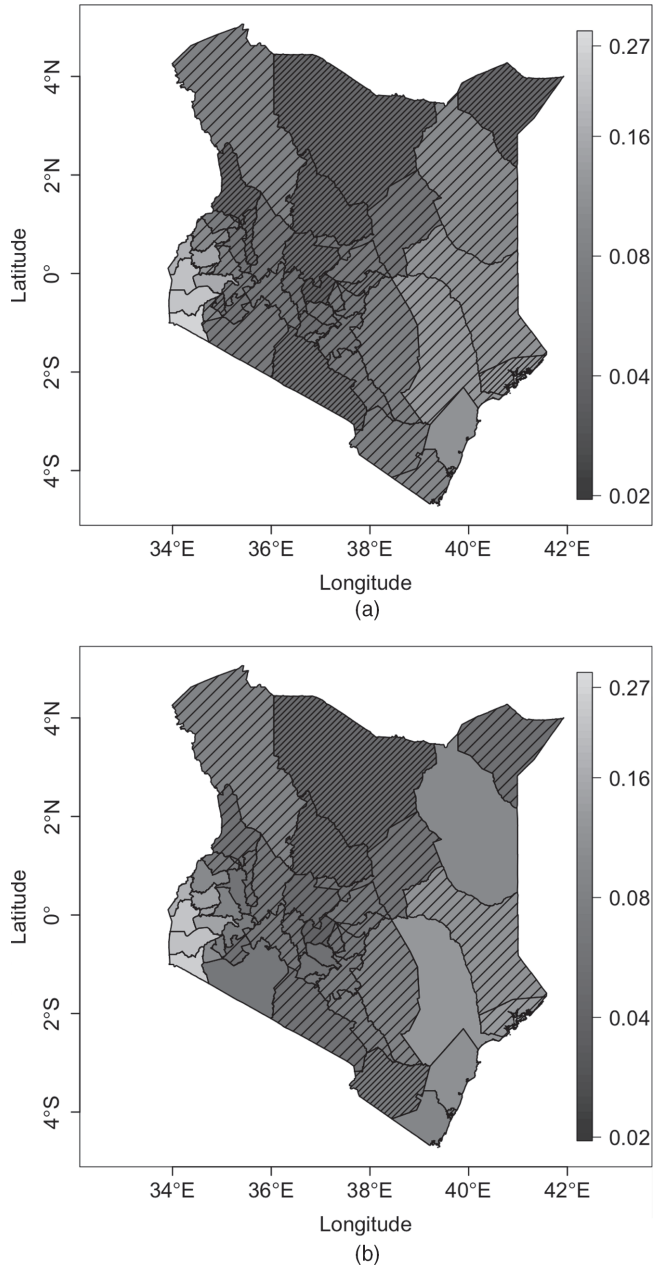



**Fig. 9.** Kenya U5MR-estimates in 2000 by using (a) a discrete spatial model and (b) a continuous spatial model (the shading indicates uncertainty showing the size of the standard deviation relative to the value of the estimate measured in per cent, i.e. standard deviation/median×100): □, [8%, 12%]; ▨, [12%, 16%]; ▨, [16%, 24%]; ▨, [24%, 35%]; ▨, >35%

We have been working on modelling the under 5 years mortality rate U5MR in a developing world context (Wakefield *et al.*, 2018), where important data sources are demographic and health survey programme surveys. For Kenya, the relevant spatial scales are admin 1, which consists of 47 counties, and admin 2, which consists of 262 divisions, and the relevant temporal scale is yearly. The hazard cannot be assumed to be independent of the age of the child. We used a discrete hazards model with hazards for each of the (monthly) age bands $[0, 1), [1, 12), [12, 24), [24, 36), [36, 48)$ and $[48, 60)$. Relevant covariates such as vaccination rates are not reliably observed at such fine temporal and spatial scales for each age group; modelled estimates are available, but the use of such values creates complications. In this setting, many age–time–space combinations do not have any deaths and reliable direct (weighted) estimates cannot be formed, and borrowing strength across space and time is essential.

If the spatial locations of the clusters are known, the locations can be used within a model-based approach with a 'pixel level' model. This endeavour is popular, with outcomes including malaria (Gething *et al.*, 2016), U5MR (Golding *et al.*, 2017) and vaccination (Utazi *et al.*, 2018). These analyses depend on household survey data, but the survey design is not explicitly acknowledged and we are currently investigating the implications of this (Paige *et al.*, 2018). We are also examining at which spatial resolutions reliable estimates can be produced, by aggregating contiguous pixels.

In Wakefield *et al.* (2018), stratification was explicitly recognized via a fixed effect for urban–rural, and we investigated the pros and cons of discrete and continuous spatial models. Point estimates of U5MR were very similar for discrete and continuous spatial models, but the uncertainty that is associated with the discrete spatial model estimates was larger, as shown in Fig. 9. A comparison of the discrete and continuous spatial models by using cross-validation showed a dramatic improvement of both model-based approaches over direct estimates at admin 1 level. However, the continuous model performed only slightly better than the discrete model.

**Seongho Kim** (*Wayne State University, Detroit*) **and Weng Kee Wong** (*University of California at Los Angeles*)

We congratulate Tzavidis and his colleagues for their detailed practical guidelines on small area estimation.

We focus our comments on data transformations. The Box–Cox transformation is a common data-driven transformation and equation (8) is a generalized version with a shift parameter '$c$' that allows for non-positive data. However, the range of the transformed outcome is restricted to a left-truncated domain with a bounded support not covering the entire range (Sakia, 1992). Thus, the choice of $c$ could influence parameter estimation. Two variants of the Box–Cox transformation that cover the entire range $(-\infty, \infty)$ are available. One that covers the entire range was proposed by Manly (1976). The other is Bickel and Doksum's (1991) modification that transforms the original data $y_{ik}$ to $|y_{ik} + c|^{\lambda}$ if $\lambda \neq 0$ and $\log(y_{ik} + c)$ if $\lambda = 0$ and the transformed response is multiplied by the sign of $y_{ik} + c$. However, none of them is reliable when there is a substantial proportion of 0s in the data. Such zero-inflated data sets may require different approaches other than data-driven transformations (Pfeffermann *et al.*, 2008; Chandra and Sud, 2012).

