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

Nikos Tzavidis[*], Li-Chun Zhang[*], Angela Luna[*], Timo Schmid[**], and Natalia Rojas-Perilla[**]

[*]Southampton Statistical Sciences Research Institute, University of Southampton, UK
[**]Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany

Abstract

Small area estimation is a research area in official and survey statistics of great practical relevance for national statistical institutes and related organisations. Despite rapid developments in methodology and software, researchers and users would benefit from having practical guidelines for the process of small area estimation. In this paper 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/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 light of the available data. Model-free and model-dependent methods are described with 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 using real data for the estimation of non-linear deprivation indicators. Linear statistics, for example 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 one. We shall follow the custom in this paper and use the terms area and domain interchangeably. In the last decades an increasing number of national statistical institutes (NSIs) and other organisations across the world have recognised the potential of producing small area (SA) statistics and their use for informing policy decisions. Some SA 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 (UALADs) by gender and age groups, and the estimates of average income for electoral wards. Other organisations and research groups have promoted the use of SAE techniques via the development of new methodologies and computational tools available for public use. An excellent example is the work by the World Bank (WB) and the use of its software PovMap (The World Bank, 2013). In collaboration with country teams, the WB has used SAE techniques for producing poverty maps in more than twenty developing countries. This is perhaps the most widespread application of SAE to date. Case studies can be found in The World Bank (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, for example measures of deprivation and inequality. Meeting the increasing complexity of users' needs requires specialised 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 adopted or developed are appropriate for the actual users' needs. In this paper we propose a framework based on three broadly defined stages, namely (i) specification, (ii) analysis/adaptation and (iii) evaluation, which are summarised in Figure 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. That is, 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/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, semi or non-parametric 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. While such a definition of parsimony is not exact, we believe 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 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 finalised 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 precision gains, 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 small area estimates is found, the process may need to return to the specification stage for defining alternative geographies, target indicators and/or data sources.

To keep a practical focus it is important to illustrate the application of the proposed framework using real data. The data we use in this paper come from Mexico. While being 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

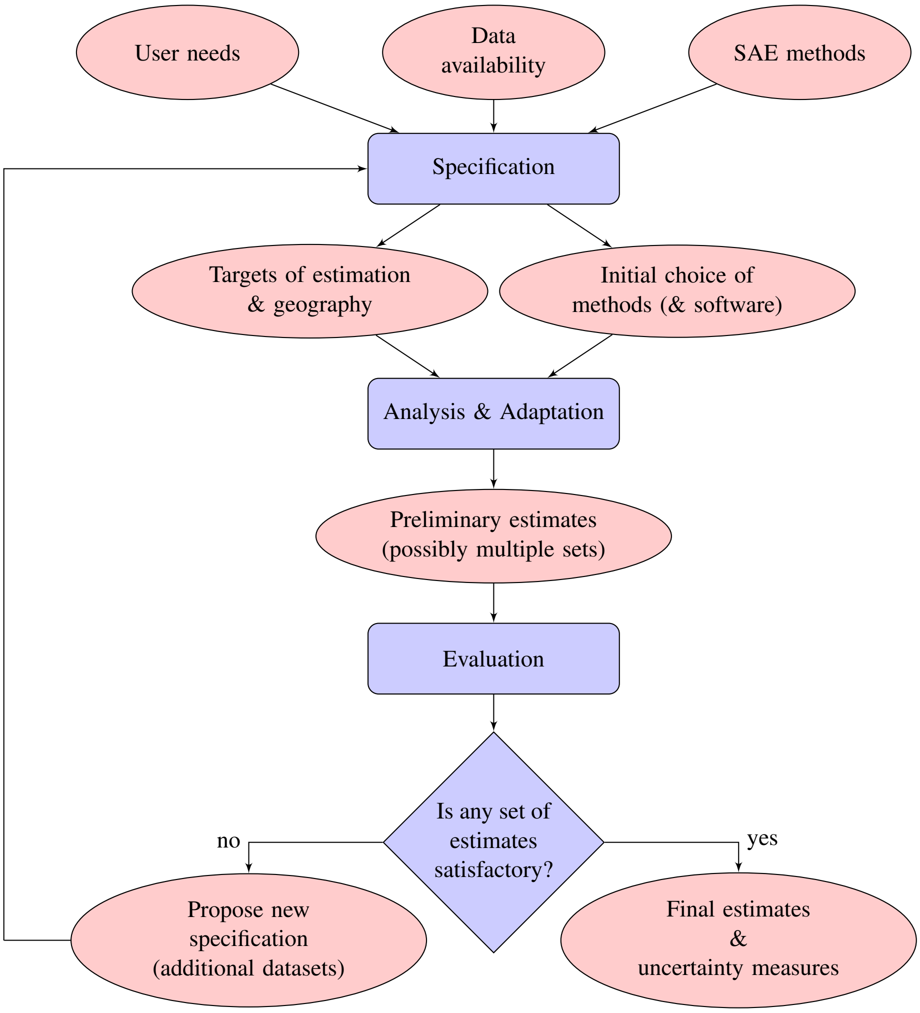
User needs
Data availability
SAE methods
Specification
Targets of estimation & geography
Initial choice of methods (& software)
Analysis & Adaptation
Preliminary estimates (possibly multiple sets)
Evaluation
Is any set of estimates satisfactory?
no
yes
Propose new specification (additional datasets)
Final estimates & uncertainty measures

Figure 1: Framework for the production of SA statistics: Stages of the project are represented by blocks. Inputs and outputs of each stage are represented by ellipses. Decisions to be made are represented by diamonds. Arrows indicate the direction of the relationship. Text in parenthesis indicates optional items

distribution of income deprivation and inequality. The National Council for the Evaluation of Social Development Policy (CONEVAL *Consejo Nacional de Evaluación de la Política de Desarrollo Social*) is responsible for estimating measures of poverty, social deprivation and inequality in Mexico. Furthermore, the general social development law (LGDS *Ley General de Desarrollo Social*) requires measures at the national and state levels to be obtained every two years and measures at the municipal level every five years. For the purposes of empirical analysis in this paper we use a sample from the household income and expenditure survey, ENIGH (*Encuesta Nacional de Ingreso y Gasto de los Hogares*) and a large sample of census micro-data. Both datasets are produced by the National Institute of Statistics and Geography (INEGI *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 based on continuous outcomes, recognising 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.

