

University of Wollongong

## Research Online

---

Faculty of Engineering and Information  
Sciences - Papers: Part B

Faculty of Engineering and Information  
Sciences

---

2019

# A Comparison of Methods for Poverty Estimation in Developing Countries

Sumonkanti Das

*Shahjalal University of Science and Technology, University of Wollongong, Maastricht University,*  
sdas@uow.edu.au

Stephen Haslett

*Australian National University, Massey University*

Follow this and additional works at: <https://ro.uow.edu.au/eispapers1>



Part of the [Engineering Commons](#), and the [Science and Technology Studies Commons](#)

---

### Recommended Citation

Das, Sumonkanti and Haslett, Stephen, "A Comparison of Methods for Poverty Estimation in Developing Countries" (2019). *Faculty of Engineering and Information Sciences - Papers: Part B*. 2906.  
<https://ro.uow.edu.au/eispapers1/2906>

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: [research-pubs@uow.edu.au](mailto:research-pubs@uow.edu.au)

---

## A Comparison of Methods for Poverty Estimation in Developing Countries

### Abstract

Small area estimation is a widely used indirect estimation technique for micro-level geographic profiling. Three unit level small area estimation techniques-the ELL or World Bank method, empirical best prediction (EBP) and M-quantile (MQ) - can estimate micro-level Foster, Greer, & Thorbecke (FGT) indicators: poverty incidence, gap and severity using both unit level survey and census data. However, they use different assumptions. The effects of using model-based unit level census data reconstructed from cross-tabulations and having no cluster level contextual variables for models are discussed, as are effects of small area and cluster level heterogeneity. A simulation-based comparison of ELL, EBP and MQ uses a model-based reconstruction of 2000/2001 data from Bangladesh and compares bias and mean square error. A three-level ELL method is applied for comparison with the standard two-level ELL that lacks a small area level component. An important finding is that the larger number of small areas for which ELL has been able to produce sufficiently accurate estimates in comparison with EBP and MQ has been driven more by the type of census data available or utilised than by the model per se.

### Disciplines

Engineering | Science and Technology Studies

### Publication Details

Das, S. & Haslett, S. (2019). A Comparison of Methods for Poverty Estimation in Developing Countries. *International Statistical Review*, 87 (2), 368-392.

# A Comparison of Methods for Poverty Estimation in Developing Countries

Sumonkanti Das<sup>1,2,3</sup>  and Stephen Haslett<sup>4,5</sup>

<sup>1</sup>*Department of Statistics, Shahjalal University of Science and Technology, Sylhet, Bangladesh*

<sup>2</sup>*School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia*

<sup>3</sup>*Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands*

<sup>4</sup>*Statistical Consulting Unit & School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, Australia*

<sup>5</sup>*School of Fundamental Sciences & Centre for Public Health Research, Massey University, Palmerston North, New Zealand*

E-mail: [sd535@uowmail.edu.au](mailto:sd535@uowmail.edu.au)

## Summary

Small area estimation is a widely used indirect estimation technique for micro-level geographic profiling. Three unit level small area estimation techniques—the ELL or World Bank method, empirical best prediction (EBP) and M-quantile (MQ) — can estimate micro-level Foster, Greer, & Thorbecke (FGT) indicators: poverty incidence, gap and severity using both unit level survey and census data. However, they use different assumptions. The effects of using model-based unit level census data reconstructed from cross-tabulations and having no cluster level contextual variables for models are discussed, as are effects of small area and cluster level heterogeneity. A simulation-based comparison of ELL, EBP and MQ uses a model-based reconstruction of 2000/2001 data from Bangladesh and compares bias and mean square error. A three-level ELL method is applied for comparison with the standard two-level ELL that lacks a small area level component. An important finding is that the larger number of small areas for which ELL has been able to produce sufficiently accurate estimates in comparison with EBP and MQ has been driven more by the type of census data available or utilised than by the model per se.

*Key words:* empirical best prediction; ELL; M-quantile; small area estimation; unit record census data.

## 1 Background

Traditional national surveys are designed to obtain national and regional statistics. Micro-level administrative unit statistics are either not directly estimable (if they contain no sample) or have estimates with standard errors too large to be useful (because sample sizes are small).

Model-based methods for small area estimation (SAE) estimate local parameters with greater efficiency by borrowing strength from other ‘similar’ areas (Rao & Molina, 2015). Generally, the survey data include both the variable(s) of interest and the potentially related auxiliary variables. Whether or not census data is used in any form, model-based SAE methods use regression-type techniques and are usually classified into two broad groups based on data availability. Unit level SAE methods are used when suitable data are available at unit level;

otherwise, area level SAE methods are applied. A linear or non-linear mixed model then uses the survey data to obtain the small area statistics via the auxiliary information.

National and local administrative authorities place increasing importance on small area statistics for policy development and implementation. In the developing world, where available, aid agencies also use small area estimates of poverty-related indicators to focus their programmes. Basic model-based SAE methods estimate small area linear parameters such as means and totals and are not suitable for estimating non-linear functions such as quantiles and distribution functions. In poverty mapping studies based on equivalised expenditure or income measured at household (HH) level, the most used poverty indicators: poverty incidence (more properly called poverty prevalence), poverty gap (PG) and poverty severity (PS) are non-linear functions of the income or expenditure distribution function (Foster *et al.*, 1984). Modified SAE methods are required for estimating such non-linear poverty indicators.

Not all SAE methods do so, but the three methods discussed in this paper [World Bank method (ELL), empirical best prediction (EBP) and M-quantile (MQ)] use linear rather than non-linear models and, in addition to survey data, use auxiliary variables available as unit records (or model-based unit records) from a census and/or administrative database. The small area estimates are then based on aggregates of predictions from the survey model, of the variable of interest or some linear or non-linear function of it, for every census unit record. The principal advantages of incorporating unit record census data can include ability to model even non-linear functions of unit level responses using a linear model and considerably increased accuracy. Conditional on the model being correct, an adequate match between survey and census area codes and between the auxiliary variables' survey and census data, this in turn allows small areas to be much smaller for a given accuracy than when there is only survey data. However, actual unit record census data are not universally available, so a model-based unit record census dataset may need to be reconstructed from a limited number of census cross-tabulations that are assumed to provide sufficient statistics for the actual census. This same model-based census reconstruction is also commonly used in spatial microsimulation (see Haslett *et al.*, 2010). Consequences of using model-based unit record census data include (i) no continuous auxiliary variables are available for the modelling of the survey data, (ii) unknown additional error is introduced into the SAEs through an implicit and often untestable census model and (iii) (cf. when the actual unit record census data are available) it is not possible to use cluster level or sub-cluster averages as contextual variables to reduce the size of random effects because they are known only for survey clusters not for every cluster in the census. There is also an intermediate situation where actual unit record data are available so that not all auxiliary variables need to be categorical, but either the required match between survey and census areas at cluster (or sub-cluster) level is not available, or for some other reason, contextual variables are not used.