During implementation of the data-driven transformations to the State of Mexico, it was observed that the estimates of the Gini index were much more sensitive to the transformations than those for the mean and head count ratio, but no discussion was given for the underestimated Gini indices. We performed small simulation studies to investigate the effect of the transformations on the estimate of the Gini index by using two data sets. One concerns simulated annual incomes and the other is the wage data in the R package `ISLR` called Wage. Figs 10(a) and 10(b) plot the Gini index estimate and the skewness of the transformed data by the power transformation against the power parameter '$a$' in the form of data$^a$ for the simulated data and the wage data respectively. Figs 10(c) and 10(d) display the same plots but with the log-transformation $\log(\text{data} + c)$. For the power transformation, the estimate of the Gini index is proportional to the skewness of the distribution of the transformed data. It is interesting that this relationship is opposite to that for the log-transformation, and the estimates from the log-transformed data are always underestimated compared with that from the untransformed data. The implication is that extra caution is required to transform the data for estimating the Gini index.

**Ralf Münnich** (*Trier University*)

In the past 20 years, many methodological developments have been achieved in small area statistics. However, during this time, and especially in the past few years, it was pointed out that, in contrast with the many outstanding theoretical developments, fewer advances have been observed in applications. I congratulate Tzavidis and his colleagues for providing a very inspiring and outstanding contribution to

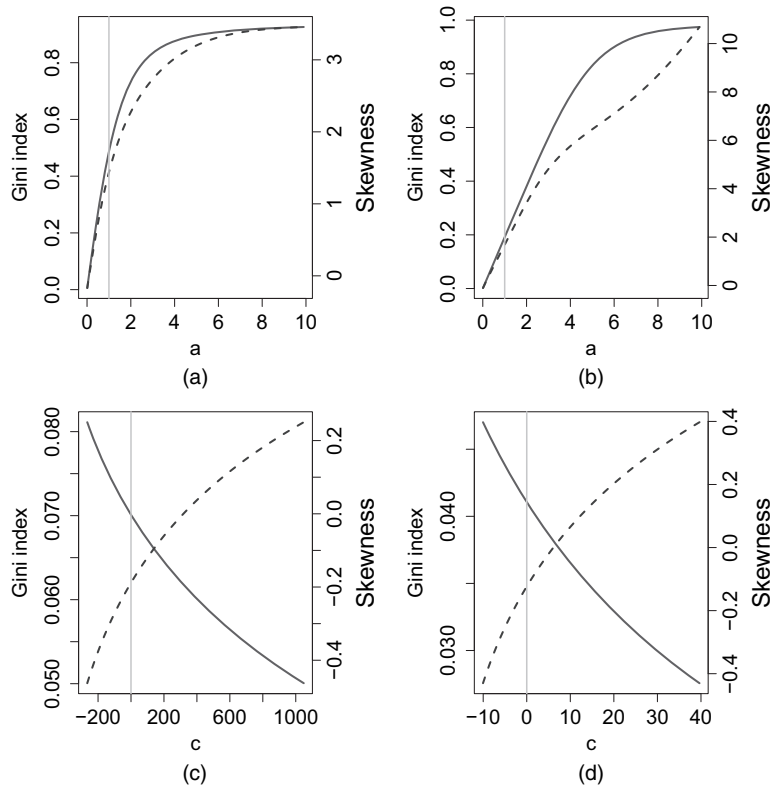



**Fig. 10.** Relationship between (a), (b) the power transformation (data$^a$) and the Gini index estimate/skewness and between (c), (d) the log-transformation (log(data + $c$)) and the Gini index estimate/skewness (the simulated income data, (531, 786, 1363, 2011, 2321, 2435, 3138, 4310, 5157, 5301, 6382, 8204, 10904, 15901, 21261), are used for (a) and (c), and (b) and (d) are based on the wage data for a group of 3000 male workers in the mid-Atlantic region available in the R package ISLR called Wage): |, cases when $a = 1$ or $c = 0$; ——, Gini index; – – –, skewness

small area statistics applications. The present paper provides a very valuable framework for setting up standards for future applications in small area statistics.

The use of small area statistics is a harbinger of a change of paradigm in official statistics from design-based methods towards model-based methods. Though many of the methods presented try to find a compromise between the two directions, model evaluation techniques will have to be integrated into the quality framework of the statistical production process. Further, instead of optimizing methods towards one single estimate, a vector of estimates must be produced, where optimizing towards single areas or all areas simultaneously leads to contradicting goals. Understanding these conflicts may be very important when using small area estimates as a basis for fund allocation.

Two aspects of data quality are also evident in small area statistics production. In official statistics practice, many estimates must be produced on very different levels of aggregation. The authors point out that it is likely that different methods must be recommended at the different levels of aggregation. However, this results in a very complex problem of coherence with increasing complexity by using different sources of data. Further, accuracy goals may have to be rewritten. Many European statistics are accompanied by recommendations regarding coefficients of variation that should be fulfilled; for example the regulations on European household surveys contain goals depending on the size of the areas, which must be adjusted accurately when we are dealing with very small areas. This discussion is currently taking place in Germany with respect to preparation of the legislation process for the German 2021 census.

Finally, small area statistics has already furnished common research between official statistics and universities. Providing better unit level or mixed data would help to advance open and reproducible research

to everyone's benefit and towards future needs. However, in the age of 'fake news' small area methods must be further integrated into statistical literacy and statistical education. An understanding of these methods and their outcomes becomes vital when small area estimates are introduced into political or economic decision processes.

**Marcin Szymkowiak** (*Poznan University of Economy and Business and Statistical Office in Poznan*)
The paper provides an excellent overview of the use of small area estimation (SAE) methodology in survey statistics and demonstrates the ability of Tzavidis and his colleagues to explain how to apply techniques of indirect estimation, which are still not widely used by national statistical institutes (NSIs). The content of the paper and problems raised by the authors are particularly close to me because of my work in the Center for Small Area Estimation (CSAE) in the Statistical Office in Poznan. My brief opinion will be mainly based on my experience gained in the CSAE and communication with people who are responsible for implementing new methods of estimation for official statistics.

