Freie Universität Berlin

From start to finish

A framework for the production of small area official statistics

Nikos Tzavidis
Li-Chun Zhang
Angela Luna Hernandez
Timo Schmid
Natalia Rojas-Perilla

School of Business & Economics
Discussion Paper

Economics

2016/13

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

Nikos Tzavidis*, Li-Chun Zhang*, Angela Luna Hernandez*, 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 that assist the process of small area estimation. In this paper we propose a general framework for the production of small area statistics that is based on three broadly defined stages namely, Specification, Analysis/Adaptation and Evaluation. The corner stone of the proposed framework is the principle of parsimony. Emphasis is given on the interaction between a user and a methodologist for 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 e.g. use of data transformations. The use of uncertainty measures and model and design-based simulations for method evaluation are also at the centre of the paper. We illustrate each stage of the process both theoretically and by using real data for estimating a simple and complex (non-linear) indicators.

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

1 Introduction

Small area estimation (SAE) has been and still predominately is a very fertile area of academic research in official statistics with important theoretical and applied contributions. In the last decade an increasing number of National Statistical Institutes (NSIs) and other organisations across the world have recognised the importance of producing small area (SA) statistics and their potential use for informing policy decisions. Examples of NSIs with well-developed programmes in the production of SA statistics include the US Bureau of Census, the UK Office for National Statistics (ONS) and the Statistical Office of Italy (ISTAT). After appropriate evaluation, SA estimates may gain accreditation as national official statistics as it is the case for the annual set of unemployment estimates for unitary authorities and local authority districts (UALADs), or the estimates of average income for electoral wards both produced by the ONS in the UK. Other organisations and research groups have promoted the use of small area estimation techniques via the development of new methodologies and computational tools available for public use. An excellent example of this approach 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 wide spread application of SAE to-date and country 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 statistics for example, 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 that extends beyond conventional survey operations within NSIs. This has created opportunities for closer collaboration between researchers and NSIs and for transferring research into practice hence, maximising the impact of research for society. At this point we should mention that the term researchers can include both researchers working outside an NSI and researchers working as part of an NSI's methodology unit. A common model used by many NSIs includes both academic researchers and methodologists collaborating for producing SA statistics of interest.

Despite the fast development of SAE methods and software researchers and users of SA statistics would benefit by a more open discussion on the practical aspects of SAE and by having a general framework that can guide the SAE process. The lack of such a framework is evident by the type of queries researchers receive from practitioners. Examples include decisions about specifying the target indicators, and disaggregation levels, the types of data required for estimation, the use of models and diagnostic analysis, model adaptations, measuring uncertainty and evaluation. There are many possible ways one can use to describe a general framework for producing SA statistics. In the present paper we propose a framework based on three broadly defined stages, namely (i) Specification, (ii) Analysis/Adaptation and (iii) Evaluation, which are summarized in Figure 1. Starting with the description of user needs, the available sources of data and existing SAE methods as inputs to the Specification stage, the user jointly with the methodologist defines a set of possible target geographies, indicators and suitable small area methods that can be supported by the available data. The outputs of the Specification stage act as inputs for the subsequent stage i.e. Analysis and Adaptation. It is our view that the decision about which estimators to use should be governed by the principle of parsimony and hence we recommend that the Analysis and Adaptation stage start by using estimators that can be easily computed as part of the usual survey process within an NSI without involving explicit modelling or additional data sources. These are the initial SAE estimates (see Section 3.1). In addition, alternative estimators that involve the explicit use of models may also be used (see Section 3.2), which involves model building and model diagnostics. Depending on the results from diagnostic analysis, adaptation of the models for example, data transformations or use of alternative parametric assumptions may be required. At the end of this process multiple sets of tentative point estimates will be available to the user. Evaluating these sets of estimates is the aim of the next step in the process. Evaluation involves both producing uncertainty estimates and method evaluation (see Sections 4.1 and 4.2). The SAE process is finalised provided that at least one set of estimates is considered of acceptable quality as assessed by uncertainty measures and method evaluation. Otherwise, the process needs to return again to the specification stage for defining alternative geographies, target indicators and/or data sources.

We believe that keeping a practical focus in this paper is of paramount importance hence each step of the process outlined in Figure 1 is illustrated by using real data. The data we use in this paper come from one of the most unequal regions in the world namely, Latin America and in particular Mexico. Despite being one of the largest economies in Latin America, according to the World Bank Mexico is still among the most unequal countries in the world. Developing policies against deprivation therefore requires

a detailed description of the spatial distribution of income deprivation and inequality. The National Council for the Evaluation of Social Development Policy (CONEVAL *Consejo Nacional de Evaluacion de la Politica de Desarrollo Social*) is responsible for estimating measures of poverty, social deprivation and inequality in Mexico. Furthermore, and according to the general social development law (LGDS *Ley General de Desarrollo Social*), it is necessary to provide measures at the national and state levels 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 a household income and expenditure survey called 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 Estadistica y Geografia*) and they were provided to the authors by CONEVAL. SAE usually includes a large number of different outputs for example, SAE with continuous and discrete outcomes and estimation of a wide range of linear and non-linear indicators. Hence, it is unavoidable that the present paper will have to focus on specific applications. In particular, we illustrate the SAE process for continuous outcomes and for estimating linear and non-linear indicators. Nevertheless, most of the steps of the SAE process we describe in this paper will also be valid also in applications with discrete outcomes.

The paper is structured as follows. Section 2 presents the specification stage. Section 3 discusses the analysis and adaptation stages. The starting point in this Section is the production of a set of initial estimates the computation of which can be done as part of standard survey processes within an organisation and without access to additional data sources and use of new methods and software. This Section then covers model-based small area estimation that includes model specification, diagnostics and adaptations. Section 4 describes another important stage in the SAE process namely, evaluation. Different evaluation measures that include unconditional and conditional mean squared error (MSE) estimators are discussed alongside a number of approaches for method evaluation. Section 5 provides an up-to-date review of open source software for SAE. Finally, in Section 6 we conclude the paper with some final remarks and 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

The first step in planning for small area estimation requires the specification of the user needs. This involves specifying the targets of estimation and the target level of geography. Before proceeding it is important to clarify that in this paper we refer to geography by using interchangeably either the term area or the term domain. The latter term is more general as it may also include non-geographic dimensions in its definition. Specifying the user needs is an important step that will impact upon the subsequent stages of small area estimation. It is the responsibility of the statistician/methodologist to guide the user through the specification stage and explain the consequences of the alternative options. 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.

It may seem as if the user has a clear idea about the target level of geography and the targets of estimation. However, this is not true in general. To start with, the complexity of the targets of estimation

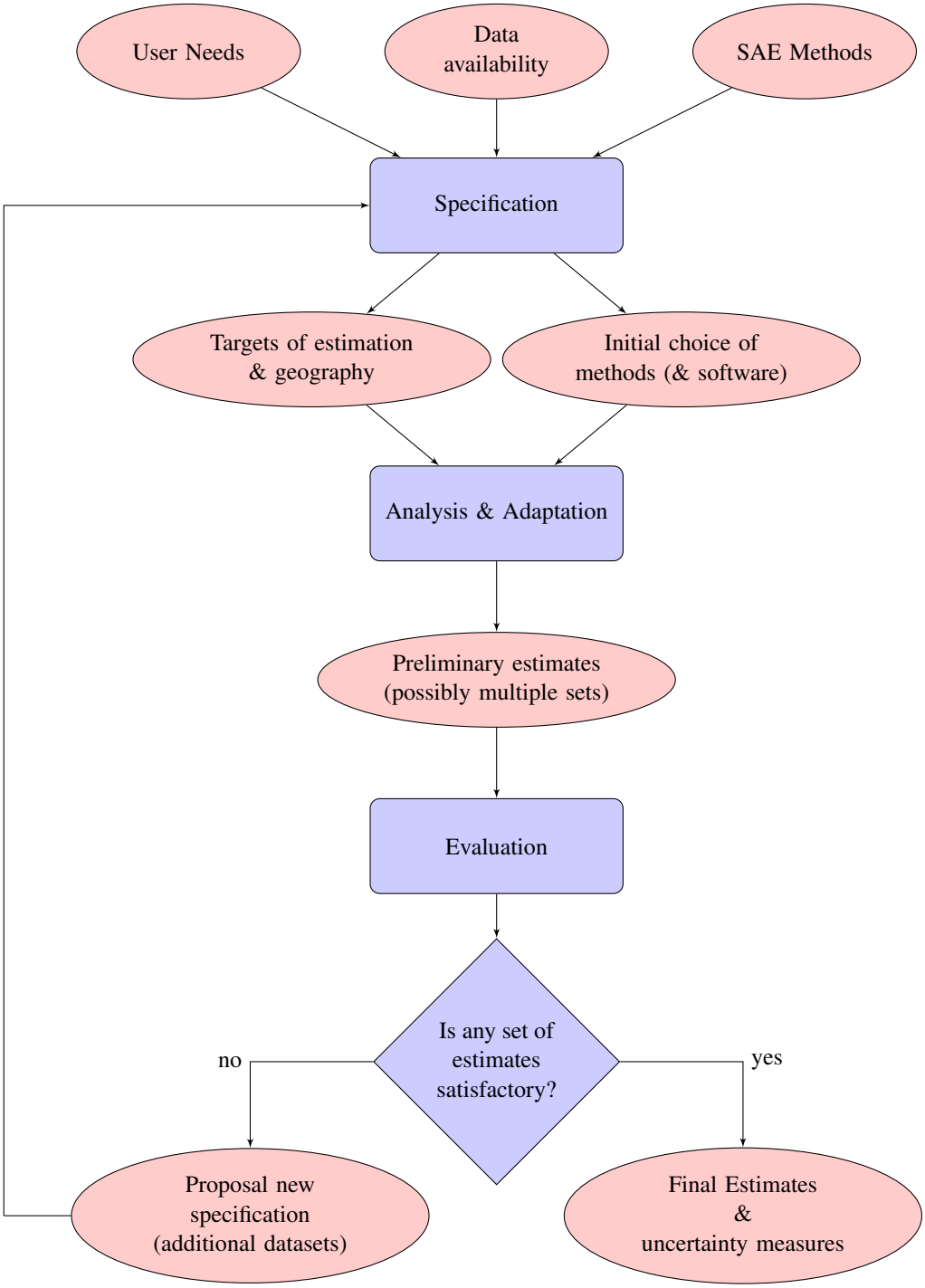

Figure 1: Framework for the production of SA statistics.
Stages of the project are represented by blocks. Inputs and outputs of each stage are represented by ellipses. Decisions to be made are represented by diamonds. Arrows indicate the direction of the relationship, i.e., an input/output. Text in parenthesis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

and whether the user is interested in cross-sectional estimates of estimates of changes over time will affect both the data required for small area estimation and the estimation methodologies that are available. Users may be interested in estimating simple linear indicators such as averages and proportions or more complex, non-linear indicators for example, the percentiles of the income distribution locally. Increasing the complexity of the target of estimation also increases the detail of the data one needs to have access to. On many occasions users have unrealistic expectations about the level of geography at which they wish to produce small area estimates for and what the available data can support. To start with, the chosen level of geography should provide useful estimates for the user. For example, estimates should be produced at a level which can inform policy. It is very tempting for the user to specify a target geography that is unrealistically low. As we will see in the next Section doing so will affect the quality (precision) of the estimates and the methods and assumptions used in computing the estimates. Hence, the recommended approach is to start from a relatively high level of geographical aggregation, at which direct estimation of acceptable precision is supported by the survey data, and move on to more disaggregated levels of geography by assessing the feasibility of producing small area estimates at each level in turn.