A set of six papers, all of which use EBP or MQ, illustrate the underlying data-dependent issues. Tzavidis *et al.* (2010), Giusti *et al.* (2012), Marchetti *et al.* (2012), Tzavidis *et al.* (2013), Marchetti *et al.* (2018) and Marhuenda *et al.* (2017) all use some form of unit record census data. Marhuenda *et al.* (2017) use EBP with categorical data and model-based unit record census data and consider a three-level model that includes variation at domain and subdomain level and seek small area estimates at both these levels. However, their subdomains are not survey clusters but are much larger and without actual unit record census data they cannot model cluster level variation. The other five papers all use some form of MQ and have available actual unit record census data. Marchetti *et al.* (2018) mention the possibility of using but do not use auxiliary variables at cluster level (i.e. contextual variables) and extend MQ to include cluster level as well as small area level random effects in their three-level MQ model. Giusti *et al.* (2012), Marchetti *et al.* (2012) and Tzavidis *et al.* (2013) do not explicitly mention contextual variables,

while Tzavidis *et al.* (2010) mention them but do not use them at cluster or sub-cluster level. Given the utility of fine level contextual variables for reducing the size of random effects and reducing small area root mean square error (RMSE) (as discussed in detail in Haslett, 2016), the central question is why they are not being used except in ELL-type models.

In answer, one reason contextual variables are so seldom used is that in some countries, there are often a relatively limited number of candidate or utilised auxiliary variables for small area models, none of which are aggregated at the required cluster or sub-cluster level for the entire census. To use such finely aggregated variables as contextual requires the clusters (or sub-clusters) used for survey design be known and are entirely matchable by area to the census data. However, survey and census often use a different frame, even if the survey and census are essentially contemporaneous, but the survey is conducted before or at around the same time as the census. Time differences between survey and census exacerbate this problem as well as raising issues concerning retention of model structure between survey and census. Recording or recovering all changes in area designations, while critical and time consuming for SAE where contextual variables are to be used, is a task of rather more limited importance for government statistical agencies usually focused on delivering survey and census results separately and not on these two data sources' fine level interconnection. Consequently, specifying clusters and sub-clusters so that census aggregates can be used in survey models as contextual variables can be very time consuming and is not always feasible, even if full actual unit record census data are available.

So whether the problem is census unit record data access or that there are major and perhaps insurmountable difficulties matching survey and census area codes at cluster or sub-cluster level, the underlying issue is data availability for either confidentiality or logistical reasons. We return to these important data-related issues later.

The SAE method for poverty mapping using both survey and census unit record data was first developed by Elbers, Lanjouw and Lanjouw (2002, 2003) and is also known colloquially as the World Bank method or the ELL method. The ELL method develops a two-level (cluster and HH) nested error regression model (Battese *et al.*, 1988) using survey data at HH level and uses the model to make equivalised income or expenditure predictions at HH level for all census observations. These are then aggregated to small area level. Mean squared error (MSE) is estimated via a bootstrap procedure where the fitted model, the estimated regression parameters plus their modelling error and the residuals from the survey-based regression are used multiple times as inputs. Due to its apparent simplicity, extensive data availability, and the PovMap software that is freely available from the World Bank (Zhao & Lanjouw, 2009), ELL has been implemented in many developing countries. Variants of and improvements to the method have been used by the World Bank and others, for example, in Thailand (Healy *et al.*, 2003), Cambodia (Fujii, 2004; Haslett *et al.*, 2013), South Africa (Alderman *et al.*, 2002) Brazil (Elbers *et al.*, 2004), Bangladesh (BBS & UNWFP, 2004; Haslett *et al.*, 2014), the Philippines (Haslett & Jones, 2005) and Nepal (Haslett & Jones, 2006; Haslett *et al.*, 2014) The ELL method has also been extended to estimating prevalence of child undernutrition at small area level (BBS & UNWFP, 2004; Haslett & Jones, 2006; Haslett *et al.*, 2013; Haslett, Jones & Isidro, 2014; Haslett *et al.*, 2014) using a model with an additional variance component at child level.

PovMap has now been extended (version 2.5: Van der Weide, 2014) to include the EBP method of Molina & Rao (2010) who used empirical Bayes or best prediction (EBP) under a finite population to estimate the Foster, Greer and Thorbecke (FGT) poverty measures (Foster *et al.*, 1984). Their method produces estimators with minimum MSE that are 'best predictors' through Monte Carlo (MC) approximation under the assumption that the transformed welfare

variable (usually log HH income or expenditure) follows a nested error regression model with normally distributed errors. The EBP method generates simulated values by making independent draws from the conditional distribution of unobserved values given the observed values and associated values of auxiliary variables. EBP is based on an area-specific two-level nested error regression model that includes small area level and HH level effects, while the ELL method is based on cluster-specific two-level model using cluster and HH variability with contextual variables from the census (intended to substitute for any non-zero mean random effect in every small area even the non-sampled ones). EBP has been used mostly in European Union Member States (EU-MS) using the Income and Living Conditions (EU-SILC 2012) survey datasets. Molina & Rao (2010) applied the EBP method to estimate poverty incidence and gap by sex for the 50 provinces in Spain.

Both ELL and EBP methods are based on standard linear random effects models with possibly strong distributional assumptions and formal specification of the random part. They may be non-robust to outliers in the response variable, although outliers used in the survey-based model do not necessarily imply outliers or non-robustness in the aggregated small area estimates.

An alternative approach to SAE, which like ELL and EBP uses unit record (or model-based unit record) census data, is the MQ method proposed by Chambers and Tzavidis (2006) and Tzavidis and Chambers (2007), which is based on quantiles of the conditional distribution of response variable given the covariates (Breckling & Chambers, 1988). Unlike ELL and EBP, MQ is distribution-free and automatically provides outlier robust inference and relaxes assumptions of random errors in the traditional unit level nested error regression model. MQ has been used in poverty mapping studies by Tzavidis *et al.* (2010) and been applied to estimate poverty indicators at local administrative level in European countries including Albania (Tzavidis *et al.*, 2008), Tuscany (Marchetti *et al.*, 2012), Italy (Giusti *et al.*, 2011) and Poland (Marchetti *et al.*, 2018)

M-quantile does not specify small area or cluster level variation explicitly because it puts aside boundary issues when developing the model. ELL and EBP can suffer from model misspecification if there is an untenable parametric specification of random errors or under the influence of marked outliers. However, because they usually aggregate tens of thousands of units if not more, unit level based SAE methods can be remarkably stable to changes in model and/or parameter estimates, providing important auxiliary variables are not omitted.

The three SAE methods for poverty mapping are based on different assumptions. No method is uniformly best. All can be criticised: ELL for assuming small area level homogeneity (even though it uses detailed area specific contextual variables from the census or GIS sources), EBP for assuming cluster homogeneity and MQ for assuming cluster homogeneity indirectly.

The aim of this article for developing world poverty studies is to compare and contrast, to the extent possible, these three unit level SAE methods all of which use unit level or model-based unit record census data. This includes numerical comparison via a simulation study and a model-based reconstruction of 2000 and 2001 Bangladesh data. Section 2 describes the FGT poverty indicators and the three unit level SAE methods (ELL, EBP and MQ); Section 3 compares these three SAE methods; Section 4 outlines construction, based on datasets from Bangladesh, of a simulated census dataset and the sample design used to select a sample from it and compares SAEs for the three methods. Section 5 discusses findings of this simulation study, and Section 6 concludes with a summary and discussion of statistical issues around selection of a SAE method for poverty mapping in developing countries.