From my perspective, the views expressed by the authors accurately reflect how NSIs should conduct the process of SAE. Moreover, the considerations presented in the paper should be treated as an indication of good practice. Unfortunately, owing to various restrictions, the implementation of SAE methods is somewhat limited in official statistics. This creates a barrier in communication between the world of SAE experts and NSI practitioners who must make decisions about whether or not to rely on new methods of estimation to deliver information for domains for which no official statistics exist. Hopefully, this paper is on the way towards eliminating these limitations.

Although many important topics, like non-response, outliers or the use of area level models, are not covered in detail, the paper provides guidelines on how NSIs should proceed step by step to apply SAE, and how to build the awareness of users and NSI employees in terms of the pros and cons of implementing SAE in practice. I also want to emphasize one important issue, which is briefly mentioned in the paper: based on my experience of working in the CSAE, I know how crucial it is to co-operate with subject matter experts. SAE statisticians can evaluate the process of estimation mainly in terms of model quality, its diagnostics and measures of accuracy and precision. It is unlikely that an expert in SAE is also an expert on poverty, the labour market etc. For this reason, co-operation with subject matter experts is important not only at the stage of selecting auxiliary variables for a model but also at the final stage before releasing results to the public.

In summary, this paper can be treated as a 'statistical road map' for all NSIs on how to conduct carefully the estimation and evaluation process using SAE in practice. It can also be hoped that this will help to increase the level of trust towards SAE methods on the part of end-users and may change the status of such estimates by putting them on a par with official statistics obtained using the design-based approach.

**Jon Wakefield, Zehang Richard Li, Yuan Hsiao, Bryan D. Martin and Jessica Godwin** (*University of Washington, Seattle*) **and Sam J. Clark** (*Ohio State University, Columbus*)
This is a timely paper given the increasing need for small area estimation (SAE) to guide policy and health interventions. Our own interest in SAE is in a low and medium income country context and we have been developing methods for estimating the longitudinal under 5 years mortality rate U5MR at a subnational level. Such estimation is often based on data from household surveys, such as demographic health surveys, that employ a stratified cluster design. The analysis of such data is crucial for evaluating whether health targets, such as those defined by the sustainable development goals, are reached.

Our work in this area (Chen *et al.*, 2014; Mercer *et al.*, 2014, 2015) has been based on space–time hierarchical modelling of the direct (weighted) estimates, as given in equation (2) of the paper. The model is a space–time smoothing version of that proposed by Fay and Herriot (1979). We describe this hybrid approach in the context of estimating the U5MR $p_{it}$ in (subnational) area $i$ and period $t$. Let $y_{it} = \log\{\hat{p}_{it}/(1-\hat{p}_{it})\}$ be the logit of the direct estimate of U5MR, with $V_{it}$ the associated (design-based) asymptotic variance estimate. We then take the first stage (likelihood) of the hierarchical model to be

$$y_{it}|\lambda_{it} \sim N(\lambda_{it}, V_{it}),$$

where we emphasize that $V_{it}$ is known. At the second stage, we smooth over space and time via

$$\lambda_{it} = \beta_0 + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it},$$

where $\beta_0$ is the intercept and the remaining five terms are random effects. We have terms in time, $\alpha_t$, and space, $\theta_i$, that represent non-smooth 'shocks' and are respectively independent and identically distributed.

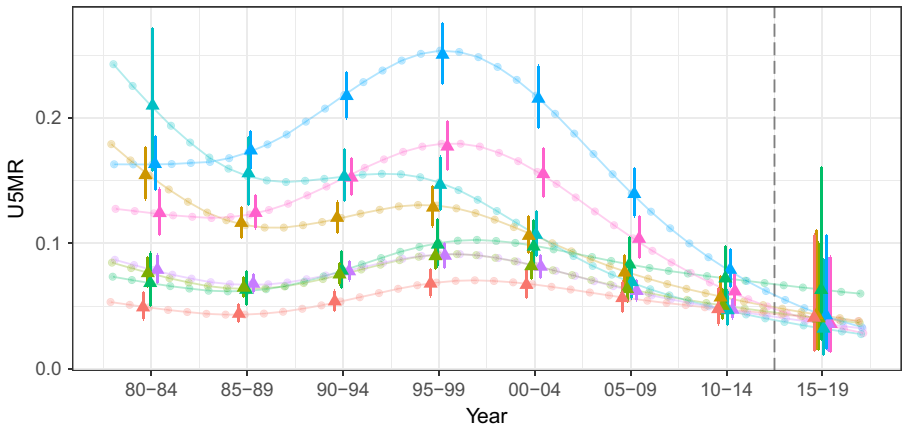



**Fig. 11.** Posterior median estimates for Kenya districts: ▲, central; ▲, eastern; ▲, north-eastern; ▲, Rift Valley; ▲, coast; ▲, Nairobi; ▲, Nyanza; ▲, western

We also have smooth terms in time, $\gamma_t$, and space, $\phi_i$, and space–time, $\delta_{it}$. This formulation allows great flexibility in the particular forms of smoothing that are chosen. We have chosen Markov random-field forms (Rue and Held, 2005), e.g. a random walk of order 2 model in time, an intrinsic conditional autoregressive model in space and a combination of the two (Knorr-Held, 2000) for the interaction. The model is implemented in the R package `SUMMER` (Martin *et al.*, 2018) with a design object being created in the `survey` package and the `INLA` package being used for Bayesian computation. It is computationally inexpensive, producing country-specific estimates in seconds.

Recently, we nave used this model to produce admin 1 estimates of U5MR for 35 countries in Africa (Li *et al.*, 2018), based on demographic health survey data. This work is supported by the United Nations Inter-Agency Group for Child Mortality Estimation. We have found that this approach produces robust estimates when the direct estimates are constructed at a spatial and temporal scale which is carefully chosen to give direct estimates that are reliable, in the sense that equation (9) is accurate. For this, in the African project we first calculate direct estimates in each subregion over 5-year time periods. The underlying temporal model is specified at the yearly level, however. An example of the results is given in Fig. 11, which displays posterior median yearly estimates for eight subregions of Kenya over 1980–2014 and predictions for 2015–2019. We also display median estimates (with 95% credible intervals) at the 5-year level; the uncertainty in the predictions is apparent. We also carry out extensive model checking and validation. We can validate against reliable direct estimates because of the choice of the spatial and temporal resolution of analysis. Full details are in Li *et al.* (2018).