2.2 Data availability and geographical coverage

As we mentioned before specifying the targets of estimation is essential for identifying what data is needed for estimation. This is a first point that needs to be clearly communicated to the user since this will have implications for the workload of staff in 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 hence it can be used for assisting with the estimation. Access to Census and administrative data sources is usually challenging due to confidentiality constraints. Most commonly, access to Census aggregate (area/domain) level data is possible but access to Census micro-data is very challenging. 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,

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

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 covariates $\bar{\boldsymbol{x}}_k$ 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 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.

A second important step is to examine the data coverage at the specified level of geography. Are sample observations available for every small area? What is the distribution of the sample size across

areas? These are important questions. For example, if the majority 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. In this case a better strategy might be to consider producing small area estimates at the next, higher, meaningful level of geography.

2.3 Illustration using the data from Mexico

We now illustrate the specification stage in practice using the data from Mexico. In the case of Mexico the targets of estimation and the required geography are specified by the LGDS (see Section 1). Figure 2 depicts a two-dimensional definition of deprivation with one dimension defined by economic welfare and the second dimension defined by social deprivation. In Figure 2 MPL and EPL defines poverty lines for moderate and extreme income poverty respectively.

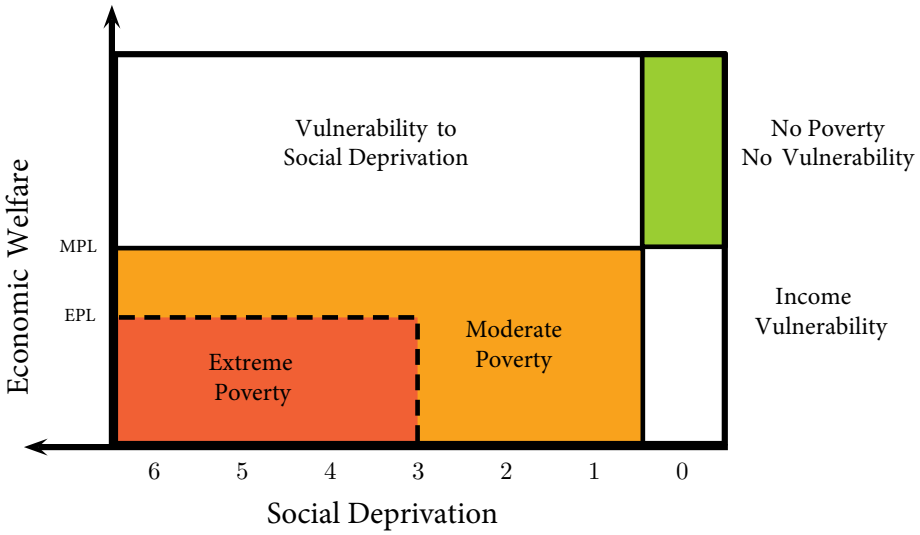

Figure 2: Two-dimensional definition of deprivation

Using Figure 2, Table 1 presents 18 indicators that the Government in Mexico is interested in estimating. We notice that the interest is both in producing estimates of proportions and totals but also in more complex, non-linear, indicators such as estimates of the Gini coefficient and the income ratio. It is therefore clear that SAE requires the use of Census micro-data. Are appropriate data available? Methodologists working 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 described in Table 1 is feasible from the point of view of data availability.

Let us now look in more detail at the data available and their geographic coverage. Mexico is divided into 32 federal entities (states), of which the State of Mexico (EDOMEX *Estado de Mexico*), in addition to having the highest population density, is regarded by the United Nations Development Programme (UNDP) to be 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 socioeconomic regions districts. This geographic information is presented on the left in Figure 3. The pilot data we have available in this paper were provided by CONEVAL and come from the 2010 ENIGH survey and the 2010 Census in EDOMEX. In particular, ENIGH survey data comprise 2748 households in 58 out of 125 municipalities. Census data come from a big sample of Census micro-data that covers all EDOMEX municipalities. The survey and Census data sources include a large number socio-demographic variables many of which are common and are measured in similar ways in both

Table 1: Targets of estimation in municipalities in Mexico

Category	Measurement
Poverty	
1.	Population in poverty
2.	Population in moderate poverty
3.	Population in extreme poverty
4.	Vulnerable population by social deprivation
5.	Vulnerable population by income
6.	Non-poor, non-vulnerable population
Social deprivation	
7.	Population with at least one social deprivation
8.	Population with at least three social deprivations
9.	Deprivation due to educational gap
10.	Deprivation due to the lack of access to health services
11.	Deprivation due to the lack of access to social security
12.	Deprivation due to the lack of quality housing services
13.	Deprivation due to the lack of access to basic housing services
14.	Deprivation due to the lack of access to food
Well-being	
15.	Population with income less than the moderate poverty line
16.	Population with income less than the extreme poverty line
Inequality	
17.	Gini coefficient
18.	Income ratio

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 (using only the ENIGH data in a municipality) for these municipalities impossible. The challenge is that in this case the target geography is determined by LGDS. As we will see later in this paper, having too many areas that are out-of-sample makes the small area estimates heavily dependent on synthetic estimation, which may introduce over-smoothing and hence bias. It is the responsibility of the methodologist to communicate to the user the implications of using a very disaggregated geography. Figure 3 on the right shows the sample size distribution for in-sample municipalities in EDOMEX with the white areas corresponding to municipalities not present in the ENIGH survey. Additionally, Table 2 shows the summary of the sample sizes for in-sample municipalities. The maximum sample size in a municipality is 527 households, the minimum is 3 households and the median is 21 households per municipality. In light of the very small sample sizes in some municipalities and the fact that there are many out-of-sample municipalities we conclude that the use of small area estimation for understanding the spatial distribution of poverty and inequality in municipalities in EDOMEX is well justified. At this stage the user with guidance from the methodologist needs to decide where there is a possibility to define a more aggregate level of geography. In the case of Mexico this alternative geography may be offered by districts. Doing so will increase the area-specific sample sizes and will considerably reduce the number of out-of-sample areas.

Table 2: Municipality sample sizes in the ENIGH survey and Census pilot data

	Min.	Q1.	Median	Mean	Q3	Max.
Survey	**3**	17	21	47.4	42	**527**
Census	**394**	2759	6852	24820	16440	**349100**

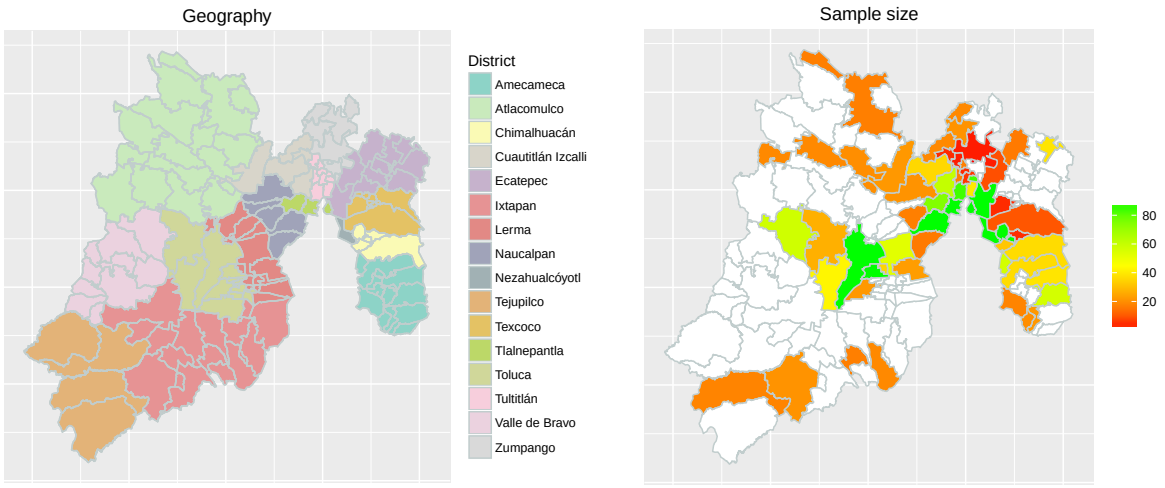

Figure 3: Geography and sample size distribution for municipalities in EDOMEX

3 Analysis/Adaptation

Having specified the user needs, the second stage in small area estimation involves the analysis of the data and the adaptation of the methods. Before starting with the detailed description in this Section, it is important to mention that in our view small area estimation should be governed by the principle of parsimony. That is, the analyst should be looking to use the simplest possible methods that provide small area estimates of acceptable precision. Section 3.1 presents a triplet of small area estimates described in 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 or skills. Ideally these 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 means of developing suitable models for small area estimation, both to improve the method of estimation and to be able to handle more complicated target parameters. Section 3.2 and 3.3 use the Mexico 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. Hence, The choice between a direct and an indirect estimator is about bias-variance trade-off, a frequently occurring dilemma in statistical inference. 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 the linear combination of the two. This defines the last estimator in the triplet of initial estimators given by

$$\hat{\theta}_k^{Composite} = \alpha_k \hat{\theta}_k^{Direct} + (1 - \alpha_k)\hat{\theta}_k^{Synthetic}, \tag{2}$$

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

Generally speaking, one may 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 Horvitz-Thompson (HT) estimator of the area population mean is

$$\hat{\bar{\theta}}_k^{Direct} = \frac{1}{N_k} \sum_{i=1}^{n_k} y_{ik}/\pi_{ik}, \tag{3}$$

where π_{ik} is the corresponding sample inclusion probability (Horvitz and Thompson, 1952). Note, when the population size for area k is unknown, N_k can be estimated by $\hat{N}_k = \sum_{i=1}^{n_k} 1/\pi_{ik}$. A synthetic estimator of the mean $\hat{\bar{\theta}}_k^{Synthetic}$ is given similarly, based on the sub-sample from a broad area including area k. The second situation is when auxiliary data are available and the estimation is based on the calibration weights (Deville and Särndal, 1992), denoted by w_{ik}, for the corresponding unit i in area k. In this case the direct estimator of the area population mean is defined by

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

The derivation of w_{ik} typically involves estimating the regression coefficients, $\hat{\boldsymbol{\beta}}$, of a so-called assisting linear model. A synthetic estimator $\hat{\theta}_k^{Synthetic}$ can thus be given by prediction using the same linear assisting model and the broad area estimates of the regression coefficients, $\hat{\boldsymbol{\beta}}$. The approach is largely the same with model-based prediction (Valliant et al., 2000). The direct estimator is derived under the prediction model fitted to the within-area sub-sample, and the synthetic estimator that under the same model fitted to a broad-area sub-sample or the entire sample.