## 2 Small Area Methodologies for Poverty Estimation

Consider a finite population of size  $N$  divided into  $D$  small domains or small areas of size  $N_1, N_2, \dots, N_D$  and let  $E_{ik}$  be the equivalised income or expenditure for individual  $k$  belong-

ing to  $i$ -th domain. If  $E_{ik}$  is less than a poverty line  $z$ , the  $k$ -th individual will be considered as below the ‘poverty line’. The Foster *et al.* (1984) FGT poverty measures for domain  $i$  are calculated as

$$F_{\alpha i} = N_i^{-1} \sum_{k=1}^{N_i} F_{\alpha ik}; i = 1, 2, \dots, D; \alpha = 0, 1, 2, \tag{1}$$

where  $F_{\alpha ik} = (1 - E_{ik}/z)^\alpha I(E_{ik} < z)$ , where  $I(\cdot)$  is an indicator function. The FGT poverty indicators  $F_{\alpha i}$  are then defined as poverty incidence (or head count ratio) if  $\alpha = 0$ , PG if  $\alpha = 1$  and PS if  $\alpha = 2$ . Poverty incidence is the proportion of the population with income/consumption level below the poverty line; PG measures the income/consumption expenditure shortfall for those below the poverty line; and PS, by squaring difference for those below the poverty line, places particular emphasis on the very poor.

When there is a random sample, each sample HH is assigned a sampling weight based on its probability of selection, non-response rate, calibration to some defined population distribution, trimming etc. HH level sampling weights need to be adjusted for a measure of HH size. So within small area  $i$ , the direct estimate of each FGT measure  $F_{\alpha i}$  is the weighted mean taken over all sampled HHs in small area  $i$  after adjusting each HH for its measure of size.

When there is no sample in a small area, synthetic estimators can be used to calculate poverty indices through model-based SAE methods based on the auxiliary information (Rao & Molina, 2015), but no small area-specific adjustments are then possible unless there are contextual variables (see Haslett, 2016). Often, a significant number of small areas are not covered in sample for ELL, but as compensation, there are contextual variables for every small area; for EBP and MQ without contextual variables, having sample in every small area is preferable so as to provide a sample-based direct area-specific adjustment (area-specific random effects in EBP and area-specific regression coefficients in MQ). If spatial information is available, for example, adjacency of all areas, partial adjustment is possible even for non-sampled areas based on the location information and areal smoothing.

ELL, EBP and MQ for poverty estimation are outlined in Sections 2.1 to 2.3.

### 2.1 The World Bank Method

In the World Bank poverty mapping methodology, a regression model is developed between the log-transformed response variable  $Y_{ijk} = \log(E_{ijk})$  and the explanatory variables ( $X_{ijk}$ ) available from a sample for an HH income or expenditure survey contemporaneous with the census. Here,  $y_{ijk}$  is the response for the  $k$ -th HH in the  $j$ -th cluster of the  $i$ -th area. Standard methods of fitting regression models cannot be used due to the complex sampling designs (e.g. stratification and/or clustering) for selecting the survey HHs. Moreover, because HHs tend to be clustered into villages or other small geographic or administrative units within small areas and are often relatively homogenous, it is common to have a cluster [primary sampling unit (PSU)] random effect in the model, along with the cluster level or sub-cluster level contextual variables available for all census clusters or sub-clusters that are part of the  $\{x_{ijk}\}$  in the survey model after area matching between survey and census. So a nested error linear regression model (Battese *et al.*, 1988) has HHs at level one and clusters at level two:

$$y_{ijk} = \mathbf{x}_{ijk}^T \boldsymbol{\beta}_{(2)} + u_{ij} + \varepsilon_{ijk} \quad i = 1, \dots, D \quad j = 1, 2, \dots, C_i \quad k = 1, 2, \dots, N_{ij} \tag{2}$$

$$u_{ij} \sim N(0, \sigma_{u(2)}^2) \quad \varepsilon_{ijk} \sim N(0, \sigma_{\varepsilon(2)}^2).$$

In the simplest form of ELL, cluster-specific and HH-specific errors  $u_{ij}$  and  $\varepsilon_{ijk}$  are assumed to follow a normal distribution with constant variance components, although generally, ELL

makes no distributional assumptions for these model errors. The sub-script 2 of  $\beta_{(2)}$ ,  $\sigma_{u(2)}^2$  and  $\sigma_{\varepsilon(2)}^2$  indicates a two-level model. The model can be extended to three-level hierarchical model if individual level data are available (say children within HH). Alternatively, a three-level ELL model could instead include a small area level error.

The FGT poverty measures are non-linear functions of  $y_{ijk}$ , generally the log expenditure or income, fitted on a per capita basis to the survey data using model (2) and then used for prediction of income or expenditure for every person in the census via back transformation. The income or expenditure predictions are then transformed to the three different FGT measures and aggregated to provide their small area estimates.

For standard error estimation conditional on the model being correct, ELL uses the bootstrap to generate  $L$  pseudo-censuses via the conditional distribution of  $y_{ijk}$ , within all  $N$  census HH units (not just the sampled ones) by adding non-parametrically simulated values of cluster-specific ( $u_{ij}^*$ ) and HH-specific errors ( $\varepsilon_{ijk}^*$ ) to each estimated fitted value ( $\mathbf{x}_{ijk}^T \hat{\beta}_{(2)}^*$ ) where  $\hat{\beta}_{(2)}^*$  is redrawn parametrically (once for each pseudo-census) from the distribution of  $\hat{\beta}_{(2)}$ . This gives a total of  $LN$  HH level predictions, each of which includes the contextual variables for the census cluster or sub-cluster to which the HH belongs.

Step 1: Fit model (2) to the sample data ( $\mathbf{y}_s, \mathbf{x}_s$ ) to obtain  $\hat{\beta}_{(2)}$ ,  $\hat{\mathbf{v}}(\hat{\beta}_{(2)})$ ,  $\hat{\sigma}_{u(2)}^2$  and  $\hat{\sigma}_{\varepsilon(2)}^2$ , allowing for the complex survey design. Some adjustments to the original ELL method are recommended (see, e.g. Haslett *et al.*, 2010).

Step 2: Generate the values of the regression parameters from the corresponding parametric distribution as  $\beta_{(2)}^* \sim N(\hat{\beta}_{(2)}, \hat{\mathbf{v}}(\hat{\beta}_{(2)}))$ .

Step 3: Generate the cluster-specific and HH-specific random errors independently and identically either parametrically or non-parametrically from the empirical distributions of the residuals (often ‘unshrunk’ to match the required variances  $\hat{\sigma}_{u(2)}^2$  and  $\hat{\sigma}_{\varepsilon(2)}^2$ ). Non-parametric methods are usually recommended, as usually the distribution of neither set of residuals is known.

Step 4: Generate  $L$  (say,  $L = 100$  to  $500$ ) pseudo-censuses each of size  $N$  with bootstrap income or expenditure  $\{y_{ijk}^{*(l)}; l = 1, 2, \dots, L\}$  from the survey model and the unit-record census data via the bootstrap superpopulation model  $y_{ijk}^* = \mathbf{x}_{ijk}^T \beta_{(2)}^* + u_{ij}^* + \varepsilon_{ijk}^*$ .