**Katie Wilson and Jon Wakefield** (*University of Washington, Seattle*)

Small area estimation (SAE) is increasingly being based on multiple data sets. The data may be collected in a relatively small geographical region (a 'cluster'), for example, in a household survey. In this case, a Global Positioning System (GPS) cluster location may be reported, or this location may be jittered for confidentiality or the administrative region containing the cluster may be given. Alternatively, averages across administrative regions may be reported, for example, from a census. The different data sets may report over differing administrative units, which may not be nested and may have boundaries that change over time.

In Wilson and Wakefield (2018), we were concerned with combining survey data with GPS co-ordinates and census data associated with larger regions and no GPS co-ordinates. By assuming a latent continuous spatial field, we can model spatial dependence in the outcome, including data with differing geographical information. Suppose that we have data for area $i$ with reported average $\bar{Y}_i$, based on $N_i$ observations. To build the area model, consider the individual model $Y_{ij} \mid \mu_{ij}, \sigma^2 \sim N(\mu_{ij}, \sigma_i^2)$, where $\mu_{ij} = \beta_0 + S(\mathbf{s}_{ij})$ and $S(\mathbf{s}_{ij})$ is the spatial random effect (assumed to follow a Gaussian random field) at location $\mathbf{s}_{ij}$. The induced area level model is $\bar{Y}_i \mid \mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2/N_i)$, where

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \{\beta_0 + S(\mathbf{s}_{ij})\}.$$

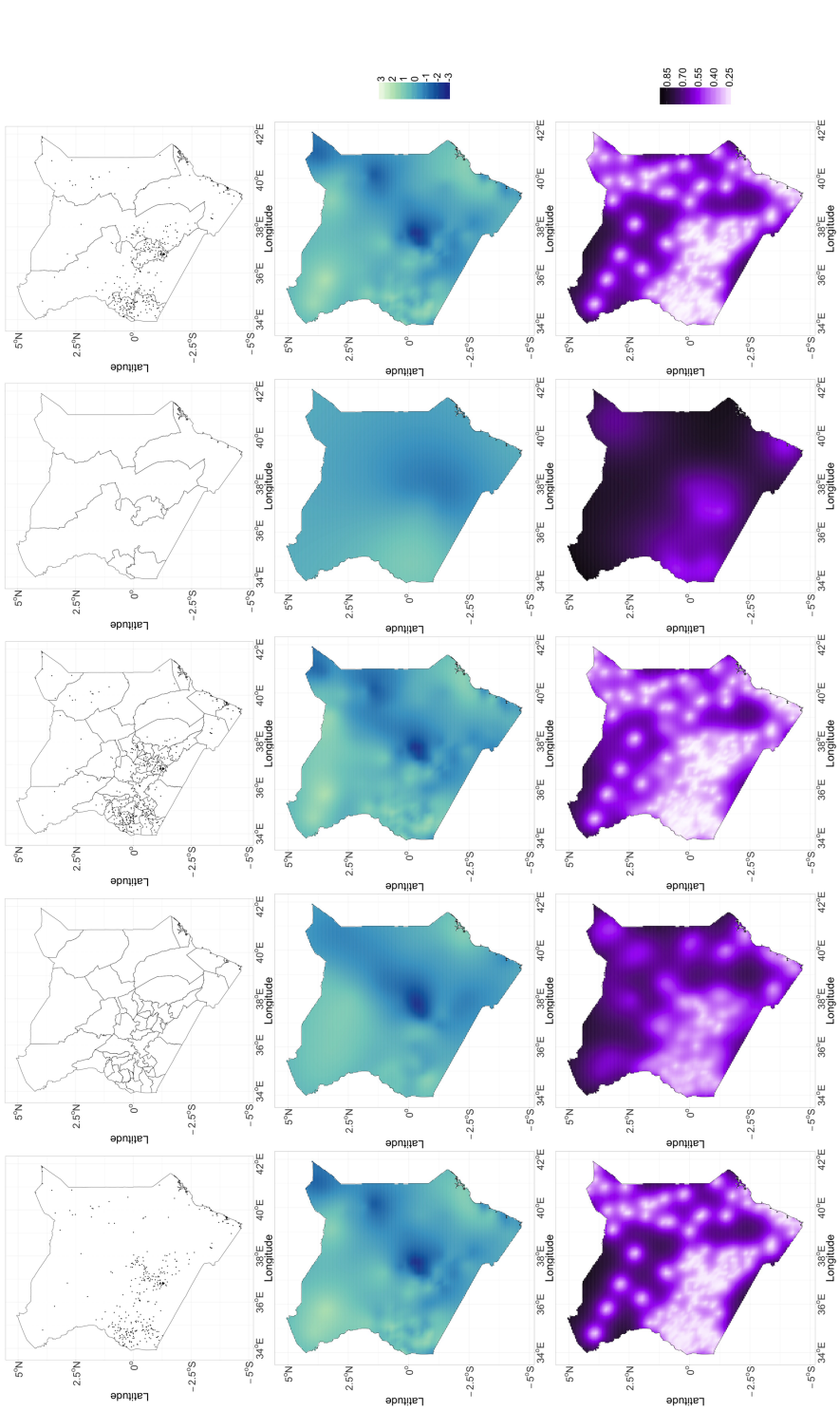



**Fig. 12.** Comparisons of latent spatial surface $S(\cdot)$ with normal data under the five scenarios (a) 400 surveys with exact location, (b) census data at the county level (47 areas), (c) both survey and county census data, (d) census data at the provincial level (eight areas) and (e) both survey and provincial census data: the top row is the data available (•, locations of the surveys; the borders correspond to the various boundaries); the middle row is the posterior mean of, and the bottom row the posterior standard deviation of, the predicted surface

A fine grid, indexed by $k$, can be used along with some measure of the population density $d(\mathbf{s})$ to give

$$\mu_i \approx \sum_{k=1}^{K_i} d(\mathbf{s}_{ik})\{\beta_0 + S(\mathbf{s}_{ik})\}.$$

To fit this model, we use the stochastic partial differential equation approach (Lindgren *et al.*, 2011) as implemented in `R-INLA` (Lindgren and Rue, 2015). For non-linear models, which would be appropriate for binomial or Poisson data, the model cannot be fitted within `R-INLA`, so we use Markov chain Monte Carlo sampling. For all likelihoods, the aggregate information can easily be combined with point level data, since they all depend on the same underlying spatial model. Fig. 12 gives an example. As expected, we find that prediction of the underlying spatial surface improves when point level survey data are combined with area level census data, compared with using one data source.