There are several choices of α_k for the composite estimator (2), including the James-Stein estimator that uses a common α in all the areas, and the area-specific minimization of the MSE. An 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. We refer to Rao and Molina (2015) for additional details. It is worth noting that the composite estimator appears more intuitive for target parameters that are linear statistics of the y_{ik}'s, like domain averages. 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). Direct (and likewise synthetic) estimates of non-linear indicators and corresponding estimates of the variance can be produced by using package laeken in R (Alfons and Templ, 2013). However, a linear combination e.g. of a direct estimate of the Gini coefficient and that from a broad area seems hardly an appropriate 'composite' estimate of the Gini coefficient. 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 below in Section 4. This would require extra effort that may not always be possible as a by-product of the standard sample survey estimation set-up. At the very minimum, the direct estimates need to be analyse and uncertainty around direct estimates should be produced as this will offer an indication of the improvement required for producing small area estimates. If the initial estimates do not provide SA statistics of acceptable quality, then one approach is to revisit the specification stage and think of alternative geographies and additional data. This is rarely the case as most organisations that produce SA statistics have predetermined ideas about the target geographies. It is more common that based on the results from the initial set of estimates the analyst will subsequently consider the use of more complex possibly model-dependent SAE methods. In this case juxtaposing the direct, synthetic and composite estimates provide 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. Composite estimation yields an indication of baseline performance upon which the model-based small area estimation must aim to improve. This is obvious in the case of the widely used mixed-effects empirical best linear unbiased predictor (Rao and Molina, 2015), which can be given in a composite form directly.

3.2 Use of models for small area estimation

Small area estimation is one of the few 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 model to generate both linear and nonlinear small area estimates of interest. In Section 3.3 we describe model building, diagnostic analysis and model adaptations in 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. Estimation of non-linear indicators such as the Gini coefficient requires the use of unit-level survey data for the outcome variable and the covariates, and unit-level Census micro-data for the covariates. In light of this, the starting point for model-based small area estimation in this case is the unit-level nested error (random effects) regression model (Battese et al., 1988) and in particular methodologies that allow for the estimates of both linear and non-linear indicators. Elbers et al. (2003) propose a methodology that allows for the estimation both of linear and non-linear indicators. The methodology is based on the use of a nested error regression model with cluster random effects that is fitted by using survey data. The response variable, which is not available in the Census, is a welfare variable, 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 alongside a defined poverty line are then used for estimating non-linear indicators for example, the HCR and the Gini coefficient. Linear statistics can also be estimated by using the synthetically-generated welfare predicted values. Please note that in the approach by Elbers et al. (2003) clusters may not coincide with target domains.

Molina and Rao (2010) propose a methodology that is similar in spirit to Elbers et al. (2003) but uses Empirical Best Prediction (EBP). 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 point of departure of the EBP method 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. 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. In many applications linked survey and Census data are unavailable. 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 $\tilde{u}_k = E(u_k|y_k)$ is the conditional expectation of u_k given the within-area sample data y_k. 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 $u_k^* + \epsilon_{ik}^*$ that are generated from the above normal distributions create the conditional covariance structure of the y_{ik}'s in the population. 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, $\boldsymbol{y}^*$. With each vector of $\boldsymbol{y}^*$ linear and non-linear indicators are computed in each domain k and the estimates are averaged over L. A moderate number of Monte-Carlo simulations, for example, $L = 50$ or $L = 100$ should suffice. Note also that the EBP approach includes small area estimation of domain averages using the unit-level model (Battese et al., 1988) as a special case. Provided that clusters coincide with the target domains, for in-sample areas Molina and Rao (2010) demonstrate the superior performance of the EBP approach when compared to Elbers et al. (2003) methodology. Estimation for out-of-sample domains requires additional discussion. For out-of-sample domains the predicted random effect is 0 and (5) becomes

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

When domains and clusters coincide, EBP point estimates for out-of-sample domains coincide with the point estimates from the Elbers et al. (2003) methodology. It is straightforward to see that these are regression synthetic estimates. For example, for domain averages, $E(y_{ik}^*) = \bar{\boldsymbol{x}}_k\hat{\boldsymbol{\beta}}$ since $E(\epsilon_{ik}^*) = 0$ and $E(u_k^*) = 0$, which is the regression synthetic estimator with regression coefficients estimated under the linear mixed model.

MSE estimation for model-based small area estimation will be discussed in some detail in Section 4.1. For now we notice that evaluation of the uncertainty both for in-sample and out-of-sample domains is performed under the unit-level nested error regression model. We return to this point in Section 4.2 that discusses methods for analytic evaluation of SAE estimates.

3.3 Model building, residual diagnostics and transformations in practice

Before considering model-based estimation, some assessment of initial estimates produced with the Mexico data is needed for motivating the use of more complex methods. In the case of Mexico the data provider did not supply the initial triplet estimates we described in Section 3.1. Our analysis below replicates such initial estimates in spirit. Figure 4 presents direct estimates and corresponding coefficients

of variation (CVs) of average equivalised household income for the municipalities in EDOMEX using the ENIGH survey data. These are produced by using the sae package in R (Molina and Marhuenda, 2015). In particular, we use the function direct, which incorporates the survey weights. The estimated variances are approximated by assuming that the joint inclusion probabilities are the product of the first order inclusion probabilities. 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 can not be produced for the out-of-sample municipalities (white coloured areas). Hence, unless the user needs change, we should explore the use of model-based methods.

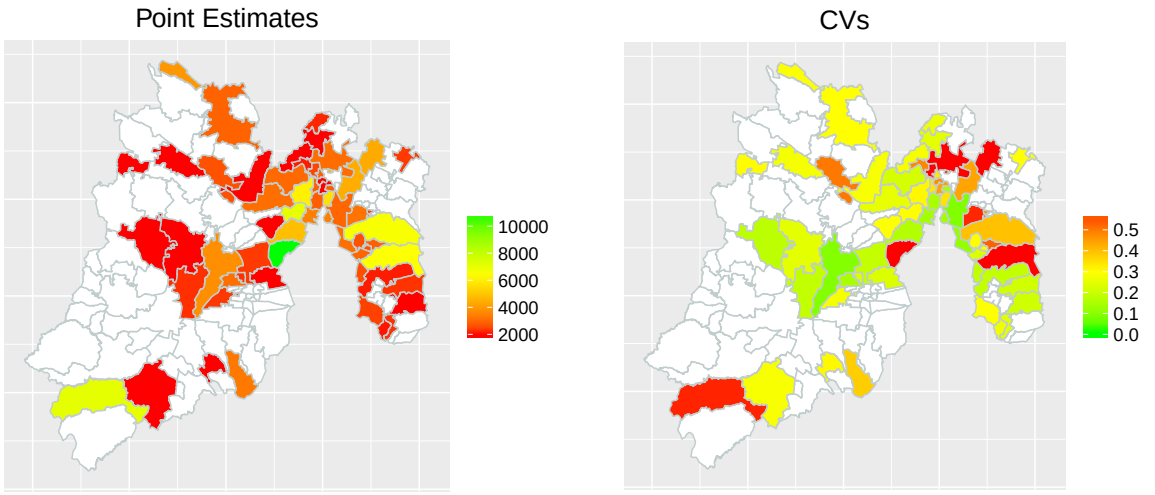

Figure 4: Direct estimates of average household equivalised income and corresponding CVs in EDOMEX municipalities

The use of models aims to improve the precision of small area estimates by making optimal use of the data available. Inference is under the model and the impact of potential model misspecification must be carefully examined. 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 practise guidelines one could follow, and illustrate these guidelines for estimating income related indicators with the data from Mexico.

For producing model-based estimates the analyst needs to build the model that as we discussed in the previous Section usually includes area random effects. However, before discussing the use of random effects the most important part of the model, in our view, is the fixed effects one. 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. There are a number of reasons for this. To name a few, random effects complicate model testing, point and MSE estimation. Hence, if random effects can be avoided, it is best to do so. An acceptable 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, the analyst can consider a marginal regression model that allows for correlated errors without including area-specific effects,

$$y_{ik} = \boldsymbol{x}_{ik}^T \boldsymbol{\beta} + \epsilon_{ik}, \boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \boldsymbol{\Sigma}). \tag{7}$$

In (7) $\boldsymbol{\Sigma}$ can be specified to have the simplest possible correlation structure i.e. an exchangeable correlation structure that corresponds to the commonly used unit-level nested-error regression model with domain random intercepts. In order to decide whether to include a covariate in the fixed part of the model one can use information criteria for example, the Akaike or the Bayesian Information Criteria (AIC, BIC) computed either under the standard linear model with uncorrelated errors or as pointed out by Vaida and Blanchard (2005) under the marginal model (7). In the case of the data from Mexico and following the recommendation by the data provider (CONEVAL), y is defined by 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 (i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$) the following covariates that are available both in the survey and Census data have been identified as good predictors of *ictpc*. Use of a marginal model with an exchangeable correlation structure leads to the same set of covariates:

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.