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 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 sub-national levels but usually have insufficient sizes to allow for 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 upon all the subsequent SAE process. It is very tempting for the user to target a geography that is unrealistically low. As we will see later, doing so will affect the methods and the assumptions 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, for example 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 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 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 needs and is well supported by the data 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 different choices and the extent to which the results will depend on finding a good enough 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 required and the models 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 minimise 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 in order to ensure that the user 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 organisations. Small area estimation is a prediction problem and typically relies on the use of survey data and data from the census or administrative/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 due to confidentiality constraints. Commonly, access to census aggregate (area/domain) level data is possible but access to census micro-data may not be possible. The question is how the type of census data available affects small area estimation. If the user is interested in estimating linear statistics, for example small area averages, access to area level census or administrative data will be sufficient for small area estimation. To illustrate this, suppose we have data on an outcome variable y_{ik} and a set of covariates $\boldsymbol{x}_{ik}$ 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 follows,

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

where n_k (N_k) denotes the sample (population) size in domain k and $\boldsymbol{x}_{ik}^T$ is the transpose of the vector $\boldsymbol{x}_{ik}$. The first summation in (1) is computed by using the survey data in domain k, assuming that sample data are available in the domain. The second summation in (1) represents the out-of-sample model predictions. It is easy to see that in order to compute (1), there is no need to have access to covariate micro-data. Instead, access to domain-level totals $\sum\limits_{i=1}^{N_k} \boldsymbol{x}_{ik}$ will be sufficient. If the interest is however in estimating non-linear indicators, then access to census or administrative micro-data is needed. Access to such data is very challenging and has implications for staff resources, in for example ensuring appropriate use of the data and respecting confidentiality constraints. Hence, the complexity of the targets of estimation determines the data requirements for small area estimation. Although the illustration of methods in this paper assumes the availability of census/administrative micro-data 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 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) develop 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) consider hierarchical Bayes approaches to fitting area level models for estimating non-linear indicators. Schmid et al. (2017) present a first attempt to use sources of big data, in particular mobile 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 micro-data for the covariates are not available or are available only for 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 realise that small area estimation 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 ENIGH data

In this case the targets of estimation and the required geography are specified by the LGDS (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 micro-data from the most recent census and survey data from the ENIGH. Hence, the estimation of the target indicators specified by the LGDS is feasible at least in principle.

Let us now look in more detail at the data available and their geographic coverage. Mexico is divided into 32 federal entities (states). The State of Mexico (EDOMEX *Estado de México*) 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 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 micro-data covers all EDOMEX municipalities. The survey and census data sources include a large number of socio-demographic variables, many of which are common and are measured in similar ways in both datasets. Total equivalised 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 impact of model assumptions and appropriately evaluates the methods and the estimates.

3 Analysis/Adaptation

The second stage in small area estimation 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 the Eurostat document ESSnet SAE (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 adaption stage but this is hardly ever the case. The analyst will most likely need to extend the triplet of estimates, by developing suitable models for small area estimation, 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 adaption 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 θ_k are the direct, synthetic and composite estimates. The direct estimator, denoted by $\hat{\theta}_k^{Direct}$, uses only the data from area k, so it is available only for an in-sample area. For areas with small sample sizes we expect that the direct estimator will have low precision. The synthetic estimator, denoted by $\hat{\theta}_k^{Synthetic}$, 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

$$\hat{\hat{\theta}}_k^{Direct} = \big(\sum_{i=1}^{n_k} y_{ik}/\pi_{ik}\big)/\big(\sum_{i=1}^{n_k} 1/\pi_{ik}\big), \tag{2}$$

where π_{ik} is the corresponding sample inclusion probability (Hajek, 1958; Brewer, 1963). A synthetic estimator of the mean $\hat{\hat{\theta}}_k^{Synthetic}$ is given similarly, based on the sub-sample from a broad area including area k, denoted by $\hat{\theta}_k^{Synthetic} = \hat{\theta}$, where $\hat{\theta}$ is a broad area estimate. The second situation is when auxiliary data are available, in which case the estimation is 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

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

where $w_{ik} = g_{ik}/\pi_{ik}$, and $g_{ik} = 1 + (X - \sum_k\sum_{i=1}^{n_k}\boldsymbol{x}_{ik}/\pi_{ik})^T(\sum_k\sum_{i=1}^{n_k}\boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^T/\pi_{ik})^{-1}\boldsymbol{x}_{ik}$, and X is the population total of $\boldsymbol{x}_{ik}$. A synthetic estimator $\hat{\theta}_k^{Synthetic} = \bar{\boldsymbol{x}}_k^T\hat{\boldsymbol{\beta}}$ is obtained by the linear model $E(y_{ik}|\boldsymbol{x}_{ik}) = \boldsymbol{x}_{ik}^T\boldsymbol{\beta}$, with $\hat{\boldsymbol{\beta}} = (\sum_k\sum_{i=1}^{n_k}\boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^T/\pi_{ik})^{-1}(\sum_k\sum_{i=1}^{n_k}\boldsymbol{x}_{ik}y_{ik}/\pi_{ik})$ and $\bar{\boldsymbol{x}}_k = N_k^{-1}\sum\limits_{i=1}^{N_k}\boldsymbol{x}_{ik}$. 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:

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

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 α_k for the composite estimator (3), including the James-Stein estimator that uses a common α 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) discuss different approaches for selecting α_k. One alternative approach is to define α_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 for example 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 too.

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 this will offer an indication of the improvement 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 already in use.

3.2 Use of models for small area estimation

Small area estimation 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 micro-data 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 only available in the survey, e.g. income or consumption. The explanatory variables, used for modelling the welfare variable, are available both in the survey and in the census datasets. After the model is fitted using the survey data, the estimated model parameters are combined with census micro-data 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, for example the HCR or the Gini coefficient. Linear statistics such as average income can also be estimated by using the same synthetically generated values.