The necessity for developing models and tools for combining data sets (including modelling data-set-specific biases) remains a top priority.

**Thomas Zimmermann** (*German Federal Statistical Office, Wiesbaden*)

I warmly welcome this paper which offers guidelines on how to apply small area estimation (SAE) methods in the production process of official statistics. Compared with the rich literature on new advancements in the field of SAE methods, the number of their applications in official statistics seems to be quite modest so far. This paper clearly has the potential to stimulate further applications of SAE methods in official statistics.

I agree with Tzavidis and his colleagues that the process of developing estimators should be governed by the principle of parsimony. In my experience, this is also helpful for communicating the methods to subject matter experts, who might be reserved about the application of sophisticated estimation methods that may seem overly complex at a first glance. A second aspect which I view as crucial for the acceptance of SAE methods is the informal evaluation approaches that are briefly discussed in Section 4.2. In some cases, design-based methods produce domain estimates that are implausible. Examples in this regard are spurious fluctuations in the estimates from one year to another. If the model analyst can provide estimates which do not suffer from these shortcomings, it is much more likely that SAE methods will be accepted.

A related issue which I want to stress is the need to produce a coherent set of estimates. From my view satisfying the benchmarking property is a prerequisite for the application of SAE methods within official statistics rather than an 'attractive property'. Furthermore, an issue of great practical importance to national statistical institutes is how to account for the benchmarking adjustment in the uncertainty measure. To the best of my knowledge, this issue has received little attention so far.

Finally, I also like the discussion on choosing a transformation such that the assumptions underlying a particular model are fulfilled. However, I am not so sure that the mean-squared error is a very meaningful measure of uncertainty in this case. Owing to the presence of a non-linear back-transformation I would prefer to report lower and upper bounds of interval estimates instead.

The **authors** replied later, in writing, as follows.

We are very grateful to all those who submitted comments on the paper. Their comments pertain to all three stages (specification; analysis and adaptation; evaluation) of the proposed framework, raising some important issues which we could not cover in sufficient detail in the paper.

Discussants explored mainly three challenging issues that relate to the specification stage. The first concerns the overall strategy governing the sampling design, estimation and validation of model-based results. Ian Gordon, Paul Smith and Alan Dorfman all point out that a large amount of planning is needed when designing surveys, which could significantly impact on how small area statistics are produced and communicated. In particular, deciding key features such as the sample size, the clustering and the geographical spread of the sample are crucial as they can potentially reduce the reliance on overtly strong model assumptions, even though model-based methods are unavoidable because of the nature of the small area estimation (SAE) problem. Obviously, in reality it is difficult to specify the exact trade-off between estimates at, say, the national and local authority level, and much should be done in terms of validation of model-based estimates based on the data that one happens to have. Nevertheless, if national statistical institutes (NSIs) and other organizations view the production of small area statistics as part of their standard output, we would strongly encourage them to consider actions at the design stage. We acknowledge, however, that this is a complex problem that involves multiple, possibly competing priorities

and can create discontinuities resulting from design changes. However, one must avoid treating SAE as an *ad hoc* activity, not properly embedded in the processes of producing official statistics, or the perception that model-based methods can fix any remaining issues that have not been adequately accounted for at the design stage. Rather, as Danny Pfeffermann points out, it is required to establish a general system for all the steps involved, 'from start to finish'. Having an internationally accepted best practice guideline for achieving such a system would greatly contribute to the transparency and trust of the processes that are involved for both statisticians and users.

The second issue that the discussants were concerned with relates to the definition of target parameters. It is evident that we should not take it for granted that area-specific prediction is always the most relevant goal of estimation. A key reason why SAE has gained popularity as a research area during the last 20 years is precisely because there is a strong policy interest behind the production of statistics at detailed levels. Yet, in many cases we may fail to answer the questions that users are interested in and, in this respect, we agree with the comment made by Thomas King. Such questions relate for example to the best approaches to allocating funds, which by definition require good ranking and other ensemble properties of the set of small area estimators. The importance of ensemble properties is raised among others in the contributions by Paul Williamson and Stefan Sperlich. As we have mentioned in the paper, triple-goal and similar constrained estimation approaches have been developed in the literature. This is an area worth investing greater research effort into, to deal with the aspects of ensemble estimation that many users are more interested in. Stefan Sperlich points to simultaneous or uniform intervals as means of addressing the multiple-comparison problem arising from the use of ensemble estimates. Being associated with more stringent criteria than the simultaneous coverage that is discussed in our paper, these intervals pay a price by being wider than the area-specific intervals that one is accustomed to. Hence, in our view the success of various ensemble estimation methods will ultimately have to depend on the user's acceptance and appropriate usage.

The third issue concerns the data requirements and, more specifically, the pros and cons of unit level *versus* area level models. This matters greatly not least from the perspective of secondary analysis and the co-operation between the data owner and analysts. The use of area level models is a topic raised by several discussants, including Enrico Fabrizi, Ian Gordon, Marcin Szymkowiak and William Bell. A key advantage of area level models is that the data that are needed for estimation are more likely to satisfy confidentiality constraints to be made accessible to secondary analysts, sometimes even within the same NSI. Estimation of non-linear parameters with unit level models is less demanding conceptually, only when relevant census and administrative microdata are available. As mentioned in the paper, there are examples where area level models can be applied for estimating both linear and non-linear parameters (Fabrizi and Trivisano, 2016). It is thus easy to see why area level models can be a preferred option. In addition, area level models offer a more straightforward approach to combining data from multiple sources while taking the corresponding sources of variation into account. This is another feature of particular interest to NSIs. Our view is that the use of area level models will increase in future SAE applications. Nevertheless, as William Bell and Enrico Fabrizi point out, there are also non-trivial difficulties with the use of area level models that include the use of transformations and the estimation of the sampling variances as input to the model fitting procedure. Although for brevity area level models are not considered in detail in our paper, the proposed framework does cover aspects of the use of these models.