Our suggestion to start by building a model without explicit use of random effects is based on the principle of parsimony. To investigate whether the use of a mixed effects model is necessary we next fitted a linear model using generalized least squares (GLS) with an exchangeable correlation structure which is equivalent to a mixed effects model with domain-specific random intercepts. In this way we explored the plausibility of composite estimation at the same time. The class of GLS models contains the standard linear model that assumes independent observations as a special case. Therefore, given a fixed effects specification, the standard linear model is nested within the model with exchangeable correlation structure and a likelihood-ratio test or information criteria may be used to decide if the model that allows for within municipality correlation fits the data better. First, we compared the GLS with an exchangeable correlation structure against a standard linear model when both models included only an intercept term. This allows us to quantify the extend of the between domain variability. The improvement by using an exchangeable correlation structure turned out to be significant. The AIC for the GLS with an exchangeable correlation structure only is 54239 whereas the corresponding AIC for the standard linear model is 54275. In the second step the GLS and standard linear regression models with the set of six covariates identified above were compared against each other. Here the improvement by allowing for an exchangeable correlation structure was marginal. The AIC suggests that the model with the exchangeable correlation structure should be preferred (AIC for GLS with an exchangeable correlation structure: 53077 vs. AIC for the standard linear model: 53079). We may, therefore, expect that the use of a mixed effects model will offer a benefit for SAE, albeit possibly small. On the one hand, a large number of areas are out-of-sample and for these areas we must rely on regression synthetic estimation. On the other hand, the covariates we decided to include in the model seem to explain a substantial part of the between municipalities variability. In particular, the intra cluster correlation (ICC) for the intercept only mixed effects model is 5.4% and for the mixed effects model that includes the six significant predictors it reduces to 1.5%. 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). The impact of including random effects is assessed in Section 4.3.2 by comparing regression synthetic to EBP estimates. Based on the results in this Section we should expect that EBP estimates will be somewhat more efficient than regression synthetic estimates.

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 utilises residual diagnostics and assessment of the predictive power of the model. Despite the inclusion of a number of significant covariates, the model may have very 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. For residual diagnostics we propose the use of graphical diagnostics in the form of 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 real applications some adaptations to 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 5 shows Normal probability plots of household-level and municipal-level residuals obtain by fitting model (4) using the raw income data, the six covariates we identified above and including municipality-specific random effects. As we should expect when using the raw income data, there are severe departures from normality. This can be seen both from the shape of the Q-Q plots and from Table 3 where the skewness and kurtosis of the two sets of residuals are clearly different from what the values we should expect under normality.

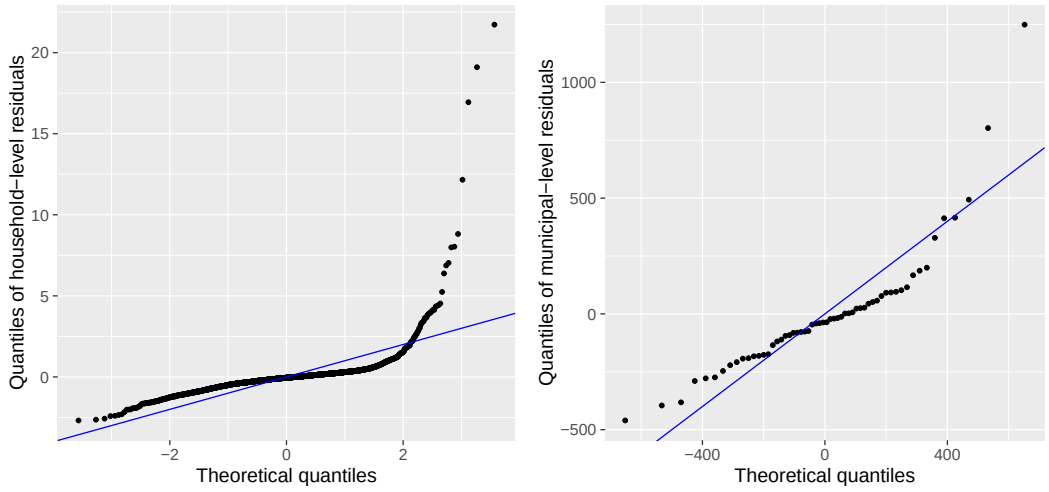

Figure 5: Normal Q-Q plots for household residuals (left) and municipality residuals (right) obtained from the model that uses raw income as the response variable

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. To name two, in the case of income data possible distributions include the Pareto distribution and the Generalised Beta distribution of the second kind (GB2). The complication with using alternative distributions is that the analyst needs to develop new estimation and inference theory for every new application. For income-related applications Graf et al. (2015) proposed an EBP methodology under the GB2 distribution. However, the general complication remains. In addition, what if there are departures from this alternative distribution? 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. Here we refer to outlier robust estimation (Datta and Lahiri, 1995; Ghosh et al., 2008; Sinha and Rao, 2009; Chambers 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 in 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 opting for more complex model adaptations.

The applications in the papers by Elbers et al. (2003) and Molina and Rao (2010) considered the use of a logarithmic transformation which is a reasonable one for income data. However, is a logarithmic transformation the most appropriate one? Can alternative transformations offer better predictive power? Rojas-Perilla et al. (2015) currently investigates the use of a wide range of transformations. For the purposes of the case study with the data from Mexico in addition to the logarithmic transformation we focus on the log-shift transformation and on power transformations (Box and Cox, 1964; Gurka et al., 2006). Denoting by $\mathrm{T}_\lambda(y_{ik})$ the transformed outcome, the log-shift transformation is defined by

$$\mathrm{T}_\lambda(y_{ik}) = log(y_{ik} + \lambda). \tag{8}$$

An empirical approach for choosing λ in (8) 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. This means select a transformation that makes the skewness of the distribution of the residuals close to 0. Note, however, that here we deal with two sets of residuals and to the best of our knowledge there is no formal approach to defining the distance from normality. For the purposes of the application in this paper we focus on the household-level residuals as these are the residuals where users detect problems with departures from normality more commonly. An alternative approach and one we are currently investigating is to use a scaled version of the log-shift transformation and the select the value of λ that maximizes the residual log-likelihood or the log-likelihood under the model. The second type of transformations we explore are scaled power transformations defined by

$$\mathrm{T}_\lambda(y_{ik}) = \begin{cases} \frac{(y_{ik}+s)^\lambda - 1}{\alpha^{\lambda-1}\lambda}, & \lambda \neq 0 \\ \alpha \log(y_{ik}+s), & \lambda = 0 \end{cases}, \tag{9}$$

for $y_{ik} > -s$ and α is the geometric mean of y_{ik}. Conditional on α, the Jacobian of the transformation is equal to 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 results standard software for fitting this model with the transformed data. 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 residual 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.

Using the data from Mexico we apply the EBP method with three transformations for the outcome namely, log, log-shift and scaled Box-Cox. Figure 6 on the right shows the graphical representation of the maximization of the residual maximal log likelihood (REML) on a grid $\lambda \in [-2; 2]$ for finding the optimal λ 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 approximately equal to 319.52. The question is

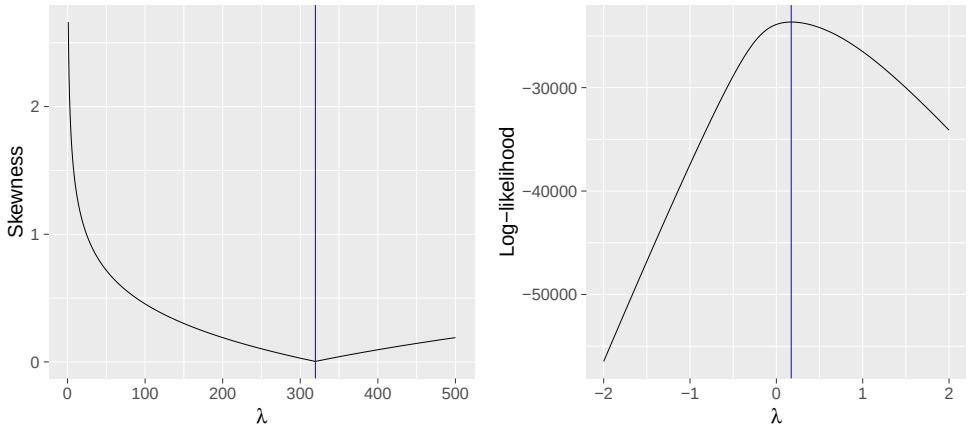

Figure 6: Shift parameter for the log-shift transformation (left) and optimal λ for the Box-Cox transformation (right)

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 8) and the distribution of the residuals in Table 3. For municipality random effects, all three transformations offer a good approximation to normality (see also Table 3). The picture is different for household-level. In particular, the household-level residual 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 3) 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 7. 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 look similar and they are available from the authors upon request.

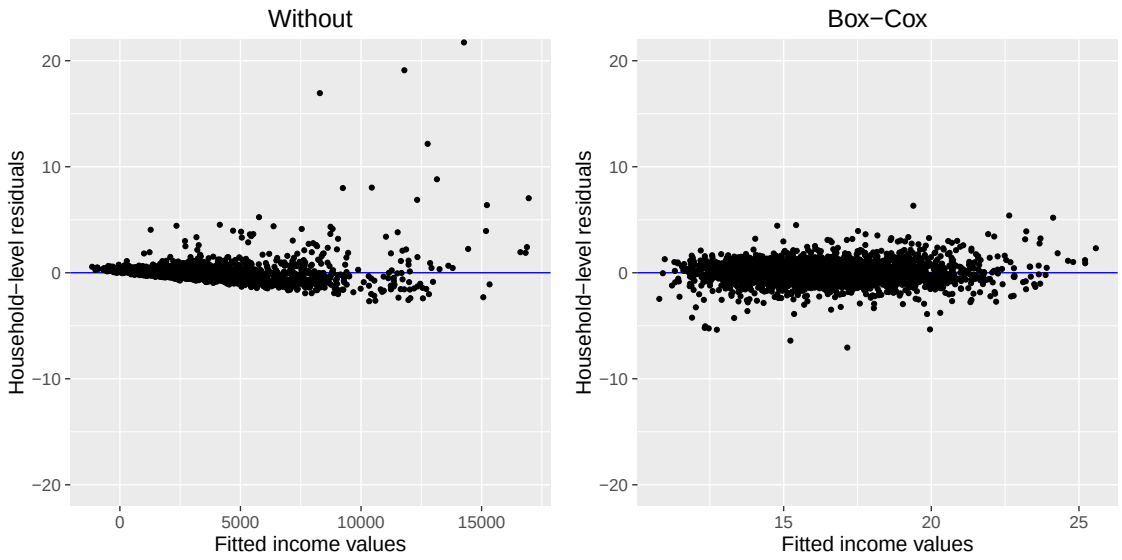

Figure 7: Standardized household residuals against fitted values without (left) and with Box-Cox transformation (right) for income.