Step 5: Calculate the FGT poverty measures  $F_{\alpha i}^{*(l)}$  from each bootstrap population or pseudo-census for each small area. The ELL estimates with their MSE are calculated as

$$\hat{F}_{\alpha i}^{ELL} = L^{-1} \sum_{l=1}^L F_{\alpha i}^{*(l)} \text{ and } mse(\hat{F}_{\alpha i}^{ELL}) = L^{-1} \sum_{l=1}^L \left\{ F_{\alpha i}^{*(l)} - \hat{F}_{\alpha i}^{ELL} \right\}^2. \quad (3)$$

The HH-specific errors (sometimes called idiosyncratic errors) are allowed to be heteroskedastic in the ELL method. Elbers *et al.* (2003) provide a specific heteroskedasticity model. In practice, this heteroskedasticity model usually has an  $R^2$  of less than 5%, so in the simulation in Section 4, HH level errors are assumed homoskedastic.

In ELL, the basic idea is to increase the predictive power of the fitted regression model (often gauged by a high  $R^2$  value) and to reduce as much as possible the ratio of between-cluster variation to total variation  $\hat{\sigma}_{u(2)}^2 / (\hat{\sigma}_{u(2)}^2 + \hat{\sigma}_{\varepsilon(2)}^2)^{-1}$ , because it is the between-cluster variation that is principally responsible for high standard errors for the SAEs. Explanatory variables at different hierarchical levels such as HH, sub-cluster, cluster and small area can be considered



in the survey regression model. In practice, it is cluster level or sub-cluster level variables (included in the regression as contextual variables) and HH level variables (all known for every census observation) that are most useful. The principal criticisms of the ELL method are that its survey model contains no small area level random effect and that the model is synthetic because it contains no small area-specific adjustments. The counterclaim is that any non-zero small area-specific effects are controlled in the regression via contextual variables that are available at (sub-)cluster level from the census for all small areas, even the non-sampled ones. We will return to these issues later.

An ELL method based on three-level model (three-level ELL) that includes small area specific random effects instead of the standard two-level model (two-level ELL) can be constructed using a three-level version of the super-population model. The three-level model solves the problem of underestimated MSE if small area specific non-zero mean random effects exist even after using contextual variables. However, in applications small areas for ELL are often much smaller than for EBP or MQ because ELL routinely uses actual rather than model-based census unit record data and contextual variables. Because few small areas contain more than one sampled cluster, for ELL, it is very difficult to get unbiased and efficient estimates of any small area level variance components. In what follows, the ELL estimators based on two-level and three-level models are denoted respectively by ELL.2L and ELL.3L. ELL.3L differs from ELL.2L in having small area level random effects in the model.

### 2.2 The Empirical Best Prediction Method

In the EBP method, the small area or domain of interest (which is an aggregate of clusters) is considered to have a non-zero mean small area specific random effect in the model for the response variable. The two-level nested error regression model is

$$\begin{aligned}
 y_{ik} &= \mathbf{x}_{ik}^T \boldsymbol{\beta}_{(2)} + \eta_i + \varepsilon_{ik} \quad i = 1, \dots, D \quad k = 1, \dots, N_i \\
 \eta_i &\sim N\left(0, \sigma_{\eta(2)}^2\right) \quad \varepsilon_{ik} \sim N\left(0, \sigma_{\varepsilon(2)}^2\right),
 \end{aligned}
 \tag{4}$$

where  $\eta_i$  and  $\varepsilon_{ik}$  are small area-specific and HH-specific random errors, respectively. The FGT estimator is split into sample and non-sample parts as

$$F_{\alpha i} = N_i^{-1} \left[ \sum_{k \in s_d} F_{\alpha ik} + \sum_{k \in r_d} F_{\alpha ik} \right] \text{ for } \alpha = 0, 1, 2,$$

and the non-sample part is then predicted based on an MC approximation. The best predictor of  $F_{\alpha i}$  under the squared error loss is given by

$$\hat{F}_{\alpha i} = E_{y_{r_i}} [F_{\alpha i} | \mathbf{y}_{s_i}] = \int F_{\alpha i} f(\mathbf{y}_{r_i} | \mathbf{y}_{s_i}) d\mathbf{y}_{r_i},$$

where  $f(\mathbf{y}_{r_i} | \mathbf{y}_{s_i})$  is the joint density of  $\mathbf{y}_{r_i}$  (vector of non-sample  $y$  in area  $i$ ) given  $\mathbf{y}_{s_i}$  (vector of sample  $y$  in area  $i$ ). The basic procedure of EBP method to obtain the estimate of  $F_{\alpha i}$  is

Step 1: Fit the nested-error model (4) to the survey data  $(\mathbf{y}_s, \mathbf{x}_s)$ . For example, Molina & Rao (2010) use ML or REML for a simple random sample, and note that more generally for a complex survey, a ‘suitable estimator’ can be used to obtain  $\hat{\boldsymbol{\beta}}_{(2)} \hat{\sigma}_{\eta(2)}^2$  and  $\hat{\sigma}_{\varepsilon(2)}^2$ .

Step 2: Generate  $L$  independent realisations of  $\mathbf{y}_{r_i} \left\{ \mathbf{y}_{r_i}^{(l)}, l = 1, 2, \dots, L \right\}$  from the conditional distribution  $f(\mathbf{y}_{r_i} | \mathbf{y}_{s_i})$  assumed to follow normal distribution with mean  $\boldsymbol{\mu}_{i_{r|s}} = \mathbf{X}_{r_i} \boldsymbol{\beta}_{(2)} + \sigma_{u(2)}^2 \mathbf{1}_{N_i - n_i} \mathbf{1}_{n_i}^T \mathbf{V}_{si}^{-1} (\mathbf{y}_{s_i} - \mathbf{X}_{s_i} \boldsymbol{\beta}_{(2)})$  and exchangeable variance–covariance matrix within each small area  $i$ ,  $\mathbf{V}_{i_{r|s}} = \sigma_{u(2)}^2 (1 - \gamma_i) \mathbf{1}_{N_i - n_i} \mathbf{1}_{N_i - n_i}^T + \sigma_{\varepsilon(2)}^2 \mathbf{I}_{N_i - n_i}$  where  $\gamma_i^{-1} = \sigma_{u(2)}^{-2} (\sigma_{u(2)}^2 + n_i^{-1} \sigma_{\varepsilon(2)}^2)$ ,  $\mathbf{X}_{s_i}$  and  $\mathbf{X}_{r_i}$  are matrices of sample and non-sample explanatory variables in the  $i$ -th area. The non-sample values of  $\mathbf{X}$  are from a model-based census (i.e. a ‘census’ generated from a set of contingency tables, the margins of which are treated as sufficient statistics). The MC approximation is simplified via  $y_{ik}^{(l)} = \mathbf{x}_{ik}^T \hat{\boldsymbol{\beta}}_{(2)} + v_i^{*(l)} + \varepsilon_{ik}^{*(l)}$ ;  $ik \in r_i$  taking  $v_i^* \sim N(0, \hat{\sigma}_{u(2)}^2 (1 - \hat{\gamma}_i))$  and  $\varepsilon_{ik}^* \sim N(0, \hat{\sigma}_{\varepsilon(2)}^2)$  as independent zero correlation between small areas is assumed.