Let us first describe the EBP approach, before we provide a brief discussion of the similarities 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 following unit-level nested error regression model,

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

where u_k denotes the domain random effect. A random effect is necessary when the covariates 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 following model,

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

where $\tilde{u}_k = E(u_k|y_s)$ is the conditional expectation of u_k given the sample data y_s. In (5), $\boldsymbol{x}_{ik}^T\boldsymbol{\beta} + \tilde{u}_k$ 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 (5) requires replacing the unknown quantities $\boldsymbol{\beta}, \sigma_u, \sigma_\epsilon$, with estimates and simulating L synthetic populations of the welfare outcome, $\boldsymbol{y}^*$. 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 small area estimation will be discussed in Section 4.1. For now we notice that evaluation of the uncertainty both for in-sample and out-of-sample domains is usually performed using parametric bootstrap under (4) and (5). Alternatively, protection against model misspecification can be offered by wild bootstrap. In this case 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 realisations of the outcome from the estimated conditional distribution (5) whereas the World Bank method simulates from the marginal distribution,

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

with all parameters replaced by their estimates. We now distinguish two cases. When clusters coincide with the target domains, Molina and Rao (2010) demonstrate 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 γ_k in (5) are both zero by default so that (5) reduces to (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 to 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 to the total variance of $y_{ik} - \boldsymbol{x}_{ik}^T\boldsymbol{\beta}$. 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 relevant cluster level covariates are included in the fixed part of the model (4). Moreover, the use of the conditional distribution (5) may be impossible in most of the clusters due to 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 two-fold 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 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 described in Section 3.1. Producing appropriate sets of initial estimates and their corresponding coefficients of variation (CV) would require access to data about the sampling design beyond our reach. The analysis below, obtained 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. Figure 2 (left) presents point estimates of average equivalised household income at the municipality level calculated from the ENIGH survey data using the final weights supplied. Figure 2 (right) 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 estimated under this assumption to 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 (ENIGH, 2010). However, even under this optimistic scenario it can be seen that, with the exception of few municipalities, the CVs are clearly above usual publication thresholds of $20\% - 25\%$. Notice also that direct estimates cannot be produced for the out-of-sample municipalities (white coloured areas). Hence, in order to satisfy the current user needs we should explore the use of model-based methods.

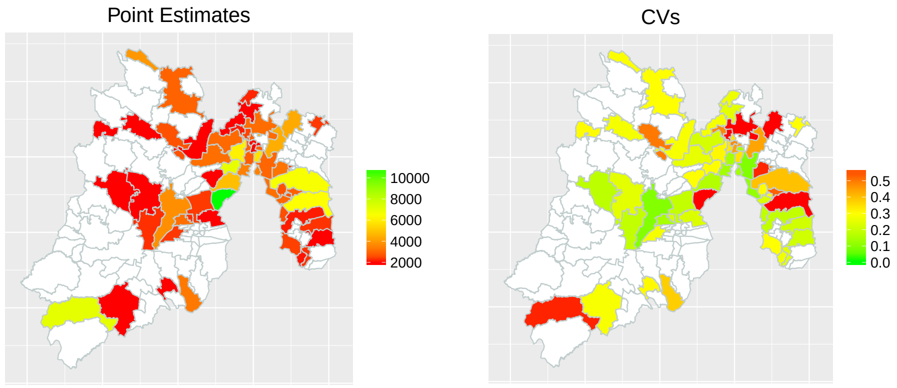

Figure 2: Direct estimates of average household equivalised income and CVs in EDOMEX municipalities

The use of models aims to improve the precision of small area estimates by making optimal use of the data available. Hence, model building, model diagnostics, sensitivity analysis and validation take central stage in model-based small area estimation. There is no single approach to model building. Here we describe some best practice guidelines 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. 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 for example, an exchangeable correlation structure in the simplest case. In order to decide whether to include a covariate in the fixed part of the model one can use simple t-statistics -computed using the correct variance under the model- or information criteria, for example the Akaike or the Bayesian Information Criteria (AIC, 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 to be 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 the AIC and a standard linear regression model the following covariates that are available both in the survey and census data have been identified as good predictors of *ictpc*:

1. Percentage of employees older than 14 years in the household;
2. Highest degree of education completed by the head of household;
3. Social class of the household;
4. Percentage of income earners and employees in the household;
5. Total number of communication assets in the household;
6. 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 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 latter 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 allows 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 vs. 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 $4.521 \cdot 10^{-10}$, 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 e.g. Snijders and Bosker, 2012). In the second step the GLS and standard linear regression models with the set of six covariates 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: 53077 vs. AIC for the standard linear model: 53079). The difference between these AIC values is very small, indicating that the covariates we included in the model explain a substantial part of the between municipalities variability. In particular, the intra cluster 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 proposed by Datta et al. (2011). The authors show 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 are 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 a number of 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 is 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, one 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 standardised 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 let us focus on the EBP method we described in Section 3.2 which relies on the normality of the residual terms. Figure 3 shows normal Q-Q plots of household-level and municipal-level residuals (random effects) obtain by fitting model (4) to income, using the six covariates we identified above 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 that expected for normal data.

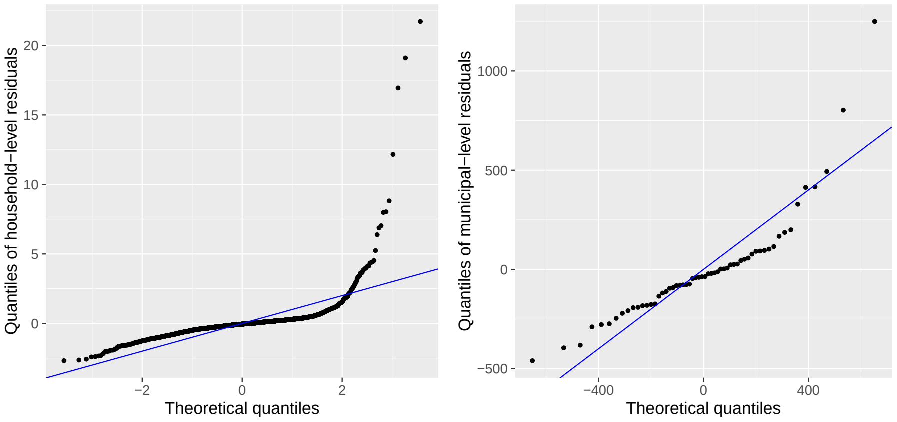

Figure 3: Normal Q-Q plots for household-level residuals (left) and municipal-level residuals (right) obtained from the model that uses raw income as the response variable

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 Generalised 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 semi-parametric approaches to model-based small area estimation have also been proposed (Weidenhammer et al., 2014). Use of semi-parametric 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 better handle real data challenges. 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 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 small area estimation. 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 some detail as an example of adapting the model. This is something we encourage prospective users to explore before deciding to use more complex models.