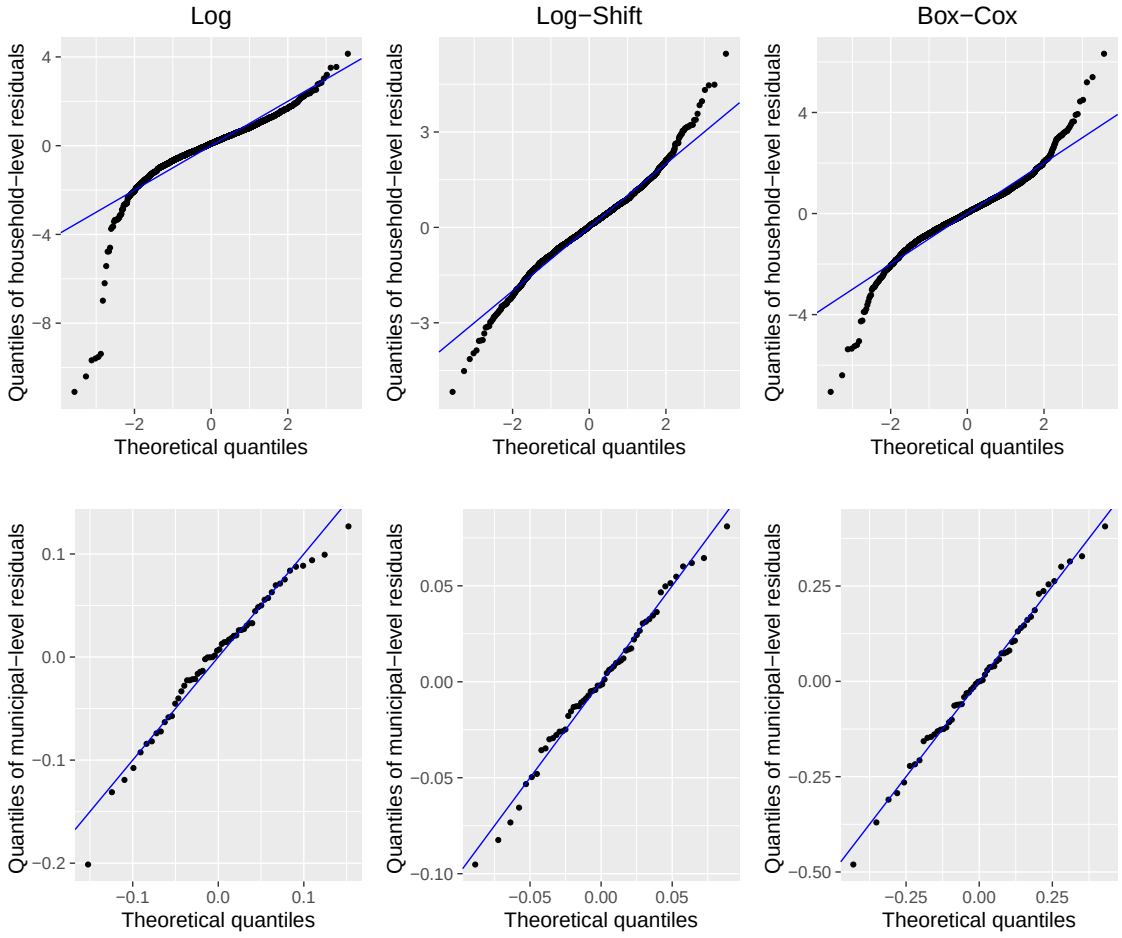

Figure 8: Normal Q-Q plots for household residuals and municipality residuals under three transformations for income

Table 3: Skewness and kurtosis for the random effects and error terms of the working models for EBP with and without transformations

	Household-level error term		Random municipality effects	
Transformation	Skewness	Kurtosis	Skewness	Kurtosis
Without	10.10	177.00	2.09	9.87
Log	-2.71	26.50	-0.60	3.52
Log-shift	0.00	4.91	-0.24	3.03
Box-Cox	-0.24	7.95	-0.12	3.00

The proportion of variability explained under each model is quantified by the coefficients of determination R^2 summarized in Table 4. As pointed out before, using the raw values of income in the EBP nested-error regression model provides clearly unsatisfactory normal Q-Q plots and a R^2 equal to 31%. The use of transformations improves the predictive power of the model with the log-shift and Box-Cox transformations giving the best predictive power.

Based on the results from the diagnostics analysis above we conclude that two transformations namely, log-shift with shift parameter $\lambda = 319.52$ and Box-Cox with $\lambda = 0.17$ provide the most promising predictive power and approximation to the model assumptions of the EBP method. 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 and less severe departures from the model assumptions translate to more precise small area estimates? 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 Figures 9-11 that show maps of point estimates of average income, Gini coefficients and HCR for municipalities in EDOMEX produced by the EBP approach using different transformations.

Table 4: Coefficients of determination from different linear regression model

Transformation	R^2	λ
Without	0.31	-
Log	0.43	-
Log-shift	0.51	319.52
Box-Cox	0.49	0.17

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 (see Figure 3). 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 to be less affected by the choice of transformation. The same holds true to a large extend 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 interested in non-linear indicators that depend

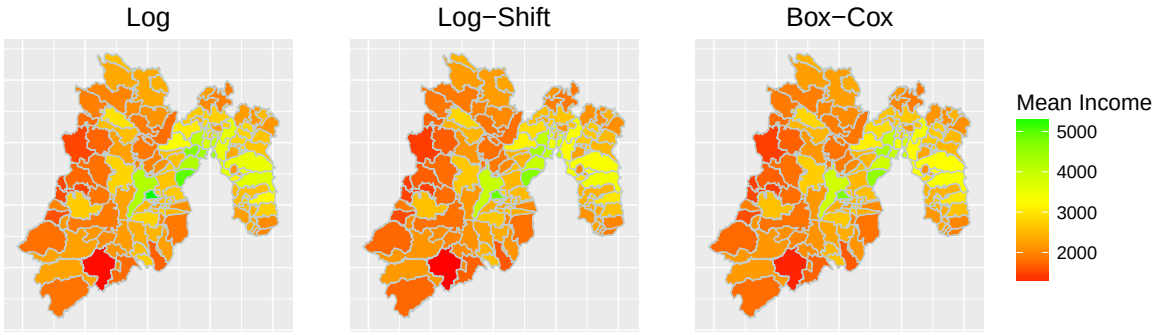

Figure 9: Map of municipal estimates of average income in EDOMEX using the EBP method under the log, log-shift and Box-Cox transformations

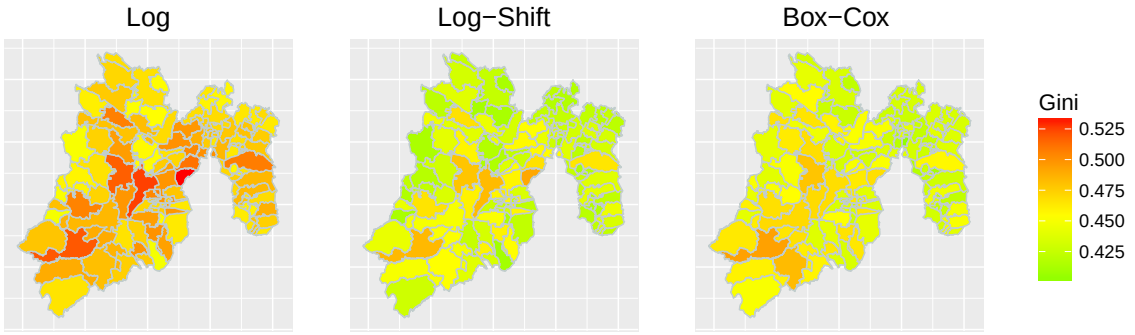

Figure 10: Map of municipal estimates of Gini coefficients in EDOMEX using the EBP method under the log, log-shift and Box-Cox transformations

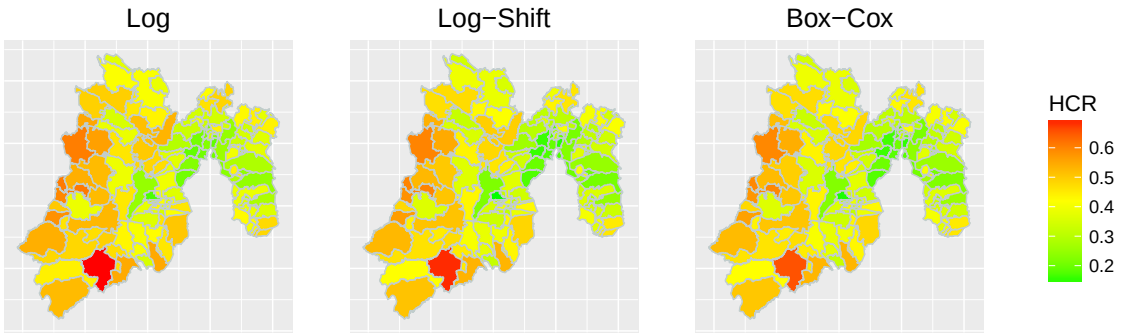

Figure 11: Map of municipal estimates of HCR in EDOMEX using the EBP method under the log, log-shift and Box-Cox transformations

on the entire distribution. The comments in this part refer only to point estimation. In the next Section we provide a framework that allows for the estimation of the uncertainty around point estimates and for method evaluation. We will therefore open this discussion again at the end of Section 4.

4 Evaluation

The small area estimates are a set of numbers of identical definition and simultaneous interest. In what sense can one set of estimates considered to be better than another? Is it enough that the underlying model is the preferred one according to some model selection criterion? Given two sets of 'optimal' estimates, derived under two different models, is it meaningful to compare the respective MSEs to each other directly? Is it enough that the average MSE of one set is better than the other? Should one rely on the design- or model-based MSE? Are ensemble properties of the small area estimates such as the range or ranks of the estimates relevant? These are all examples of questions, to which, in our opinion, there are hardly any definite answers. A detailed discussion on evaluation is beyond the scope of this paper. Our approach below is in Section 4.1 to describe some aspects of evaluation, which we believe are *necessary* to be taken into consideration in any application. In Section 4.2 we highlight the importance of seeking a common ground when comparing estimators under different models, which in our experience is a matter that is often either misunderstood or overlooked. Some aspects of evaluation are illustrated with the data from Mexico in Section 4.3.

4.1 Some general aspects of evaluation

Let θ_k be the generic target parameter of area k, for $k = 1, ..., m$. Let $\boldsymbol{\theta} = \{\theta_1, ..., \theta_m\}$ denote the collection of them. Let $\tilde{\theta}_k$ be the estimator of θ_k and $\tilde{\boldsymbol{\theta}}$ the collection of them. Here we assume that one is interested in *all* elements of $\boldsymbol{\theta}$ at the same time. In other words, one can not fix only on one particular θ_k, or a few of them, and disregard how estimators perform in the rest of the areas.