Step 3: Calculate  $F_{\alpha i}^{(l)}$  using the vector  $\mathbf{y}_i^{(l)} = \left\{ \mathbf{y}_{s_i}^T, \mathbf{y}_{r_i}^T \right\}$  and then average over  $L$  replicates to obtain the EBP estimates as  $\hat{F}_{\alpha i}^{EBP} \approx L^{-1} \sum_{l=1}^L F_{\alpha i}^{(l)}$ .

In EBP, the MSE estimator of  $\hat{F}_{\alpha d}^{EBP}$  is calculated separately via a parametric bootstrap method (González-Manteiga *et al.*, 2008). The basic steps are

Step 1: Obtain  $\hat{\boldsymbol{\beta}}_{(2)}$ ,  $\hat{\sigma}_{\eta(2)}^2$  and  $\hat{\sigma}_{\varepsilon(2)}^2$  following a suitable survey-based estimation method.

Step 2: Generate domain effects  $\eta_i^* \sim N(0, \hat{\sigma}_{\eta(2)}^2)$  and HH effects  $\varepsilon_{ik}^* \sim N(0, \hat{\sigma}_{\varepsilon(2)}^2)$  independently and then calculate bootstrap population income values  $y_{ik}^*$  via  $y_{ik}^* = \mathbf{x}_{ik}^T \hat{\boldsymbol{\beta}}_{(2)} + \eta_i^* + \varepsilon_{ik}^*$  for all model-based census HHs.

Step 3: Calculate domain-specific  $F_{\alpha i}^*$  by aggregating the bootstrap population values.

Step 4: Then refit the nested error model (4) with the bootstrap elements  $y_{ik}^*$  belonging to the sample and calculate the bootstrap model parameters denoted as  $\hat{\boldsymbol{\beta}}_{(2)}^*$ ,  $\hat{\sigma}_{\eta(2)}^{2*}$  and  $\hat{\sigma}_{\varepsilon(2)}^{2*}$ ; obtain the bootstrap EBP estimator of  $F_{\alpha i}^*$  denoted by  $\hat{F}_{\alpha i}^{*EBP}$  through MC approximation method.

Step 5: Repeat Steps 2–4 a large number of times, say,  $B = 1\,000$ .

Step 6: Let  $F_{\alpha i}^{*(b)}$  be the true value calculated at Step 3 and  $\hat{F}_{\alpha i}^{*EBP(b)}$  be the EBP estimate calculated at Step 5 in the  $b$ -th ( $b = 1, \dots, B$ ) bootstrap replicate; then the estimated MSE of  $\hat{F}_{\alpha i}^{EBP}$  is defined by  $mse(\hat{F}_{\alpha i}^{EBP}) = B^{-1} \sum_{b=1}^B \left[ F_{\alpha i}^{*(b)} - \hat{F}_{\alpha i}^{*EBP(b)} \right]^2$ .

Molina and Rao (2010) pointed out that, in the simplest case of estimating a small area mean, the EBP approach leads to  $E(\bar{y}_i^*) = N_i^{-1} \left( \sum_{k \in s_i} y_{ik} + \sum_{k \in r_i} \mathbf{x}_{ik}^T \boldsymbol{\beta}_{(2)} + \eta_i^* \right)$ , while the ELL

method provides a synthetic regression estimators  $E(\bar{y}_i^*) = N_i^{-1} \sum_{ijk \in s_i + r_i} \mathbf{x}_{ijk}^T \boldsymbol{\beta}_{(2)}$  because

for ELL,  $E(u_{ij}^*) = 0$  and  $E(\varepsilon_{ijk}^*) = 0$  even for a specific small area,  $i$ . Thus, unless ELL incorporates contextual variables, the ELL estimator will be less efficient than EBP estimator. However, if the number of small areas is large so that there are small areas with no sampled clusters, the EBP method will essentially reduce to a type of ELL method (Haslett, 2013). In reality, if a large number of small areas are not sampled in the survey then ultimately the EBP approach will produce synthetic estimates for these, like ELL, but without adequate contextual variables. This is why EBP requires sample data in every small area, but ELL does not

necessarily. ELL with unit level census data and matching between survey and census cluster (or sub-clusters) codes can consequently produce small area estimates with acceptable MSE for areas much smaller in size than EBP when EBP has only model-based census unit record data, because ELL can generate cluster or sub-cluster means as contextual variables for every small area not just the sampled ones. However, this is a twin-edged sword as ELL small area estimates require much greater care than often taken with survey model fitting, and there can be a temptation to make small areas much too small. The relative merits of EBP and ELL are not however universal. For example, the simulation in Molina and Rao (2010) uses sample and census sizes that are very much smaller than in developing world application, and unit record census data that being model-based lacks contextual variables at cluster (or even small area) level. Then, as Molina and Rao (2010) note, ELL estimates are entirely synthetic and biased, and EBP outperforms ELL.

### 2.3 The M-quantile Method

The MQ of order  $q$  for the conditional distribution of a random variable  $y$  given a vector of  $P$  covariates  $\mathbf{x}$ , is denoted by  $Q_q(y|\mathbf{x}; \psi)$  and defined as the solution of the estimating equation  $\int \psi_q[\varphi^{-1}(y - Q_q(y|\mathbf{x}; \psi))] f(y|\mathbf{x}) dy = 0$  where  $\psi_q$  is an asymmetric influence function. Suppose  $(\mathbf{x}_k, y_k), k = 1, \dots, n$  denotes the observed sample values of  $(\mathbf{x}, y)$  for individual  $k$ . Then a linear MQ regression model for  $y_k$  given  $\mathbf{x}_k$  assumes that the  $q$ -th MQ satisfies  $Q_q(y|\mathbf{x}, \psi) = \mathbf{x}^T \boldsymbol{\beta}_\psi(q)$ , where the  $k$ -th element of  $\mathbf{y}$  is  $y_k$  and the  $k$ -th row of  $\mathbf{x}$  is  $\mathbf{x}_k$ .

For a specified  $q$  and continuous  $\psi$ , the MQ regression parameters  $\boldsymbol{\beta}_\psi(q)$  are estimated by solving the estimating equations  $\sum_{k=1}^n \psi_q[\varphi^{-1}\{r_k(q)\}] \mathbf{x}_k = \mathbf{0}$  where  $r_k(q) = y_k - \mathbf{x}_k^t \hat{\boldsymbol{\beta}}_\psi(q)$ ;  $\psi_q(r_k(q)) = 2\psi(\varphi^{-1}r_k(q))\{q * I(r_k(q) > 0) + (1 - q) * I(r_k(q) \leq 0)\}$ ; and  $\varphi$  is a suitable robust estimate of the scale of the residuals (say, mean absolute deviation estimate, i.e.  $\varphi = (0.6745)^{-1} * \text{median}|r_k(q)|$ ). Here,  $\psi$  is an appropriate influence function such as Huber function  $\psi(u) = u * I(-c \leq u \leq c) + c * \text{sgn}(u) * I(|u| > c)$  used with tuning constant  $c = 1.345$  by Tzavidis et al. (2010). Straightforward application of iteratively reweighted least squares provides the solution of the estimating equations.