The papers by 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 on 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, denoted by λ, which must be estimated. The logarithmic transformation is then a special case of this family of transformations when $\lambda = 0$. Denoting by $\mathrm{T}_\lambda(y_{ik})$ the transformed outcome, the log-shift transformation is defined by

$$\mathrm{T}_\lambda(y_{ik}) = log(y_{ik} + \lambda). \tag{7}$$

The Box-Cox transformation is defined by

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

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 κ 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 κ, the Jacobian of the transformation is 1. Using the scaling by the geometric mean allows for 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 λ in (7) is to define a grid of values for λ, fit the nested error regression model by using each of the transformed outcomes $\mathrm{T}_\lambda(y_{ik})$ and select the transformation that makes 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) studies the use of different scaled transformations and estimation methods for λ in small area estimation. A general algorithm for implementing the EBP method with power transformations is as follows:

1. Define a parameter interval for λ;
2. Set λ to a value inside the interval;
3. Maximize the restricted log-likelihood function with respect to the vector of model parameters conditional on the fixed value of λ;
4. Repeat 3 and 4 until the value of λ that maximises the likelihood is found;
5. Apply the EBP method with the chosen value of λ.

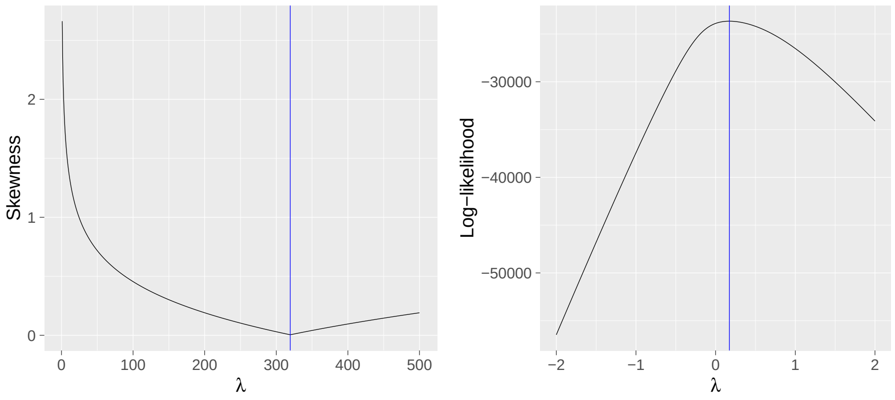

Figure 4: Shift parameter for the log-shift transformation (left) and optimal λ for the Box-Cox transformation (right)

Using the ENIGH data we apply the EBP method with three transformations for the outcome, namely log, log-shift and scaled Box-Cox. Figure 4 on the right 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 λ is approximately equal to 0.17. A similar graph on the left shows the shift parameter that minimises 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 identified above improve the diagnostic analysis and the predictive power of the model. We start with comments on the normal Q-Q plots (Figure 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 the other transformations.

In order to assess the assumption of homoscedasticity, we produce plots of the fitted values (x-axis) against the standardised residuals (y-axis) obtained by fitting model (4) using the raw income data (left) and the Box-Cox power transformation (right) in Figure 6. It can be observed that using transformations helps to stabilise the variance of the residuals. The corresponding plots for the log and the log-shift transformations are similar.

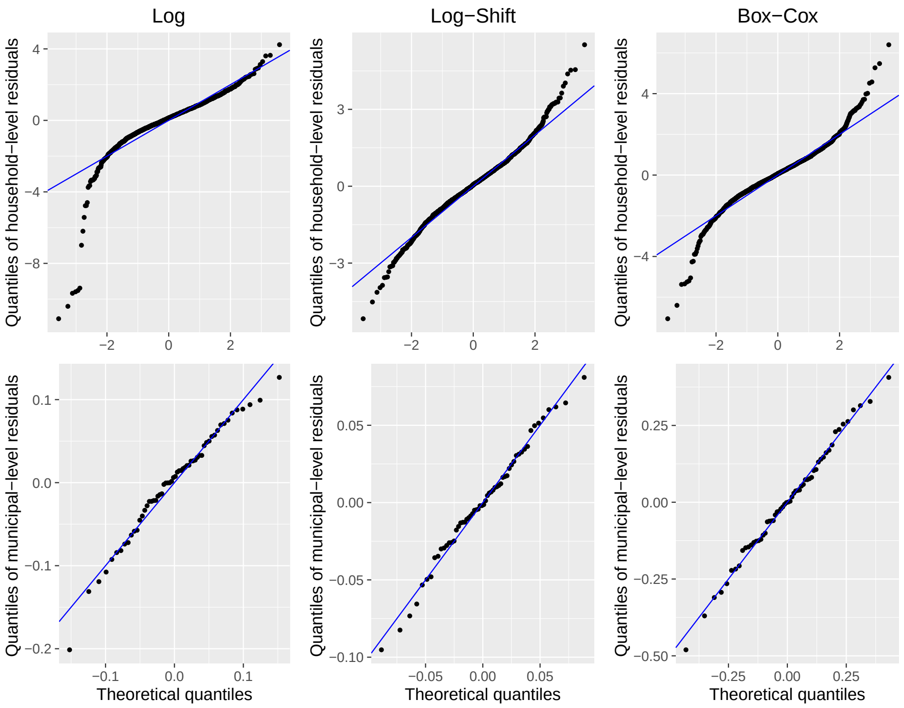

Figure 5: Normal Q-Q plots for household-level residuals and municipal-level residuals under three transformations for income

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

	Household-level residuals		Municipal-level residuals		
Transformation	Skewness	Kurtosis	Skewness	Kurtosis	R^2
Without	10.10	177.00	2.09	9.87	0.31
Log	-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

The proportion of variability explained under each model is quantified by the coefficients of determination R^2 summarised in Table 1. Note that as R^2 is computed based on 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 a R^2 equal to 31%. The use of transformations improves the predictive power of the model for the transformed variables.