Generally speaking, in small area estimation one may distinguish between the ***area-specific*** and ***ensemble*** properties of $\tilde{\boldsymbol{\theta}}$. 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 $N = \sum_{k=1}^m N_k$, and let $G = \sum_{k=1}^m(\theta_k - \bar{\theta})^2/(m-1)$ be the dispersion (or empirical 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 $\boldsymbol{\theta}$. The various ensemble properties of $\tilde{\boldsymbol{\theta}}$ are important for purposes such as benchmarking, subgroup analysis, fund allocation, evaluation and monitoring (see e.g. Ghosh, 1992; Shen and Louis, 1998). Although ensemble characteristics have attracted some interest, area-specific prediction has been the focus in the majority of applications. The most common area-specific uncertainty measure is MSE. Below we describe three *types* of MSE in common use.

Denote by y_k be the observed data in area k, for $k = 1, ..., m$. Let $\boldsymbol{y} = \{y_1, ..., y_m\}$ denote the collection of them. Provided a population model of $\boldsymbol{\theta}$, the *(unconditional) MSE* is given by $E[(\tilde{\theta}_k - \theta_k)^2]$, where the expectation is over both $\boldsymbol{\theta}$ and $\boldsymbol{y}$. Prasad and Rao (1990) develop second-order accurate analytic MSE estimator under the linear mixed model. Jackknife methods have been developed for the same purpose under a wider range of models (Jiang et al., 2002; Lohr and Rao, 2009). 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). 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 unconditional MSE estimation uses parametric bootstrap which proceeds as follows. Generate B bootstrap populations using the fitted super-population model $\boldsymbol{y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{u}^* + \boldsymbol{e}^*$. Compute the population value of the target parameter from each bootstrap population, θ_k^*. From each bootstrap population select a bootstrap sample and compute bootstrap estimates of the target parameter, $\tilde{\theta}_k^*$, by using the same method as the one used with the

original sample. Finally, compute the average of the B squared bootstrap errors -defined by the difference between $\tilde{\theta}_k^*$ and θ_k^*- as an estimate of the unconditional MSE.

Booth and Hobert (1998) argue for the use of *conditional MSE of prediction (CMSEP)* in the case of non-normal-theory mixed models including generalised linear mixed models. The CMSEP is given by $E[(\tilde{\theta}_k - \theta_k)^2|y_k]$, where the corresponding within-area y_k is held fixed. When all the model parameters are known, the best predictor is $\tilde{\theta}_k = E(\theta_k|y_k)$, and the only natural measure of its uncertainty is the CMSEP that reduces to $V(\theta_k|y_k)$. 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 the *finite-population (FP) MSE* given by $E[(\tilde{\theta}_k - \theta_k)^2|\boldsymbol{\theta}]$, where only the observed data $\boldsymbol{y}$ are allowed to vary. The FP-MSE is often evaluated with respect to the sampling design (Rivest and Belmonte, 2000). Chambers et al. (2011) develop an FP-MSE estimator where the sampling distribution of $\boldsymbol{y}$ is specified with respect to different models in common use. The fact that $\boldsymbol{\theta}$ is held fixed means the FP-MSE is an uncertainty measure for the actual population from which the sample is selected instead of over all possible populations from which the sample could have been selected. This is a familiar interpretation that appeals to many users. However, 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.

Interval estimation may be considered besides MSE estimation. Let $C_k = (\tilde{\theta}_{kL}, \tilde{\theta}_{kU})$ be an interval estimator of θ_k, where $\tilde{\theta}_{kL} < \tilde{\theta}_{kU}$. Let $\delta_k = 1$ if $\theta_k \in C_k$ and 0 otherwise. Analogous to unconditional MSE, the unconditional *coverage* of C_k is given by $\alpha_k = E(\delta_k) = P(\theta_k \in C_k)$, 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 $E(\delta_k|\theta)$. Again, any unbiased C_k can have rather erratic area-specific FP-coverage compared to the nominal level of confidence. Zhang (2007) defines $\alpha = \sum_{k=1}^{m} E(\delta_k|\theta)/m$ to be the FP *simultaneous coverage* of the C_k's, all 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 $m \to \infty$, α converges to the nominal level of the C_k's, provided the underlying population model of $\boldsymbol{\theta}$ is correct.

Up to this point, in this Section we described different measures of uncertainty. In addition to measuring the uncertainty around $\tilde{\boldsymbol{\theta}}$, an analyst is more generally interested in *method evaluation*. An analyst may be interested in 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. In general there are three approaches to method evaluation namely, model-based simulation, design-based simulation and analytic method evaluation. Model-based and design-based simulations are common in practice and an example of a design-based simulation using the data from Mexico is presented in Section 4.3. Analytic method evaluation is a very powerful tool but less common in practice. We recommend that whenever possible analysts should also consider using analytic method evaluation. An example of this for comparing estimators under two different models is presented in Section 4.2.

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 FP 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. For each simulated sample, a given estimation method is applied to obtain a replicate set of small area estimates. In particular, within a

design-based simulation study different estimation methods or models can be directly compared to each other in terms of their FP 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 data from Mexico 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 from Mexico.

Unlike in a design-based simulation study, where the different methods are subjected to the same source of uncertainty and usually the population is not generated under a model, ***model-based*** method evaluation generally requires the use of a model for generating a population. Model-based studies are more common when researchers develop new methods and they are interested in evaluating the properties of estimators (point and MSE) both when the model assumptions hold and under specific assumptions about model misspecification (sensitivity analysis). The design of model-based studies requires careful thinking about the choice of the *evaluation model* used for generating the population. An example relevant to this paper is when using parametric bootstrap for MSE estimation under the EBP method. The simulation of a bootstrap population will differ depending on whether the super-population model we assume holds. There arises a question i.e. whether it is meaningful to compare the MSE of an estimator θ_{kA} of θ_k derived under model M_A to that of another estimator θ_{kB} of θ_k under model M_B directly. Notice that it is always possible to evaluate the MSE of θ_{kA} under model M_B even though the estimator is motivated and computed under model M_A and *vice versa*. Since the MSE of θ_{kA} will differ according to whether the evaluation model is M_A or M_B, there is a need to compare the estimators within a common framework in order to avoid misleading comparisons. Model-based evaluation is not the focus of this paper, the matter will be discussed more closely in Sections 4.2 and 4.3.

Before concluding this Section we must mention that in addition to the evaluation methods mentioned above i.e. uncertainty assessment and design, model-based and analytic method evaluation, there are additional informal evaluation approaches that are of relevance to practitioners. A set of small area estimates is expected to be numerically consistent and more efficient than unbiased direct estimates. One can further compare the aggregated area estimates to the corresponding direct aggregate estimates for the same purpose. If aggregated model-based (indirect) estimates are not consistent with aggregate direct estimates, an analyst can use benchmarking techniques to achieve consistency. Benchmarking offers estimates with built-in consistency an attractive property for NSIs (see Pfeffermann, 2013, for a discussion on benchmarking methods). Notice, however, that direct estimates can not necessarily be assumed to be unbiased when it comes to ensemble characteristics of area parameters. For example, direct estimates can generally be expected to over-estimate the dispersion of $\boldsymbol{\theta}$. Use of informal evaluation approaches such as compatibility with external data, quantitative or qualitative, evaluation by subject-matter experts, bias and goodness of fit diagnostics from the point of view of practitioners are described in Brown et al. (2001).

4.2 Levelling the common ground for model-based evaluation

To illustrate the idea of levelling the common ground for evaluation, assume estimation of θ_k under a fixed effects model, but allow in addition an associated mixed effects model of θ_k for evaluation. Let the

two models be given by, respectively,

$$\begin{aligned}\mathcal{F}: &\quad \theta_k = \mu_k \\ \mathcal{M}: &\quad \theta_k = \mu_k + u_k\end{aligned}$$

where $\mu_k = \mu_k(\boldsymbol{\beta})$ is the fixed effects predictor, such as e.g. $\mu_k = \bar{\boldsymbol{x}}_k\boldsymbol{\beta}$ for small area means, and the u_k's are IID random effects with $E(u_k) = 0$ and $V(u_k) = \sigma_u^2$. Table 5 below summarises how the MSE of the synthetic estimator of θ_k varies according to whether the evaluation model is the fixed effects or mixed effects model, with or without conditioning on the observed data.

Table 5: MSE of the synthetic estimator under a fixed and a mixed effects model

Evaluation		Unconditional MSE	Conditional MSE
BSE	$\mathcal{F}$	0	0
	$\mathcal{M}$	σ_u^2	$E(u_k^2\|y_k)$
EBSE	$\mathcal{F}$	$V_{\mathcal{F}}(\hat{\mu}_k)$	$E_{\mathcal{F}}[(\hat{\mu}_{0k} - \mu_k)^2]$
	$\mathcal{M}$	$\sigma_u^2 + V_{\mathcal{M}}(\hat{\mu}_k)$	$E(u_k^2\|y_k) + E_{\mathcal{M}}[(\hat{\mu}_{0k} - \mu_k)^2] - 2E(u_k\|y_k)E_{\mathcal{M}}(\hat{\mu}_{0k} - \mu_k)$

To start with, provided the parameter $\boldsymbol{\beta}$ is known, the best synthetic estimator (BSE) is simply μ_k. Therefore, under model $\mathcal{F}$, its MSE is 0, both conditionally and unconditionally. Meanwhile, under model $\mathcal{M}$, we have $\theta_k - \mu_k = u_k$. Therefore, the unconditional MSE of μ_k is σ_u^2, and it is $E(u_k^2|y_k)$ conditionally. Notice that the conditional MSE is only applicable to an area that is present in the sample. In any case, the MSE of μ_k is clearly larger under the mixed effects model $\mathcal{M}$ both conditionally and unconditionally. In practice, however, the parameter $\boldsymbol{\beta}$ needs to be estimated, yielding the empirical BSE (EBSE) $\hat{\mu}_k = \mu_k(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the estimator of $\boldsymbol{\beta}$. The unconditional MSE is then given by $E_{\mathcal{F}}[(\hat{\mu}_k - \mu_k)^2] = V_{\mathcal{F}}(\hat{\mu}_k)$, where we assume $E_{\mathcal{F}}(\hat{\mu}_k) = \mu_k$, and the subscript indicates that the expectation is calculated under model $\mathcal{F}$. On the other hand, conditional on y_k, $E_{\mathcal{F}}[(\theta_k - \hat{\mu}_k)^2|y_k] = E_{\mathcal{F}}[(\hat{\mu}_{0k} - \mu_k)^2]$, where $\hat{\mu}_{0k}$ is the estimator of μ_k when y_k is held fixed while the data from all the other areas are allowed to vary.