The basic idea of SAE with the MQ model is that, in the mixed effects model, the variability in the conditional distribution of  $y$  given  $\mathbf{x}$  does not depend on the pre-defined hierarchical structure, but rather is characterised by the MQ coefficients of the population units. If there are small area effects, units in the same cluster will tend to be similar, and this will be reflected in the quantiles the units belong to. Similar data within clusters will result in such clusters having a similar mixture of quantiles. This is how MQ captures area level random effects without explicitly modelling them. The MQ coefficient  $q_k$  for the population unit  $k$  with values  $y_k$  and  $\mathbf{x}_k$  is obtained via  $Q_{q_k}(\mathbf{x}_{ik}; \psi) = y_{ik}$ . When the conditional MQs are assumed to follow a linear model,  $Q_q(\mathbf{x}; \psi) = \mathbf{x}^T \boldsymbol{\beta}_\psi(q)$ , Tzavidis et al. (2010) proposed a bias corrected MQ predictor of area-specific mean ( $\mu_i$ ) based on smearing method of Chambers and Dunstan (1986):

$$\hat{\mu}_i^{MQ.CD} = N_i^{-1} \left( \sum_{k \in S_i} y_{ik} + \sum_{k' \in r_i} \mathbf{x}_{ik'}^t \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) + \frac{N_i - n_i}{n_i} \sum_{k \in S_i} (y_{ik} - \hat{y}_{ik}) \right), \quad (5)$$

where  $\hat{\theta}_i$  is the estimator of MQ coefficient ( $\theta_i$ ) for area  $i$  calculated by averaging the MQ coefficients of the units in area  $i$ ,  $\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)$  is the area-specific regression coefficients and  $\hat{y}_{ik} = \mathbf{x}_{ik}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)$  is an appropriate linear combination of the auxiliary variables.

Straightforward estimation of FGT poverty indicators (1) is not possible in MQ method. Like EBP, estimation of these non-linear parameters corresponds to the problem of estimating the out-of-sample observations. Thus, MQ estimator of FGT indicators can be written as

$$\hat{F}_{\alpha i}^{MQ} = N_d^{-1} \left[ \sum_{k \in s_i} F_{\alpha ik} + \sum_{k' \in r_i} \hat{F}_{\alpha ik'} \right] \quad i = 1, 2, \dots, D \quad \alpha = 0, 1, 2, \quad (6)$$

where  $\hat{F}_{\alpha ik'} = n_i^{-1} \sum_{k \in s_i} \left( \frac{z - \hat{E}_{ikk'}}{z} \right)^\alpha I(\hat{E}_{ikk'} \leq z)$ , and because the model is based on log per capita income or expenditure, the estimate of per capita expenditure for the  $k$ -th non-sampled HH is  $\hat{E}_{ikk'} = \exp \left[ \mathbf{x}_{ik'}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) + e_{ik} \right]$  where  $e_{ik}$  is a sample residual defined by the MQ fitted model to  $(y_s, x_s)$ . Marchetti *et al.* (2012) proposed an alternative procedure to calculate (6) following an MC simulation approach parallel to the EBP approach, and Marchetti *et al.* (2018) have extended the usual two-level MQ model to a three-level model.

The MSE of  $\hat{\mu}_i^{MQ.CD}$  can be estimated analytically (Chambers *et al.*, 2011), but the procedure can be unstable when small area sample sizes are very small (Marchetti *et al.*, 2012). A non-parametric bootstrap was proposed instead by Marchetti *et al.* (2012) to estimate MSE for small area means, poverty indicators and quantiles. The method uses an MQ small area model to estimate the target small area parameter, and a bootstrap that may smooth the empirical distribution function for the residuals conditioned on small area. The bootstrap is used to generate pseudo-censuses, and bootstrap estimates of the parameters are estimated from each of these. Because there is more than one pseudo-census, this procedure permits estimation of both bias and variance, and hence MSE, but it is very computationally intensive.

### 3 Comparison of World Bank, Empirical Best Prediction and M-quantile Methods

The three census unit level SAE methods of poverty estimation use different assumptions and so work better in different situations. None of the three methods is uniformly better. All have benefits and downsides. Which method is optimal or even usable depends on the available data. A comparison of the technical aspects of SAE methods is given in Table 1. There ‘standard implementation’ indicates the situation where all the level-specific residuals are assumed homoskedastic. More generally, ELL uses either homoskedastic or heteroskedastic HH level random errors. MQ is also categorised as a two-level rather than a three-level model (cf. Marchetti *et al.*, 2018).

All three methods, ELL, EBP and MQ, also require unit level survey data used to create a model to predict the outcome variable at unit-record level. ELL uses contextual variables at cluster or sub-cluster level as a standard and actual unit record census data. EBP usually uses model-based unit record data based on categorical variables only so does not have available any contextual variables. More often than for EBP, in applications MQ has unit record census data but, possibly, because of lack of access to cluster or sub-cluster code matching of survey and census, has to the authors’ knowledge, never used contextual variables at cluster or sub-cluster level.

This distinction is driven not by the methods per se but by data availability. The complexity and utility of the survey model and hence the accuracy of the SAEs or the number of small areas for EBP and MQ would be greatly enhanced if there were actual unit record census data, more variables that appear in both survey and census as candidate variables for modelling and contextual variables. Contextual variables, as used in ELL at cluster level, if used in EBP and

Table 1. Comparison of standard implementation for SAE of ELL, EBP and MQ.

Requirements and assumptions	ELL-2	ELL-3	EBP	MQ
Unit level survey data	Yes	Yes	Yes	Yes
Model-based unit level census data	No	No	Yes	Yes
Actual unit level census data	Yes	Yes	No	No
Model-based census data created from available cross-tabulations and an implicit log-linear model	No	No	Yes	Yes
Matching of survey and census households based on implicit census log-linear model	NA	NA	Yes	Yes
Testing of implicit census log-linear model	NA	NA	No	No
Complete matching of survey and census clusters	Yes	Yes	No	No
Use of contextual variables available for all census clusters in survey model and SAEs	Yes	Yes	No	No
Small area-specific random effects	No	Yes	Yes	No
Cluster-level random effects	Yes	Yes	No	No
Household-level random effects	Yes	Yes	Yes	No
Non-zero mean random effects conditional on each small area	No	No	Yes	No
Sample required in every small area for small area-specific adjustment	No	No	Yes	Yes
Distributional assumptions in survey model	No	No	Yes	No
Transformation to normality of survey data and model residuals required	No	No	Yes	No
Sub-setting of survey data with separate models for each subset	No	No	No	No
Influence of household level outliers in survey model fit	Limited	Limited	Limited	No
Influence of household level outliers in SAEs	Low	Low	Low	No
Influence on survey model fit of outliers at cluster level or for some cell in model-based census table	High	High	High	No
Influence of cluster level outliers on SAEs	Low	Low	Low	No
SAE RMSEs include estimation error in survey model regression parameters	Yes	Yes	No	No
Parametric bootstrap for RMSE of survey model regression parameter estimates	Yes	Yes	No	No
Error in estimation of survey parameter estimates included in RMSE for the SAEs	Yes	Yes	No	No
Parametric bootstrap for RMSE of random error components and hence SAEs	No	No	Yes	No
Non-parametric bootstrap for RMSE of random error components	Yes	Yes	No	Yes
Sensitivity of SAEs to survey model misspecification	High	High	Moderate	Moderate
Computational demand	Moderate	Moderate	High	High

Note 'Standard implementation' indicates homoskedasticity of all the level-specific residuals and a two-level model for MQ. EBP, empirical best prediction; ELL, World Bank method; MQ, M-quantile; RMSE, root mean square error; SAE, small area estimation.