Turning now to the comments on the analysis and adaptation stage, these clearly capture the diversity of opinions on the topic. Here, we would like to respond on four broad topics.

The first topic concerns the role of sampling design (or selection mechanism in general) in model-based SAE. This is a challenge that methodologists must face in practice and often raise questions about. It has been a key topic in statistical analysis of complex survey data, on which there is a large literature that is relevant to SAE as well. The contribution by Maria Giovanna Ranalli offers some practical ways to account for the sampling design, which includes the use of design variables and the sampling weights in the small area model. A more ambitious approach can be found in Sverchkov and Pfeffermann (2018), who consider SAE under informative sampling and non-response. Meanwhile, as mentioned in our paper, another difficult problem is the estimation of finite population mean-squared error (MSE), in particular design MSE, which is of interest in official statistics. The SAE problem originates from the lack of design-based estimation methods that are both feasible and of acceptable precision. Underpinning this observation, albeit implicitly, is the adoption of design MSE as a relevant measure of uncertainty. The model-based estimators are often superior to direct design-based estimators in terms of design MSE. In other words, the acceptance of model-based estimators does not imply that design MSE cannot or should not remain a relevant measure of uncertainty. We expect many developments in this direction.

The second topic concerns extensions and modifications of the linear mixed model (LMM) and the even broader class of generalized linear (mixed) models (GLMMs). There is an extensive literature on the use of GLMMs in SAE (Ghosh *et al.*, 1998). Both LMMs and GLMMs have been extended to allow for covariance between the random effects where the correlation exists over time or space or both. Specifying an appropriate parametric correlation structure for the problem at hand can be challenging. Exploring semiparametric or non-parametric approaches that flexibly accommodate the temporal–spatial correlations in the data is an active area of research. The use of scaled transformations, which we have illustrated in our paper, is a form of model adaptation. Unlike model extensions, this approach allows us to apply previously worked-out procedures and fits well with the principle of parsimony. Seongho Kim, Weng Kee Wong and Ian Gordon comment that transformations may not apply uniformly for all target parameters, as we have noted in our analysis, despite the fact that a carefully chosen transformation can demonstrably improve the model fit. It is thus important to reiterate that transformations should not be blindly applied. Method evaluation (design-based simulations) and external validation of the estimates should remain at the centre of one's attention.

The third topic, which as we expected is touched on by many, is Bayesian methods; see for example the contributions by Bernard Baffour and William Bell. Empirical Bayes and hierarchical Bayesian (HB) methods have always had a central place in academic SAE literature. The James–Stein estimator is acknowledged as one of the key theoretical contributions to SAE. In the most recent SAE conference in Shanghai (June 16th–18th, 2018), two-keynote talks by James Berger and Malay Ghosh focused on new developments in Bayesian methods. A Bayesian version of the empirical best predictor method that is described in the paper was proposed by Molina *et al.* (2014). Current work by Browne, Tzavidis and Schmid, funded by the National Centre for Research Methods in the UK, also looks at Bayesian computation for the empirical best predictor method in Stat-JR. Here we would like to make two observations. First, an empirical Bayes estimator is just an empirical best prediction method from the frequentist perspective, and hierarchical modelling is not Bayesian in nature. HB methods are properly Bayesian because of the introduction of prior distributions and inference based on the posterior distribution. Bayesian methods offer computational advantages for fitting complex models and inference using the Bayesian posterior distribution, which is not always the case with frequentist procedures such as maximum likelihood estimation. However, given that in principle any frequentist method using an LMM or GLMM has an HB counterpart, Bayesian approaches face the same conceptual challenges as any other model-based approach. If anything, the additional steps of model specification of the priors that are required for the HB methods can place an even heavier burden on analysis. Second, as pointed out in our paper, there are many SAE applications where the target parameters are analytic in concept, such as in disease mapping and environmental and demographic studies. The traditional scepticism towards model-based methods for descriptive inference that are common in official statistics, under either the frequentist or HB approach, appears unjustified in such situations and, as pointed out by Peter Thall, Fuglstad and his colleagues, Wakefield and his colleagues and Wilson and Wakefield, Bayesian methods can have much to offer as alternatives to frequentist inference.

The fourth topic concerns the use of machine learning techniques, in particular the use of supervised learning algorithms. Kuldeep Kumar makes reference to machine learning tools such as random forests. As examples in the context of SAE we would also like to mention the research programme by Andy Tatem's group at the University of Southampton (`http://www.worldpop.org.uk`). In addition to the usual criticisms of model-based methods—also applicable in the case of machine learning methods—researchers and users are wary of the use of so-called black box methods. One can only hope to overcome this obstacle with additional training in these methods and better understanding of their underlying mathematical principles and statistical properties. In our opinion, the key issue with these methods is how to obtain valid estimates of uncertainty for the finite population parameters of interest. The accuracy of an official statistic cannot be formally assessed and the statistic published unless a valid estimate of uncertainty can be computed.

The comments in this last section relate to the evaluation stage of the framework proposed. To start with, we would like to reiterate the distinction between uncertainty assessment and method evaluation that is outlined in our paper, and the importance of design-based method evaluation, which provides an acceptable common ground for evaluating competing estimators regardless of their underlying inference outlook. Such evaluations have the merit of both simplicity and familiarity in communicating the performance of estimators; see for example Beaumont and Bocci (2016) for a recent study conducted at Statistics Canada. Another issue of importance is the role of benchmarking in SAE: a point also raised in the contribution by Thomas Zimmermann. Coherence with known or accepted benchmark totals is an indicator of the quality of SAE. When the benchmarks are obtained from external sources, it is possible to improve SAE

if these benchmarks are appropriately incorporated in the model. Nevertheless, as we have mentioned in our paper, there is currently an apparent contrast between the prevalent preference for design-based approaches to estimation at higher levels of aggregation and model-based approaches at lower levels. An approach worth researching is how to let benchmarking drive SAE. Starting from the accepted design-based estimates at an aggregated level, we can view SAE at the next lower level as a problem of how best to allocate the corresponding aggregated total by using model-based methods. Successively iterating the allocation would then yield a pyramid of estimates from the national total to the totals at the most detailed level. Building a pyramid like this would yield an analogy to the established process of deriving seasonally adjusted statistics, where model-based smoothing is obtained from the initial design-based point estimates, while offering a coherent account of the sets of point, trend and seasonally adjusted estimates.