Evaluated under model $\mathcal{M}$, unconditionally we have

$$E_{\mathcal{M}}[(\theta_k - \hat{\mu}_k)^2] = E_{\mathcal{M}}[u_k^2 + (\hat{\mu}_k - \mu_k)^2 - 2u_k(\hat{\mu}_k - \mu_k)] = \sigma_u^2 + V_{\mathcal{M}}(\hat{\mu}_k),$$

provided $E_{\mathcal{M}}(\hat{\mu}_k) = \mu_k$ despite $\hat{\mu}_k$ is estimated under model $\mathcal{F}$. On the other hand, conditionally

$$\begin{aligned}E_{\mathcal{M}}[(\theta_k - \hat{\mu}_k)^2|y_k] &= E_{\mathcal{M}}[u_k^2 + (\hat{\mu}_k - \mu_k)^2 - 2u_k(\hat{\mu}_k - \mu_k)|y_k] \\ &= E(u_k^2|y_k) + E_{\mathcal{M}}[(\hat{\mu}_{0k} - \mu_k)^2] - 2E(u_k|y_k)E_{\mathcal{M}}(\hat{\mu}_{0k} - \mu_k)\end{aligned}$$

provided area k is present in the sample, since u_k is independent of $\hat{\mu}_{0k}$ given y_k. Moreover, notice that $E_{\mathcal{M}}[(\hat{\mu}_{0k} - \mu_k)^2]$ is different from $E_{\mathcal{F}}[(\hat{\mu}_{0k} - \mu_k)^2]$. Provided all the within-area sample sizes remain bounded, as $m \to \infty$, the leading terms are σ_u^2 for the unconditional MSE and $E(u_k^2|y_k)$ for the conditional MSE, respectively, which do not vanish asymptotically, such that the MSE of EBSE $\hat{\mu}_k$ is larger under the mixed effects model both conditionally and unconditionally.

The previous discussion illustrates how the performance of an estimator varies according to the evaluation model in general. To appreciate why this matters for method choice, simply compare the synthetic estimator derived under model $\mathcal{F}$ to the best predictor (BP) derived under model $\mathcal{M}$ assuming all the

model parameters are known. The CMSEP of the BP is $V(u_k|y)$ under model $\mathcal{M}$, provided area k is represented in the sample, and it is σ_u^2 when the area is not represented in the sample. It is thus clear that it does not make sense to compare the conditional or unconditional MSE of the BSE under model $\mathcal{F}$, which is 0 (Table 5), with the CMSEP of the BP under model $\mathcal{M}$, which is not 0. Instead, one can compare the MSEs when both are evaluated under the same model $\mathcal{M}$, and see right away that the BP has a smaller MSE as long as the area is present in the sample, whereas for an out-of-sample area the MSE will be σ_u^2 for both the BSE and the BP.

4.3 Illustrating aspects of SAE evaluation using the data from Mexico

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 6 presents summaries over municipalities of point, root MSE (RMSE) and CV estimates computed using the original data supplied to us by CONEVAL. 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. Note that synthetic estimates (point and MSE) are produced under the mixed model (see discussion in Section 4.2).

Why have we decided to use the EBP method? This is because the model building and diagnostics results in Section 3.3 showed that the inclusion of municipality mixed model provided a better fit to the data. Why have we decided to use both the raw income and three different transformations? The model diagnostic results in Section 3.3 showed that the best transformations are the log-shift and the Box-Cox. Do these results, however, translate in estimates with better efficiency?

The results in Table 6 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. 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 transforma-

tion) are met. In reality there are always small departures from the model assumptions. One question is whether small departures can have an impact on MSE estimation. The second 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 the properties of MSE estimation under model misspecification requires the design of model-based or design-based simulations. This can be very computer intensive because a computer-intensive method such as parametric bootstrap is implemented for a large number of Monte-Carlo runs. We discussed this issue again in the next Section and we draw some preliminary conclusions of relevance for the prospective analyst.

Table 6: One sample analysis. Summaries of point estimates, estimated RMSEs and CVs over municipalities in EDOMEX

58 In-sample municipalities							
	Indictor	Mean		Head Count Ratio		Gini	
	Estimator	Median	Mean	Median	Mean	Median	Mean
Point Estimates	EBP	2730	2875	0.380	0.374	0.949	1.020
	EBP Log	2699	2927	0.363	0.354	0.477	0.481
	EBP Log-shift	2600	2782	0.329	0.323	0.433	0.439
	EBP Box-Cox	2617	2780	0.336	0.326	0.435	0.441
RMSE	EBP	449.2	435.8	0.040	0.038	0.177	1.110
	EBP Log	249.7	257.2	0.039	0.037	0.011	0.012
	EBP Log-shift	202.3	207.0	0.036	0.035	0.010	0.010
	EBP Box-Cox	185.2	186.2	0.034	0.032	0.010	0.010
CV	EBP	0.163	0.171	0.104	0.104	0.187	0.570
	EBP Log	0.095	0.091	0.108	0.108	0.024	0.024
	EBP Log-shift	0.080	0.077	0.112	0.112	0.022	0.023
	EBP Box-Cox	0.071	0.069	0.103	0.103	0.022	0.022
67 Out-of-sample municipalities							
Point Estimates	EBP	2042	2130	0.436	0.433	1.261	1.344
	EBP Log	2244	2296	0.439	0.436	0.474	0.473
	EBP Log-shift	2151	2196	0.409	0.410	0.432	0.433
	EBP Box-Cox	2171	2216	0.409	0.410	0.440	0.445
RMSE	EBP	523.4	518.6	0.048	0.048	0.400	8.102
	EBP Log	256.1	256.3	0.050	0.049	0.013	0.013
	EBP Log-shift	209.3	209.9	0.048	0.048	0.011	0.011
	EBP Box-Cox	188.4	187.7	0.043	0.043	0.011	0.011
CV	EBP	0.251	0.265	0.114	0.111	0.313	3.366
	EBP Log	0.111	0.112	0.119	0.116	0.027	0.028
	EBP Log-shift	0.095	0.096	0.122	0.120	0.025	0.026
	EBP Box-Cox	0.085	0.085	0.110	0.108	0.025	0.026

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. How much are the results affected by the error of the model parameter estimate, assuming that the model is correct? What if the model assumptions are not all correct? These natural questions suggest why it is helpful to perform design-based method evaluation that does not depend on the model assumptions

altogether. We now present an approach for designing 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 what fixed population to use. In the Mexican Census we identified variable *inglabpc - earned per capita income from work* which is highly correlated with the variable of interest *ictpc* that is only available in the survey data. Variable *inglabpc* is not considered to provide income data of high quality 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. We further wanted to perform method evaluation in a situation where the covariates available explain a moderate part of the variance. Table 7 presents summary statistics for variable $inglabpc$ (used in the design-based simulation) and the variable $ictpc$ (used in the one sample analysis). As expected the distribution of both variables is similar and the total per-capita income $ictpc$ is 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 $ictpc$ in Table 7, 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 provided by *inglabpc*. Note that since the data for the design-based simulation are coming from the Census, no model-based assumptions are being made.

Table 7: Summary statistics for the income variables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$inglabpc$ (Census)	0	1000	1700	2717	3000	100000
$ictpc$ (survey)	0	1310	2142	3243	3518	98070

Table 8: Summary statistics over municipalities

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA
Population size	394	2759	6852	24820	16440	349100	-
Sample size	3	17	21	47.4	42	527	67

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 8.

Using each sample selected from the fixed population we compute estimates of average equivalised household income from work, HCR and Gini coefficients using direct and indirect estimators. In particular, we use the direct (3), the EBP based on different transformations and regression synthetic estimators. Indirect estimators are computed with a model that uses the same six covariates identified via model building in Section 3.3. The R^2 was on average -over repeated sampling- around $40 - 50\%$, which is consistent with the results we obtained from one sample analysis.

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

squared error (RMSE) given by

$$\text{Relative Bias}(\hat{\theta}_k) = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\theta}_{tk} - \theta_{tk}}{\theta_{tk}}$$

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

where $\hat{\theta}_k$ is a 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 CVs as a additional performance indicator.

Table 9 reports the results split by the 58 in-sample and the 67 out-of-sample municipalities. The table presents mean and 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 to 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). For in-sample areas we note that the synthetic estimates are somewhat less efficient than the model-based estimates. Despite the small between-area variability we obtain from the full model we used for estimation, including random effects is recommended for the in-sample municipalities. The comparison between synthetic and EBP estimates, however, highlights the importance of building a model that has good predictive power. Doing so means that use of a synthetic estimator may not be that inefficient.

It is important to also 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. An informal approach to evaluating the performance of parametric bootstrap is by comparing the empirical FP RMSEs reported in Table 9 to the estimated unconditional RMSEs reported in Table 6. Generally, it can be shown that, across all the areas, the mean (or median) of the FP RMSE is often comparable to that of the unconditional RMSE provided the population model of the small area parameters is correct. This seems to be case for estimates of average income and HCR but not for the Gini Coefficient where we see that the estimated unconditional RMSEs are much smaller than the empirical FP RMSEs. We conjecture that even small departures from the model assumptions can have an adverse effect on MSE estimation of non-linear indicators computation of which depends on the entire target distribution. We are currently investigating the use of non-parametric methods for MSE estimation with non-linear indicators. The preceding analysis shows that practitioners must be particularly careful when using parametric methods and should always employ design-based method evaluation.