Based on the results from the diagnostic analysis we conclude that two transformations, namely log-shift with shift parameter $\lambda = 319.52$ and Box-Cox 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 one. The following questions are raised at this stage. How important is the choice of transformation in small area estimation? 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 associated with the centre of the distribution and parameters associated with tails of the distribution? We attempt to address these questions in Section 4 that presents an evaluation framework for SAE. For now, we comment on Figure 7 that show maps of point estimates of average

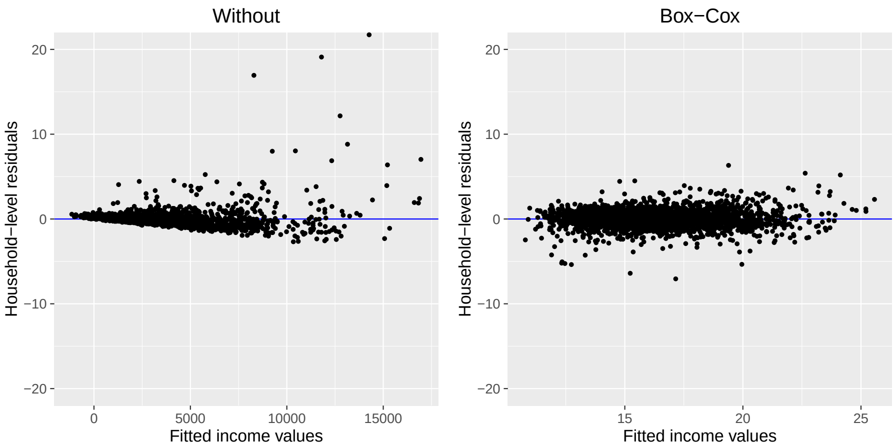

Figure 6: Standardized household-level residuals against fitted values without (left) and with Box-Cox transformation (right) for income

income, Gini coefficients and HCR for municipalities in EDOMEX produced by the EBP approach using different transformations.

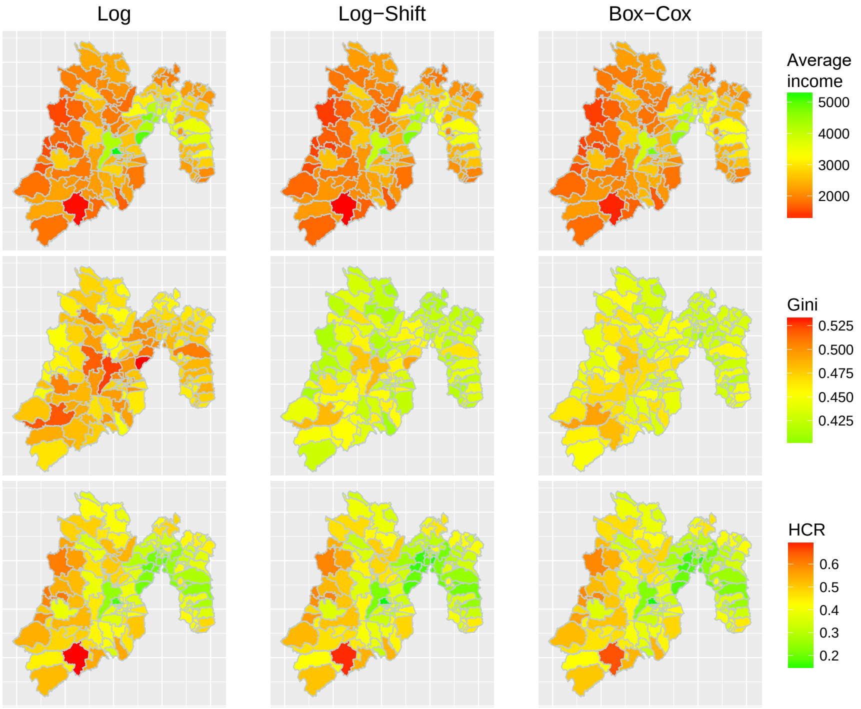

Figure 7: Map of municipal estimates of average income, Gini coefficients and HCR in EDOMEX using the EBP method under the log, log-shift and Box-Cox transformations

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 intra-regional 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 located in the central and south-west parts of EDOMEX. There are, however, some differences in the maps of point estimates 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. On the other hand, 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 an impact on point estimation especially when interest is in non-linear indicators that depend on the entire distribution. We will 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 using a design or a model-based MSE causes lively debates among researchers and practitioners. Comparing sets of optimal small area estimates produced under alternative models and deciding whether one set is better than another can be also a challenging task. Assessing ensemble properties of small area estimates such as the range or ranks of the estimates is 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 θ_k be the target parameter of area k, for $k = 1, ..., m$. Let $\boldsymbol{\theta} = \{\theta_1, ..., \theta_m\}$ be the collection of them. Let $\hat{\theta}_k$ be the estimator of θ_k and $\hat{\boldsymbol{\theta}}$ the collection of them. We assume that one is equally interested in all elements of $\boldsymbol{\theta}$ and cannot fix only on one particular θ_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 small area estimation one may distinguish between the area-specific and ensemble targets of $\boldsymbol{\theta}$. An ensemble characteristic of $\boldsymbol{\theta}$ is defined by using all θ_k's. For example, let $\bar{\theta}_w = \sum_{k=1}^m N_k\theta_k/N$ be the population mean, where N_k is the population size in area k and $N = \sum_{k=1}^m N_k$, or let $G = \sum_{k=1}^m (\theta_k - \bar{\theta})^2/(m-1)$ be the dispersion (i.e. population variance) of $\boldsymbol{\theta}$, where $\bar{\theta} = \sum_{k=1}^m \theta_k/m$. 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 e.g. Ghosh, 1992; 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 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 $k = 1, ..., m$. Let $\boldsymbol{y} = \{y_1, ..., y_m\}$ 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 $\boldsymbol{y}$. Prasad and Rao (1990) develop 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). Bootstrap (most commonly parametric) 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 bootstrap is particularly relevant for uncertainty estimation of indicators such as the Gini coefficient and the HCR. For example, for the EBP method 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 using the fitted marginal model (6). Compute the population value of the target parameter from each bootstrap population, denoted by θ_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 θ_k^* – as an estimate of the unconditional MSE. Notice that the procedure here is not second-order accurate, unlike the more sophisticated, but more computer intensive, bootstrap methods cited above. In case of using a transformation, the bootstrap populations are generated 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 λ 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 $j = 1, ..., m$. They argue particularly for its use under the generalised linear mixed models, and elaborate 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 one with respect to $\hat{\boldsymbol{\psi}}$ that varies only with the rest y_j's, for $j \neq k$, given y_k, where u_k and $\hat{\boldsymbol{\psi}}$ are conditionally independent (Booth and Hobert, 1998). Lohr and Rao (2009) propose a second-order accurate jackknife estimator of the conditional MSE. For a practical example, Zhang (2009) applies the CMSEP to estimates of small area compositions subjected to informative missing data.