MQ survey-based models could make error variances at small area and cluster level smaller; the use of contextual variables would also allow the possibility that the small area effects in the survey model are zero mean even when conditioned on each small area. For EBP, estimation of a non-zero mean random effect for every small area is instead required, which is why EBP requires every small area to be sampled. The population size of the small areas from EBP using model-based unit record census data is consequently large, as it is for MQ with actual unit record census data but without contextual variables even if MQ is three-level rather than

two-level, when EBP and MQ are compared with ELL using actual unit record census data and contextual variables. So ELL can produce accurate estimates for a greater number of small areas than EBP and MQ, but this is data rather than model driven. More developed countries, where EBP and MQ have been most used, are more likely to necessitate the use of model-based census data. In comparison, ELL has frequently been used in under-developed countries for which unit record census data have been made available along with the linkage between all the survey and census area codes, under strong confidentiality agreements.

Model-based unit record census data has an additional disadvantage over actual unit record census data even without contextual variables. The margins used in a model-based census, either as measured variables or their interactions, are from a limited set of cross-tabulations and so are necessarily categorical. They provide sufficient statistics for an implicit log-linear model, one which can only be tested using the limited number of tabulations available as margins. The risk then without actual unit record census data is untestable, omitted variable bias in both the model-based census data and the survey model, driven by limited data availability usually for confidentiality reasons.

Although ELL can provide finer level SAEs than EBP and MQ due to data availability, there are aspects of SAE modelling using ELL that can be difficult. Most are due indirectly to the richer data sources and the model sensitivities this can introduce.

Firstly, ELL requires matching every survey cluster or sub-clusters with the identical area in the census, and all census clusters or sub-clusters to be defined as they were at the time of the survey for using the contextual variables in both modelling and prediction. If the survey has been conducted using the previous census frame and there have been administrative area updates for the new census, area code matching can be very time consuming. The theoretical development of EBP and MQ methods instead also requires matching of census and survey HHs which is impossible in most real-world applications. The sampling ratios are very small however so that such matching for EBP and MQ is functionally redundant.

Secondly, because extensive unit record census data are used in ELL rather than model-based census data, the number of variables available for the survey model can be large. This plethora makes choice of a good model (including contextual variables) a non-trivial task. For example, if all possible two-way and three-way interactions are included in the survey model, even as a starting point, then the number of parameters can be close to the number of survey observations. This drives up  $R^2$ , the proportion of variance explained by the fitted model, and can too easily lead to model overfitting. Sub-setting of survey data into regions and fitting separate models to each (Fujii, 2004; cf. Haslett *et al.*, 2013) can further complicate this problem and, even if  $R^2$  is high, can lead to disjointed small area estimates, and non-smooth small area estimates at the boundary between regions that use different subset survey models. This is one reason that spatial models, which smooth any type of SAEs over small area based on proximity, need to be used with caution as they can mask even major underlying model misspecification. Whether  $R^2$  is an appropriate or useful measure for a mixed model also needs to be considered. Fitting MQ models that require parameter estimates for a given predictor variable to be ordered by increasing quantile may be beneficial. However, different proportions of the quantiles are usually found in different small areas, so MQ SAEs are much more stable; each comes from a mixture of subset models rather than from only one.

Thirdly, for ELL in particular (although also for EBP and MQ), a question in census and survey needs to be the same in the language used to administer the question, and yield averages that are not statistically different. To achieve this, variables can be transformed by a categorisation or re-categorisation within the question. Not being statistically different is required unless the variable is a cluster or sub-cluster census mean used as a contextual variable.

Empirical best prediction assumes the variable being modelled is normally distributed or can be transformed to normality. It also has little choice, given it has only model-based census unit record data, but to include a non-zero mean normally distributed random effect for each small area level based on the sampled observations. ELL does not actually require normality but has been criticised for not including a similar small area level random effect in the model (ELL.2L). This could be a very strong criticism, because it could otherwise bias SAEs and underestimate their standard error, but ELL also uses contextual variables available for every census HH based on the cluster to which it belongs.

All three methods ignore either cluster-variability and/or area-variability if a three-level superpopulation model holds. The MQ method is free from distributional assumptions for random effects; however, small area-specific MQ coefficients behave similarly to area-specific random effects. The ELL method uses contextual variables to control for or eliminate small area level random variation but may nevertheless give biased SAEs with underestimated standard errors if a non-negligible amount of between-area variation remains in the distribution of the response variable even after inclusion of contextual variables. A similar problem is also expected in both EBP and MQ methods. In practice, for ELL, this is difficult to check from national sample surveys because of its larger number of small areas than EBP and MQ. Where this has been checked to the extent possible (Haslett & Jones, 2005, 2006; Haslett *et al.*, 2013; Haslett, Jones & Isidro, 2014; Haslett *et al.*, 2014), the small area variance has been found to be negligibly small. Nevertheless, in the simulation study of Section 4, non-negligible small area effects are considered explicitly. Generally, design of national sample surveys with SAE in mind would help (Haslett, 2012). The problem is not completely avoided by EBP and MQ (although at cluster rather than small area level); cluster membership of the model-based census HHs is not known from census cross-tabulations, so cluster level random effects and variances are not estimable for all of the census clusters and hence cannot be included in the SAEs.

In developing countries like Bangladesh, the between-cluster variation in explanatory variables in the survey regression is often significantly higher than the between-area variation. When the cluster level is ignored in the multilevel survey model specification, the cluster level variance component merges with both the individual and small area level variance components (Tranmer & Steel, 2001). In such situations, for EBP, the individual level and small area level variance components will mislead about the distribution of the corresponding random errors. In a similar manner, if the small area level random effect is found significant but ignored in ELL method, the small area level variance component will be merged with cluster variance component, and MSE of the SAE will be underestimated, possibly severely underestimated (Das & Chambers, 2017).

Geographical specification of area and cluster are not needed to develop the MQ regression model. Instead of area-specific random effect ( $\eta_i$ ), the area-specific MQ coefficient ( $\theta_i$ ) is calculated by taking the average of MQ coefficient of individual units ( $q_k$ ) belonging to the small area and then small area-specific MQ regression coefficients  $\beta_\psi$  ( $\theta_i$ ) are estimated. The small area-specific regression coefficients  $\beta_\psi$  ( $\theta_i$ ) are then used to predict response values for the individuals belonging to the area. In this sense, the MQ method also incorporates a form of the between-area variation and ignores the between-cluster variation, except when a three-level model is applied as in Marchetti *et al.* (2018).

The EBP method provides efficient estimates based on nested-error regression model under the normality assumption for the target variable. For implementation, a correct transformation of the skewed expenditure or income variable is needed to induce normality of the random errors. Log transformations have been used, although not exclusively. Diallo and Rao (2014) instead consider a skew-normal distribution for the random errors. Both the ELL and MQ methods are formally free from requiring transformation because they do not assume normally

distributed errors. This is why generalised least squares estimation is often used instead of ML or REML in ELL (see Haslett *et al.*, 2010).