Finally, discussants commented on the relationships between the framework proposed and several other relevant issues. David Matz questions the connection to the code of practice, which is the broad quality framework for the European statistical system. The three stages proposed in our framework support directly principle 7 'Sound methodology (A/A-stage)', principle 11 'Relevance (S-stage)' and principle 12 'Accuracy and reliability (E-stage)'. The principle of parsimony contributes to improving transparency, establishing widely acceptable production processes and best practice. In his contribution, Stefano Falorsi describes the SMART2 system developed at the Italian National Statistical Institute, which provides a nice example of how to enhance relevance by allowing user specification of SAE and facilitates direct communication of uncertainty. Paul Williamson points out rightly that SAE remains a relatively specialist and ghettoized activity. We would add to this that research areas that develop and use small-area-type methods are to a large extent also disconnected. Ralf Münnich stresses that SAE methods need to be more integrated with statistical literacy and training in survey and official statistics. Our current work under a grant (`https://www.ncrm.ac.uk/research/ISAEM/`) funded by the National Centre for Research Methods and the Economic and Social Research Council in the UK aims to enable statisticians and geographers to work together for linking different SAE methodological techniques. The deliverables from this grant, including the development of open-source statistical software (see for example Kreutzmann *et al.* (2018)), will form the basis of training courses in the future.

## References in the discussion


Beaumont, J.-F. and Bocci, C. (2016) Small area estimation in the Labour Force Survey. *Advisory Committee on Statistical Methods Meet., May 2nd*. Statistics Canada, Ottawa.

Bell, W. R. (1999) Accounting for uncertainty about variances in small area estimation. *Bull. Int. Statist. Inst.*, 52nd sess.

Bickel, P. J. and Doksum, K. A. (1981) An analysis of transformations revisited. *J. Am. Statist. Ass.*, **76**, 296–311.

Boonstra, H. J. (2012) hbsae: hierarchical Bayesian small area estimation. *R Package Version 10*. (Available from `https://CRAN.R-project.org/package=hbsae`.)

Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.

Breslow, N. and Clayton, D. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Carpenter, J. and Bithell, J. (2000) Bootstrap confidence intervals: when, which, what?: A practical guide for medical statisticians. *Statist. Med.*, **19**, 1141–1164.

Centre for Data Science and Public Policy (2016) Data maturity framework. University of Chicago, Chicago. (Available from `https://dsapp.uchicago.edu/wp-content/uploads/2018/05/Data_Maturity_Framework_4.28.16.pdf`.)

Chandra, H. and Sud, U. C. (2012) Small area estimation for zero-inflated data. *Communs Statist. Simuln Computn*, **41**, 632–643.

Chen, C., Wakefield, J. and Lumley, T. (2014) The use of sample weights in Bayesian hierarchical models for small area estimation. *Spatl SpatTemp. Epidem.*, **11**, 33–43.

Crawley, M. J. (2014) *Statistics: an Introduction using R*, 2nd edn. Chichester: Wiley.

Department for Communities and Local Government (2015) *The English Index of Multiple Deprivation (IMD) 2015—Guidance*. London. Department for Communities and Local Government.

Dorfman, A. H. (2018) Towards a routine external evaluation protocol for small area estimation. *Int. Statist. Rev.*, **86**, 259–274.

DuMouchel, W. H. and Duncan, G. J. (1983) Using sample survey weights in multiple regression analyses of stratified samples. *J. Am. Statist. Ass.*, **78**, 535–543.

Fabrizi, E. and Trivisano, C. (2016) Small area estimation of the Gini concentration coefficient. *Computnl Statist. Data Anal.*, **99**, 223–234.

Falorsi, S. (2017) Census and social surveys integrated system. *Working Paper 23*, item 8, 'Integration between census and social surveys'.