The third type of MSE we describe is given by $E[(\hat{\theta}_k - \theta_k)^2|\boldsymbol{\theta}]$, where only the observed data $\boldsymbol{y}$ 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 $\boldsymbol{\theta}$ are now held fixed, and for this reason one may refer to this MSE as the finite-population (FP) MSE. There are several variations of the FP-MSE 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 $\boldsymbol{y}|\boldsymbol{\theta}$. The FP-MSE becomes the well-known design-based MSE, when $\boldsymbol{\theta}$ are population quantities such as the area means and $\boldsymbol{y}$ vary according to the sampling design (Rivest and Belmonte, 2000, e.g.). 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 one may not have access to the details required to implement the sampling design. Chambers et al. (2011) calculate the FP-MSE under the model for $\boldsymbol{y}|\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ are the theoretical area means rather than the population area means. Notice that these authors use the term "conditional" MSE, where it is the θ_k's that are treated as fixed not y_k as under the CMSEP. Finally, because the FP-MSE is a small area parameter itself, unbiased estimation is unstable whether it is with respect to the sampling design or model. Hence, one needs to treat the estimation of FP-MSE as a small area estimation 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" (Discussion of Nelder, 1977). There are times when the target parameter θ_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 mortality rate. But there are also many other situations, such as when θ_k is the area unemployment rate, where it is clearly defined as a descriptive statistic of the given population. One can still treat u_k as a random effect in order to achieve a sensible bias-variance trade-off, e.g. using model (4) to motivate a choice of α_k in the composite estimator (3). Without introducing the random effects model, one would have to resort to other means for deriving α_k. However, we believe that while it is inferentially consistent to report the model-based MSE here, which treats $\boldsymbol{\theta}$ as random, one is entitled to question its relevance when θ_k is a descriptive statistic and the assumption $E(u_k) = 0$ may be doubtful for a given k. In such a case, the FP-MSE is attractive for many survey practitioners. However, as explained above, the estimation of the FP-MSE 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 $C_k = (\hat{\theta}_{kL}, \hat{\theta}_{kU})$ be an interval estimator of θ_k, where $\hat{\theta}_{kL} < \hat{\theta}_{kU}$. The simplest procedure is to set the bounds such as $\hat{\theta}_k \pm 1.96 \cdot \hat{\text{MSE}}(\hat{\theta}_k)^{1/2}$, aimed at the 95% nominal confidence level. See Pfeffermann (2013, Section 6.2) for a review of interval estimation methods. Let $\delta_k = 1$ 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 $\boldsymbol{y}$ are allowed to vary. Similarly, one can speak about conditional coverage of C_k given by $E(\delta_k|y_k)$, and FP-coverage given by $E(\delta_k|\boldsymbol{\theta})$. Notice that any model-based C_k that treats θ_k as random can have rather erratic area-specific FP-coverage compared to the nominal level of confidence. Zhang (2007) defines $\varsigma = \sum_{k=1}^{m} E(\delta_k|\boldsymbol{\theta})/m$ to be the FP 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, ς converges to the nominal level, provided the underlying population model of $\boldsymbol{\theta}$ is correct.

4.2 Method evaluation

In the previous section we described different uncertainty measures. In addition to measuring the uncertainty associated with $\hat{\boldsymbol{\theta}}$ under the assumed model, an analyst may be interested in method evaluation. This might include comparing different point estimators, assessing how a 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 $\boldsymbol{y}$ from the sampling design. The evaluation can be performed analytically provided 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 hard to imagine that an NSI will produce any small area statistics on a regular basis without validating the design-based performance of the adopted method under realistic conditions. Typically, a census or similar population dataset 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 pseudo-population for the in-sample areas from the sample data at hand (e.g. Sverchkov and Pfeffermann, 2004). However, a model will be necessary in order to generate a pseudo-population 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 to 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 used for generating the population. A general question is whether it is meaningful to compare directly the MSE of an estimator $\hat{\theta}_{kA}$ of θ_k derived under model M_A to that of another estimator $\hat{\theta}_{kB}$ of θ_k under model M_B, which may involve different random effects or correlation structure. Notice that it is always possible to evaluate the MSE of $\hat{\theta}_{kA}$ under model M_B even though the estimator is motivated and computed under model M_A and vice versa. Since the MSE of $\hat{\theta}_{kA}$ will differ according to whether the evaluation model is M_A or M_B, there is a need to level the ground in order to avoid misleading comparisons. One may, for example, carry out simulation of both $\hat{\theta}_{kA}$ and $\hat{\theta}_{kB}$ under the model M_B if M_A is nested in M_B. When M_A and M_B are not nested in each other but are from the same class of models, one 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 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 to 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; 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 SAE evaluation using the ENIGH data

In this section we illustrate some of the aspects of SAE evaluation we discussed in Sections 4.1 and 4.2. In particular, using the results of model selection and diagnostics we described in Section 3.3, we present results for the estimation of average household equivalised income, 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 using the original data supplied to us by CONEVAL and estimated MSEs under the assumed model. 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 using four versions of the EBP method i.e. with untransformed income and three transformations (Log, 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 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 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 impact on point estimation. The impact 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 bootstrap estimator with the original sample. 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 impact on MSE estimation. Another question is whether the impact 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 using design-based simulation

In Section 4.3.1 above the MSE was calculated under the model estimated based on 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) suggest generating a pseudo-population by using the sample data. In some cases a variable that is highly

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

	Municipalities	58 In-sample			67 Out-of-sample		
	Indicator	Mean	HCR	Gini	Mean	HCR	Gini
Point Estimates	EBP	2730	0.380	0.949	2042	0.436	1.261
	EBP Log	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 Log	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 Log	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

correlated with the target variable is available in the census. This is the case 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 only available 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 as the target variable and *inglabpc* can play this role. A first reason as to 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. A 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 $inglabpc$ (used in the design-based simulation) and $ictpc$ (used in the one sample analysis). The distribution of both variables is similar and the total per-capita income $ictpc$ is generally higher compared to 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 micro-data and modelling of proxy household income *inglabpc*.