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

Table 9: Performance of predictors over municipalities in design-based simulations

58 In-sample municipalities							
	Indictor	Mean		Head Count Ratio		Gini	
	Estimator	Median	Mean	Median	Mean	Median	Mean
RMSE	EBP	180.2	236.3	0.095	0.091	0.497	0.538
	EBP Log	187.5	229.5	0.049	0.051	0.026	0.031
	EBP Log-shift	156.6	225.1	0.038	0.045	0.022	0.028
	EBP Box-Cox	171.7	227.5	0.045	0.049	0.025	0.029
	EBP Syn	188.2	257.6	0.093	0.091	0.486	0.554
	EBP Log Syn	160.7	256.2	0.054	0.062	0.026	0.031
	EBP Log-shift Syn	159.4	266.0	0.041	0.054	0.022	0.028
	EBP Box-Cox Syn	168.5	263.2	0.051	0.059	0.025	0.029
	Direct	543.6	631.1	0.097	0.102	0.083	0.093
RB [%]	EBP	2.39	3.93	34.77	30.40	109.6	116.8
	EBP Log	2.96	2.52	12.54	11.85	3.89	3.16
	EBP Log-shift	0.93	0.95	6.49	5.57	0.08	-0.34
	EBP Box-Cox	1.98	1.66	11.18	10.06	2.32	1.45
	EBP Syn	2.79	4.68	34.45	30.74	110.1	114.9
	EBP Log Syn	1.84	2.98	16.65	13.89	3.89	3.16
	EBP Log-shift Syn	0.80	1.52	9.59	7.13	0.10	-0.42
	EBP Box-Cox Syn	1.41	2.22	14.67	11.90	2.35	1.45
	Direct	-0.13	-0.32	-0.35	-0.27	-7.92	-10.21
CV	EBP	0.082	0.098	0.262	0.254	0.534	0.512
	EBP Log	0.078	0.096	0.145	0.158	0.058	0.066
	EBP Log-shift	0.073	0.093	0.123	0.141	0.048	0.062
	EBP Box-Cox	0.076	0.095	0.137	0.152	0.056	0.063
	EBP Syn	0.088	0.106	0.260	0.254	0.530	0.530
	EBP Log Syn	0.072	0.105	0.174	0.184	0.058	0.066
	EBP Log-shift Syn	0.078	0.109	0.144	0.166	0.049	0.062
	EBP Box-Cox Syn	0.074	0.107	0.161	0.179	0.055	0.063
	Direct	0.239	0.274	0.291	0.362	0.203	0.246
67 Out-of-sample municipalities							
RMSE	EBP	210.6	224.4	0.073	0.079	0.846	109.3
	EBP Log	216.3	234.5	0.061	0.069	0.032	0.034
	EBP Log-shift	200.7	222.9	0.062	0.075	0.031	0.032
	EBP Box-Cox	212.6	229.7	0.060	0.070	0.032	0.033
RB [%]	EBP	11.28	10.46	-0.69	2.98	152.6	1236.0
	EBP Log	12.43	13.96	-5.27	-2.83	2.25	2.16
	EBP Log-shift	11.19	12.83	-9.86	-7.69	-0.21	-0.26
	EBP Box-Cox	11.91	13.49	-6.60	-4.24	1.09	1.23
CV	EBP	0.109	0.123	0.179	0.187	0.693	4.045
	EBP Log	0.112	0.128	0.146	0.167	0.071	0.074
	EBP Log-shift	0.107	0.123	0.166	0.189	0.068	0.071
	EBP Box-Cox	0.110	0.126	0.154	0.172	0.071	0.072

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, 2004, 2012) and sampling (Tillè 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. 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. The package also includes code for implementing the EBP approach of Molina and Rao (2010) we discussed in Section 3.2. An alternative code in R for computing EBP estimates that also includes an option for using the transformations discussed in the present paper, visualization and export of the results to Excel is proposed by Kreutzmann (2016). 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. Small area estimation from a Bayesian perspective is provided, among others, 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 an 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 additional code in SAS that allows for the computation of direct and indirect small area methods. For additional details and examples in SAS we refer the prospective 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 Conclusion

In this paper we propose a general framework for the production of SA statistics and attempt to illustrate the SAE process in practice. As part of this framework we have touched upon four inter-related topics namely, specification of the problem, model selection and testing and finally method evaluation. While much can be said for each of these four 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 four stages is in many ways the wisdom of applied statistics, which holds true also for SAE. The illustrations we have included in this paper offer some practical ways of keeping this balance. Our empirical illustrations show 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 also very crucial steps for improving estimation without the need for additional data sources. Finally, having access to 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 still be applied in most cases.

Nevertheless there are questions that remain unresolved and which feel we should 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 the 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

Tzavidis, Zhang, Luna Hernandez and Schmid 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 M. Templ (2013). Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software 54 (15)*, 1–25.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association 83 (401)*, 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 J. Hobert (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association 93 (441)*, 262–272.

Box, G. E. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B 26 (2)*, 211–252.

Brown, G., R. Chambers, P. Heady, and D. Heasman (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.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B 76 (1)*, 47–69.

Chambers, R., J. Chandra, and N. Tzavidis (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology 37 (2)*, 153–170.

Datta, G. and P. Lahiri (1995). Robust hierarchical bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis 54 (2)*, 310–328.

Datta, G. S., P. Hall, and A. Mandal (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 (493)*, 362–374.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87 (418)*, 376–382.

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

Elbers, C., J. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. *Econometrica 71 (1)*, 355–364.

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.

Ghosh, M. (1992). Constrained bayes estimation with applications. *Journal of the American Statistical Association 87 (418)*, 533–540.

Ghosh, M., T. Maiti, and A. Roy (2008). Influence functions and robust bayes and empirical bayes small area estimation. *Biometrika 95 (3)*, 573–585.

Graf, M., J. Marin, and I. Molina (2015). Estimation of poverty indicators in small areas under skewed distributions. In *Proceedings of the 60th World Statistics Congress of the International Statistical Institute*, The Hague, Netherlands.

Gurka, M. J., L. J. Edwards, K. E. Muller, and L. L. Kupper (2006). Extending the box–cox transformation to the linear mixed model. *Journal of the Royal Statistical Society Series A 169 (2)*, 273–288.

Hall, P. and T. Maiti (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B 68 (2)*, 221–238.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47 (260)*, 663–685.

Jiang, J., P. Lahiri, and S. Wan (2002). A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics 30 (6)*, 1782–1810.

Jiang, J. and T. Nguyen (2012). Small area estimation via heteroscedastic nested-error regression. *Canadian Journal of Statistics 40 (3)*, 588–603.

Kreutzmann, A.-K. (2016). Poverty mapping using small area estimation: An application with R. Master's thesis, Freie Universität Berlin.

Lohr, S. and J. Rao (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika 96 (2)*, 457–468.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software 9 (1)*, 1–19. R package version 2.2.

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

Meindl, B., M. Templ, A. Alfons, A. Kowarik, , 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 Y. Marhuenda (2015). sae: An R package for small area estimation. *The R Journal 7 (1)*, 81–98.

Molina, I. and J. N. K. Rao (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics 38 (3)*, 369–385.

Mukhopadhyay, P. and A. McDowell (2011). Small area estimation for survey data analysis using SAS software. SAS Global Forum 2011.

Opsomer, J., G. Claeskens, M. Ranalli, G. Kauermann, and F. Breidt (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society Series B 70 (1)*, 265–283.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science 28 (1)*, 40–68.

Pfeffermann, D. and S. Correa (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika 99 (2)*, 457–472.

Prasad, N. G. N. and J. N. K. Rao (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association 85 (409)*, 163–171.

Pratesi, M. and N. Salvati (2009). Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics 25 (1)*, 37–53.

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

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

Rivest, L.-P. and E. Belmonte (2000). A conditional mean squared error of small area estimators. *Survey Methodology 26*, 67–78.

Rojas-Perilla, N., T. Schmid, and N. Tzavidis (2015). A comparison of small area estimation methods for poverty mapping under box-cox type transformations. In *4th Italian Conference on Survey Methodology*, Rome, Italy.

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

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

Shen, W. and T. Louis (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society Series B 60 (2)*, 455–471.

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

Sinha, S. K. and J. N. K. Rao (2009). Robust small area estimation. *The Canadian Journal of Statistics 37 (3)*, 381–399.

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.

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

Tillè, Y. and A. Matei (2012). *sampling: Survey Sampling*. R package version 2.5.

Tzavidis, N., S. Marchetti, and R. Chambers (2010). Robust estimation of small area means and quantiles. *Australian and New Zealand Journal of Statistics 52 (2)*, 167–186.

Ugarte, M., T. Goicoa, A. Militino, and M. Durban (2009). Spline smoothing in small area trend estimation and forecasting. *Computational Statistics and Data Analysis 53 (10)*, 3616–3629.

Vaida, F. and S. Blanchard (2005). Conditional akaike information for mixed-effects models. *Biometrika 92 (2)*, 351–370.

Valliant, R., A. Dorfman, and R. Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

Warnholz, S. and T. Schmid (2016). Simulation tools for small area estimation: Introducing the R package saeSim. *Austrian Journal of Statistics 45*, 55–69.

Weidenhammer, B., N. Tzavidis, T. Schmid, and N. Salvati (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 (2)*, 223–237.

Zhang, L.-C. (2009). Estimates for small area compositions subjected to informative missing data. *Survey Methodology 35 (2)*, 191–201.

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2016 erschienen:

2016/1 BARTELS, Charlotte und Maximilian STOCKHAUSEN
Children's opportunities in Germany – An application using multidimensional measures
Economics

2016/2 BÖNKE, Timm; Daniel KEMPTNER und Holger LÜTHEN
Effectiveness of early retirement disincentives: individual welfare, distributional and fiscal implications
Economics

2016/3 NEIDHÖFER, Guido
Intergenerational Mobility and the Rise and Fall of Inequality: Lessons from Latin America
Economics

2016/4 TIEFENSEE, Anita und Christian WESTERMEIER
Intergenerational transfers and wealth in the Euro-area: The relevance of inheritances and gifts in absolute and relative terms
Economics

2016/5 BALDERMANN, Claudia; Nicola SALVATI und Timo SCHMID
Robust small area estimation under spatial non-stationarity
Economics

2016/6 GÖRLITZ, Katja und Marcus TAMM
Information, financial aid and training participation: Evidence from a randomized field experiment
Economics

2016/7 JÄGER, Jannik und Theocharis GRIGORIADIS
Soft Budget Constraints, European Central Banking and the Financial Crisis
Economics

2016/8 SCHREIBER, Sven und Miriam BEBLO
Leisure and Housing Consumption after Retirement: New Evidence on the Life-Cycle Hypothesis
Economics

2016/9 SCHMID, Timo; Fabian BRUCKSCHEN; Nicola SALVATI und Till ZBIRANSKI
Constructing socio-demographic indicators for National Statistical Institutes using mobile phone data: estimating literacy rates in Senegal
Economics

2016/10 JESSEN, Robin; ROSTAM-AFSCHAR, Davud und Sebastian SCHMITZ
How Important is Precautionary Labor Supply?
Economics

2016/11 BIER, Solveig; Martin GERSCH, Lauri WESSEL, Robert TOLKSDORF und Nina KNOLL
Elektronische Forschungsplattformen (EFP) für Verbundprojekte: Bedarfs-, Angebots- und Erfahrungsanalyse
Wirtschaftsinformatik

2016/12 WEIDENHAMMER, Beate; Timo SCHMID, Nicola SALVATI und Nikos TZAVIDIS
A Unit-level Quantile Nested Error Regression Model for Domain Prediction with Continuous and Discrete Outcomes
Economics