Fay, R. and Herriot, R. (1979) Estimates of income for small places: an application of James–Stein procedure to census data. *J. Am. Statist. Ass.*, **74**, 269–277.
Field, A. (2018) *Discovering Statistics using IBM SPSS Staistics*, 5th edn. London: Sage.
Franco, C. and Bell, W. R. (2013) Applying bivariate binomial/logit normal models to small area estimation. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 690–702.
Fuller, W. A. (2011) *Sampling Statistics*. Hoboken: Wiley.
Gething, P. W., Casey, D. C., Weiss, D. J., Bisanzio, D., Bhatt, S., Cameron, E., Battle, K. E., Dalrymple, U., Rozier, J., Rao, P. C., Kutz, M. J., Barber, R. M., Huynh, C., Shackelford, K. A., Coates, M. M., Nguven, G., Fraser, M. S., Kulikoff, R., Wang, H., Naghavi, M., Smith, D. L., Murray, C. J. L., Hay, S. I. and Lim, S. S. (2016) Mapping plasmodium falciparum mortality in Africa between 1990 and 2015. *New Engl. J. Med.*, **375**, 2435–2445.
Ghosh M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998) Generalized linear models for small-area estimation. *J. Am. Statist. Ass.*, **93**, 273–282.
Golding, N., Burstein, R., Longbottom, J., Browne, A., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D., Dwyer-Lindgren, L., Farag, T., Flaxman, A., Fraser, M., Gething, P., Gibson, H., Graetz, N., Krause, L., Kulikoff, X., Lim, S., Mappin, B., Morozoff, C., Reiner, R., Sligar, A., Smith, D., Wang, H., Weiss, D., Murray, C., Moyes, C. and Hay, S. (2017) Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *Lancet*, **390**, 2171–2182.
González-Manteiga, W., Lombardía, M. J., Martínez-Miranda, M. D. and Sperlich, S. (2013) Kernal smoothers and bootstrapping for semiparametric mixed effects models. *J. Multiv. Anal.*, **114**, 288–302.
Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS and Neighbourhood Targeting*. Chichester: Wiley.
Heady, P. and Ralphs, M. (2004) Some findings of the EURAREA project—and their implications for statistical policy. *Statist. Transn*, **6**, 641–653.
Judson, D. H. (2007) Information integration for constructing social statistics: history, theory and ideas towards a research programme. *J. R. Statist. Soc.* A, **170**, 483–501.
Kimball, A. W. (1957) Errors of the third kind in statistical consulting. *J. Am. Statist. Ass.*, **52**, 133–142.
Knorr-Held, L. (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Statist. Med.*, **19**, 2555–2567.
Kramlinger, P., Krivobokova, T. and Sperlich, S. (2018) Marginal and conditional multiple inference for small areas. *Discussion Paper*. Georg-August Universität Göttingen, Göttingen.
Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M. and Tzavidis, N. (2018) The R package emdi for estimating and mapping regionally disaggregated indicators. *J. Statist. Softwr.*, to be published.
Li, Z. R., Hsiao, Y., Godwin, J., Martin, B. D., Wakefield, J. and Clark, S. J. (2018) Changes in the spatial distribution of the under five mortality rate: small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. To be published.
Lindgren, F. and Rue, H. (2015) Bayesian spatial modelling with R-INLA. *J. Statist. Softwr.*, **63**, 1–25.
Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J. R. Statist. Soc.* B, **73**, 423–498.
Lombardía, M. J. and Sperlich, S. (2008) Semiparametric inference in generalized mixed effects models. *J. R. Statist. Soc.* B, **70**, 913–930.
Longford, N. T. (2005) *Missing Data and Small-area Estimation*. New York: Springer.
Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000) WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statist. Comput.*, **10**, 325–337.
Manly, B. F. J. (1976) Exponential data transformations. *Statistician*, **25**, 37–42
Martin, B. D., Li, Z. R., Hsiao, Y., Godwin, J., Wakefield, J. and Clark, S. J. (2018) SUMMER: spatio-temporal under-five mortality methods for estimation. *R Package Version 0.2.0*. Department of Statistics, University of Washington, Seattle.
Mercer, L., Wakefield, J., Chen, C. and Lumley, T. (2014) A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatl Statist.*, **8**, 69–85.
Mercer, L., Wakefield, J., Pantazis, A., Lutambi, A., Masanja, H. and Clark, S. (2015) Small area estimation of childhood mortality in the absence of vital registration. *Ann. Appl. Statist.*, **9**, 1889–1905.
Molina, I., Nandram, B. and Rao, J. N. K. (2014) Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. *Ann. Appl. Statist.*, **8**, 852–885.
Moon, G., Twigg, L. and Taylor, J. (2014) Estimating census health geographies: using synthetic estimation with secondary survey and census data. *Working Paper 1*. Estimating Census Health Geographies Research Project, University of Southampton, Southampton.
Mueller, P., Quintana, F., Jara, A. and Hanson, T. (2015) *Bayesian Nonparametric Data Analysis*. London: Springer.
Newcastle City Council (2011) Delivering cycling improvements in Newcastle—a ten year strategy 2011-22. Newcastle City Council, Newcastle upon Tyne. (Available from `https://www.newcastle.gov.uk/sites/default/files/wwwfileroot/planning-and-buildings/planning-policy/delivering_cycling_improvements_in_newcastle_-_a_ten_year_strategy.pdf`.)

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008) Non-parametric small area estimation using penalized spline regression. *J. R. Statist. Soc.* B, **70**, 265–286.
Paige, J., Fuglstad, G.-A., Riebler, A. and Wakefield, J. (2018) Model-based approaches to analysing spatial data from complex surveys. To be published.
Pfeffermann, D., Terryn, B. and Moura, F. A. S. (2008) Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Surv. Methodol.*, **34**, 235–249.
Plummer, M. (2010) JAGS—just another Gibbs sampler, JAGS 2.1.0. International Agency for Research on Cancer, Lyon. (Available from `https://sourceforge.net/projects/mcmc-jags`.)
Rao, J. N. K. and Molina, I. (2015) *Small Area Estimation*, 2nd edn. Hoboken: Wiley.
Reluga, K., Lombardía, M.-J. and Sperlich, S. (2018) Simultaneous prediction intervals for small area parameter. *Discussion Paper*. University of Geneva, Geneva.
Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman and Hall–CRC.
Sakia, R. M. (1992) The Box-Cox transformation technique—a review. *Statistician*, **41**, 169–178.
Schabenberger, O. and Gotway, C. A. (2005) *Statistical Methods for Spatial Data Analysis*. Boca Raton: Chapman and Hall–CRC.
Shi, C. and Zhang, P. (2013) BayesSAE: Bayesian analysis of small area estimation. *R Package Version 1.0-1*. (Available from `https://CRAN.R-project.org/package=BayesSAE`.)
Singh, M. P., Gambino, J. and Mantel, H. J. (1994) Issues and strategies for small area data. *Surv. Methodol.*, **20**, 3–22.
Sverchkov, M. and Pfeffermann, D. (2018) Small area estimation under informative sampling and not missing at random non-response. *J. R. Statist. Soc.* A, **181**, 981–1008.
Tsiptsis, K. and Chorianopoulos, A. (2010) *Data Mining Techniques in CRM: inside Customer Segmentation*. Chichester: Wiley.
Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Leesler, J. and Tatem, A. J. (2018) High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine*, **36**, 1583–1591.
Voas, D. and Williamson, P. (2001) The diversity of diversity: a critique of geodemographic classification. *Area*, **33**, 63–76.
Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K. and Clark, S. J. (2018) Estimating under-five mortality in space and time in a developing world context. *Statist. Meth. Med. Res*, to be published, doi 10.1177/0962280218767988.
Wang, J., Fuller, W. A. and Qu, Y. (2008) Small area estimation under a restriction. *Surv. Methodol.*, **34**, 29.
Williamson, P. (2007) Benchmarking the impact of cell adjustment on tabular outputs: the shortcomings of current approaches. *J. Off. Statist.*, **23**, 319–344.
Williamson, P. (2016) Small-area incomes: their spatial variability and the relative efficacy of proxy, geodemographic, imputed and model-based estimates. *Appl. Spatl Anal. Poly*, **9**, 463–489.
Wilson, K. and Wakefield J. (2018) Pointless continuous spatial surface reconstruction. *Biostatistics*, to be published.
Yoshimori, M. and Lahiri, P. (2014) A new adjusted maximum likelihood method for the Fay-Herriot small area model. *J. Multiv. Anal.*, **124**, 281–294.