Table 3: Summary statistics over municipalities

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$inglabpc$ (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

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 equivalised household income from work, 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 identified in Section 3.3. The R^2 from linear regression models under different transformations (log, log-shift and Box-Cox) is around $40-50\%$ over the 500 samples, which is consistent with the results we obtained with the original sample.

The performance of these estimators is evaluated by computing the relative bias (RB) and root mean squared error (RMSE) given by

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

where $\hat{\theta}_k$ is generic notation to denote an estimator of the target parameter in municipality k, θ_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: Performance of predictors over municipalities in design-based simulations

	Municipalities	58 In-sample			67 Out-of-sample		
	Indicator	Mean	HCR	Gini	Mean	HCR	Gini
RMSE	EBP	180.2	0.095	0.497	210.6	0.073	0.846
	EBP Log	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 Log	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 Log	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 Log	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 Log	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 Log	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	—	—	—

Table 4 reports the results split by the 58 in-sample and the 67 out-of-sample municipalities. The table 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 small area estimation. 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 perform somewhat better compared to the EBP Log 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 we obtained from diagnostic analysis (Section 3.3). Comparing the EBP Box-Cox and the EBP log-shift 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 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 would like to 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 sub-area 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 only for 13 districts for which design-based estimates have a CV below 30%. Figure 8 shows point estimates for district-level average household equivalised income using the direct estimator (black line) and the EBP estimators with log (blue line), log-shift (orange line) and Box-Cox (red line) 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 (descending order from left to right). We observe that for districts where the direct estimates are more unreliable (left 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 (right part of the plot), the EBP Box-Cox and EBP Log-shift 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 Log estimates. We should emphasise that this is an informal approach to evaluating the quality of model-based estimates and there is no rule of thumb as to 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 8 districts, and the largest 8 districts, and 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.

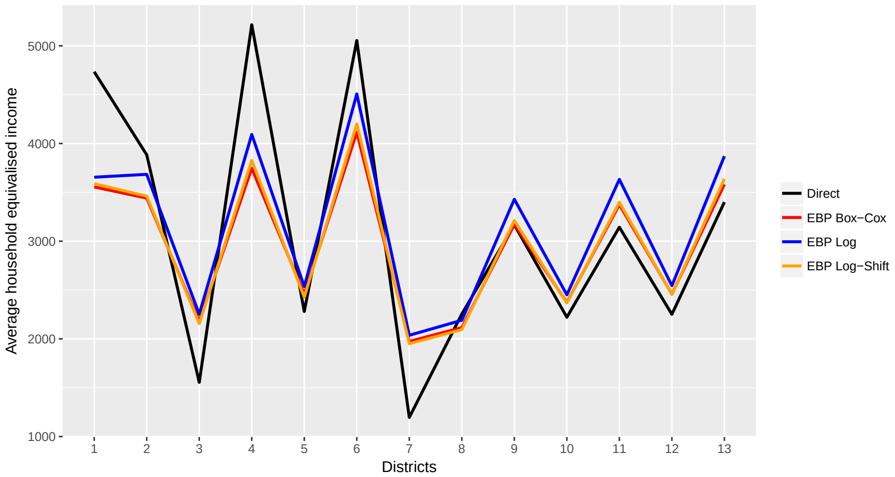

Figure 8: Estimates for average household equivalised income at district level.

5 An update on SAE software

In this section we provide a 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 are implemented in the open-source software R (R Core Team, 2015) via R packages.

A comprehensive review of relevant software is included in the CRAN 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, 2012) 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 different 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. A code in R for computing EBP estimates 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. (2017). 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 BIAS project (BIAS, 2005) which includes code for the unit-level EBLUP and spatial EBLUP with correlated random effects (Pratesi and Salvati, 2009). The SAMPLE project (SAMPLE, 2007) also provides a very wide range of code for implementing parametric, semi-parametric and outlier-robust small area estimation and allows for models with spatial and temporal correlations. We refer to Molina et al. (2010) for additional details. Small area estimation 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 (Meindl et al., 2016) and saeSim (Warnholz and Schmid, 2016) that support the prospective user in the setup 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 (The World Bank, 2013). PovMap implements the small area estimation procedure developed in Elbers et al. (2003) and is stand-alone software solution. The European funded project EURAREA (2001) delivered SAS codes for the computation of direct and indirect small area methods. For additional procedures in SAS we refer to Mukhopadhyay and McDowell (2011). Finally, all methods discussed in the paper are implemented by computationally efficient algorithms using R. The codes are available from the authors upon request.

6 Concluding remarks

In this paper we propose a general framework for the production of SA statistics and illustrate the SAE process in practice. As part of this framework we have touched upon three inter-related topics, namely specification of the problem, analysis of the data/ adaptation of the model, and method evaluation. While 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 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 data sources. Finally, obtaining uncertainty measures of good quality and designing method evaluation studies are of paramount importance for reassuring the users especially if interest is in using the estimates for official purposes, for example 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 SA statistics with discrete outcomes and use of area level models are not covered although the proposed framework 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 exists 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, as often it is, remain a side issue that one only pays attention to at the last stage of estimation?

Both area-specific and ensemble properties of a set of small area estimates are undoubtedly of interest. This is a distinctive feature of SA statistics in comparison to the national estimate that is a single number. Small area estimation is a simultaneous rather than a point estimation problem. Multi-purpose (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 impact in practice. Can users ever be ready or willing to accept multiple sets of estimates, each optimal for a particular purpose? How can one 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 used in empirical work. The views 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 only produced for illustrating the methods.

References

Alfons, A. and Templ, M. (2013) Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software*, **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. *Journal of the American Statistical Association*, **83**, 28–36.

BIAS (2005) Bayesian methods for combining multiple individual and aggregate data sources in observational studies. `http://www.bias-project.org.uk/`. Accessed: 11.04.2016.

Boonstra, H. J. (2012) *hbsae: Hierarchical Bayesian Small Area Estimation*. R package version 1.0.

Booth, J. and Hobert, J. (1998) Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, **93**, 262–272.

Box, G. E. and Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society Series 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. *Australian Journal of Statistics*, **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 *Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. Statistics Canada.

Ceriani, L. and Verme, P. (2012) The origins of the gini index: Extracts from variabilità e mutabilità (1912) by corrado gini. *The Journal of Economic Inequality*, **10**, 421–443.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014) Outlier robust small area estimation. *Journal of the Royal Statistical Society Series 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. *Survey Methodology*, **37**, 153–170.

Datta, G. and Lahiri, P. (1995) Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis*, **54**, 310–328.

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. *Journal of the American Statistical Association*, **106**, 362–374.

El-Horbaty, Y. (2015) *Model Checking Techniques for Small Area Estimation*. Ph.D. thesis, University of Southampton.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2003) Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355–364.

ENIGH (2010) Encuesta nacional de ingresos y gastos de los hogares 2010. enigh. diseño muestral. `http://www.beta.inegi.org.mx/app/biblioteca/ficha.html?upc=702825002420`. Accessed: 20.12.2017.

ESSnet SAE (2012) Small area estimation. `http://ec.europa.eu/eurostat/cros/content/sae-finished_en`. Accessed: 19.04.2016.

EURAREA (2001) Enhancing small area estimation techniques to meet european needs. `http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html`. Accessed: 11.04.2016.

Fabrizi, E., Salvati, N., Pratesi, M. and Tzavidis, N. (2014) Outlier robust model-assisted small area estimation. *Biometrical Journal*, **56**, 157–175.

Fabrizi, E. and Trivisano, C. (2016) Small area estimation of the gini concentration coefficient. *Computational Statistics & Data Analysis*, **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. *Journal of the American Statistical Association*, **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, Universitá di 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. *Journal of the Royal Statistical Society Series A*, **169**, 273–288.

Hajek, J. (1958) On the theory of ratio estimates. *Aplikace matematiky*, **3**, 384–398.

Hall, P. and Maiti, T. (2006) On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society Series B*, **68**, 221–238.

Jiang, J., Lahiri, P. and Wan, S. (2002) A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics*, **30**, 1782–1810.

Jiang, J. and Nguyen, T. (2012) Small area estimation via heteroscedastic nested-error regression. *Canadian Journal of Statistics*, **40**, 588–603.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Tzavidis, N. and Templ, M. (2017) *emdi: Estimating and Mapping Disaggregated Indicators*. R package version 1.1.1.

Lohr, S. and Rao, J. (2009) Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, **96**, 457–468.

Lumley, T. (2012) *survey: Analysis of Complex Survey Samples*. R package version 3.28-2.

Marhuenda, Y., Molina, I., Morales, D. and Rao, J. N. K. (2017) Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A*, **180**, 1111–1136.

Meindl, B., Templ, M., Alfons, A., Kowarik, A., and with contributions from Mathieu Ribatet (2016) *simPop: Simulation of Synthetic Populations for Survey Data Considering Auxiliary Information*. R package version 0.3.0.

Molina, I. and Marhuenda, Y. (2015) sae: An R package for small area estimation. *The R Journal*, **7**, 81–98.

Molina, I., Morales, D., Pratesi, M. and Tzavidis, N. (2010) Final small area estimation developments and simulations results. *Research Project Report Deliverable D12 and D16*, EU-FP7-SSH-2007-1 SAMPLE. URL: `http://www.sample-project.eu/`.

Molina, I. and Rao, J. N. K. (2010) Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, **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. *Journal of the Royal Statistical Society Series A*, **140**, 48–77.

Opsomer, J., Claeskens, G., Ranalli, M., Kauermann, G. and Breidt, F. (2008) Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society Series B*, **70**, 265–283.

Pfeffermann, D. (2013) New important developments in small area estimation. *Statistical Science*, **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. *Journal of Official Statistics*, **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 (2016) *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-126.

Prasad, N. G. N. and Rao, J. N. K. (1990) The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, **85**, 163–171.

Pratesi, M. and Salvati, N. (2009) Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics*, **25**, 37–53.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. N. K. and Molina, I. (2015) *Small Area Estimation*. New York: Wiley, 2nd edition edn.

Rivest, L.-P. and Belmonte, E. (2000) A conditional mean squared error of small area estimators. *Survey Methodology*, **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.

SAMPLE (2007) Small area methods for poverty and living condition estimates. `http://www.sample-project.eu/`. Accessed: 11.04.2016.

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. *Journal of the Royal Statistical Society: Series A*, **180**, 1163–1190.

Schmid, T., Tzavidis, N., Münnich, R. and Chambers, R. (2016) Outlier robust small area estimation under spatial correlation. *Scandinavian Journal of Statistics*, **43**, 806–826.

Shen, W. and Louis, T. (1998) Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society Series B*, **60**, 455–471.

Shi, C. and Zhang, P. (2013) *BayesSAE: Bayesian Analysis of Small Area Estimation*. R package version 1.0-1.

Sinha, S. K. and Rao, J. N. K. (2009) Robust small area estimation. *The Canadian Journal of Statistics*, **37**, 381–399.

Snijders, T. and Bosker, R. (2012) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publishers.

Sverchkov, M. and Pfeffermann, D. (2004) Prediction of finite population total based on the sample distribution. *Survey Methodology*, **30**, 79–92.

Templ, M. (2015) Cran task view: Official statistics and survey methodology. `https://cran.r-project.org/web/views/OfficialStatistics.html`. Accessed: 11.04.2016.

The World Bank (2007) *More than a pretty picture: using poverty maps to design better policies and interventions*. The international Bank for Reconstruction and Development - The World Bank.

— (2013) Software for poverty mapping. `http://go.worldbank.org/QG9L6V7P20`. Accessed: 11.04.2016.

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. *Australian and New Zealand Journal of Statistics*, **52**, 167–186.

Ugarte, M., Goicoa, T., Militino, A. and Durban, M. (2009) Spline smoothing in small area trend estimation and forecasting. *Computational Statistics and Data Analysis*, **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. *Austrian Journal of Statistics*, **45**, 55–69.

Weidenhammer, B., Tzavidis, N., Schmid, T. and Salvati, N. (2014) Domain prediction for counts using microsimulation via quantiles. In *Small Area Estimation 2014 Conference*. Poznan, Poland.

Zhang, L.-C. (2007) Finite population small area interval estimation. *Journal of Official Statistics*, **23**, 223–237.

— (2009) Estimates for small area compositions subjected to informative missing data. *Survey Methodology*, **35**, 191–201.