Compared with ELL, computational demand for estimation of RMSE in EBP and MQ can be very high, even for a relatively small number of small areas. If EBP and MQ were applied to the unit record data from developing countries with appropriate model adjustments, rather than to model-based unit record census data, the implementation of EBP and MQ method would be even more time consuming and expensive. To reduce the computational demand, the computation can be performed separately by dividing both sample and population into several large regions. Ferretti-and-Molina (2012) suggest such a fast algorithm for EBP.

The World Bank method has rather lower computational burden, but for census unit record data, there can still be tens or even hundreds of millions of HHs, each requiring multiple predictions via the bootstrap for RMSE estimation.

Even though EBP has normality assumptions that are not always met by real-world data, relative to the EBP and MQ, it is for ELL that sensitivity of SAEs to survey model misspecification can be particularly high. Reasons have been given previously, and the risks need to be balanced against the benefits of using ELL even when suitable data are available. When using the World Bank's PovMap software for ELL, short courses to relatively untrained staff are seldom an adequate basis for producing sound estimates with appropriate standard or mean square errors. Rather more than skill in PovMap's programming syntax is required.

It is only possible to examine differences among the three methods when both between-cluster and between-area variations exist in the distribution of response variable and the same data are used. Section 4 provides such a comparison, but as for earlier studies, the benefits of one method over another nevertheless remain principally dependent on whether the census data used is model-based unit record census data or actual census unit records.

There have not been previous studies to compare the three methods simultaneously. Molina and Rao (2010) compared ELL and EBP methods through a simulation study with model-based unit record census data, no contextual variables, very small census and sample size, with only area-specific random effects in the population model, very poor predictive power of the regression model and small areas that are large relative to the entire area and population. They indicated that ELL method may provide a poorer result than the direct estimator in such situations. This is however not a situation in which ELL is actually used. Betti *et al.* (2007) conducted a simulation experiment to compare ELL and MQ methods utilising two real datasets: the Living Standards Measurement Study 2002 and the Population Census 2001 of Albania. Using the survey dataset, at first, they developed a two-level regression model considering HH at 1st level and cluster at 2nd level and then generated a simulated population by drawing the regression parameters from their sampling distribution and resampling the level-specific errors from the corresponding estimated residuals. The census population is consequently not actual unit record census data, and contextual variables specific to each population cluster were not available. They found that ELL method produced more biased estimates than MQ, but the ranking of the poverty estimates was not markedly different even given the bias. They also noted that, as for EBP, there was a problem in MQ estimates for small areas (districts) not covered in the survey.

Tzavidis *et al.* (2013) explicitly discussed various technical issues of SAE methods for poverty mapping and conducted an empirical study as well as an application to Tuscany poverty data to compare ELL and MQ methods. Note that unit record census data were model-based without contextual variables. They found that MQ method is more efficient than the ELL method and that sometimes ELL provides a poorer outcome than the direct estimates, as also found by Molina and Rao (2010). However, as with the Molina and Rao (2010) simulation, the ELL method when used in the developing world for poverty mapping is not based on this frame-



work. The main reason for the Tzavidis *et al.* (2013) results is the purely synthetic behaviour of ELL method when there are no contextual variables available for every population cluster and there are small areas that include no sample. Souza *et al.* (2015) compare the ELL and EBP methods using 2010 Census data of Minas Gerais state of Brazil where, unusually, information on the target variable (per capita HH income) is collected. They conducted a simulation study by selecting 400 samples following the design of the 2008–2009 Consumer Expenditure Survey from the actual census population and then applying ELL and EBP methods to estimate poverty incidence and PG at municipality level. In the survey dataset, only 20% of municipalities (195 out of 853) were available. They observed that the ELL estimator performed better than the EBP estimator in terms of relative bias (RB) and relative root MSE (RRMSE).

Recently, Marchetti *et al.* (2018) compared the two-level and three-level MQ models with EBP based on data from Poland that included a 2011 Census 20% subsample as unit record census data, and the 2011 European Union Survey on Income and Living Conditions (EU-SILC) with  $n = 12\,871$  households. They found that for districts with sample size less than 15, EBP performed a little better in terms of coefficient of variation, but for all other sample size classes, MQ tended to perform better. They noted that their conclusions are data specific. Marchetti *et al.* (2018) also compared the two-level and three-level MQ models (MQ2 and MQ3, respectively) via simulation and found ‘an overall good performance of MQ3 versus MQ2’.

#### 4 An Empirical-based Simulation Study

Empirical best prediction and MQ have generally used model-based unit record census data. ELL usually has actual unit record census data with contextual variables at cluster level. So, it is difficult to compare the three methods directly. Some data standardisation is necessary. Further, if census data are simulated as here, then the generating model almost inevitably favour one method over another. With these provisos, the model-based simulation study below compares the three SAE methods for poverty under a scenario as similar as practical to that used in developing countries. The first poverty mapping study in Bangladesh was conducted by BBS & UNWFP (2004) using the Household Income and Expenditure Survey (HIES) 2000 (BBS, 2003) and a 5% sub-sample of unit record data from the Bangladesh Population and Housing Census 2001. The simulation is based on these two datasets. The available census data are unusual in being sub-sampled; this required a model be used to simulate both census and survey. A three-level linear model with small area, cluster and HH random effects, using parameters estimated from HIES 2000 was used to simulate the census. This favours ELL.3L and disadvantages ELL.2L, EBP and MQ. Possibly due to using a different fitting techniques, small area variance was estimated from HIES 2000 at about half the cluster level variance, a rather higher ratio than in Cambodia (Haslett *et al.*, 2013), Bangladesh (Haslett *et al.*, 2014), the Philippines (Haslett & Jones, 2005) or Nepal (Haslett *et al.*, 2014). EBP requires sample in every small area, so only small areas containing sample are used. Nevertheless, the simulation structure used provides some valuable insights.

Table 2. Summary statistics of administrative units in Census 2001 and HIES 2000.

	Division	District	Sub-district	EA	Household	Population ('000)
Census 2001	5	64	507	59 990	25 362 321	130 523
5% Census 2001	5	64	507	12 908	1 258 240	6 156
HIES 2000	5	63	295	442	7 428	38

Source: BBS and UNWFP (2004). EA, enumeration area; HIES, Household Income and Expenditure Survey.

Table 3. Summary statistics of EAs by sub-districts in Census 2001 and HIES 2000.

	Number	Mean	SD	Min	Max
5% Census 2001	12 908	25.5	12.9	3 (0.02%)	79 (0.007%)
HIES 2000	442	1.5	1.9	1 (75.3%)	10 (0.03%)

Source: BBS and UNWFP (2004). EA: enumeration area; HIES, Household Income and Expenditure Survey.

#### 4.1 Structure of Census and Survey Datasets

Sample design for HIES 2000 used the 1991 Census as a frame and divided Bangladesh into five divisions (*Barisal, Chittagong, Dhaka, Khulna, and Rajshahi*), 64 districts (*Zila*) and 507 sub-districts (*Upzila*). For 2001 Census, *Sylhet* division was created from sub-districts of the *Chittagong* division. HIES 2000 structure has been maintained in the simulation. BBS initially selected 5% of enumeration areas (EAs) from each sub-district (small domain) for analysis, ostensibly by systematic sampling (BBS & UNWFP, 2004) although mapping of the sampled HHs indicates clustering. Nevertheless, for confidentiality reasons, the 5% census data that is the basis of the simulation cover all Bangladesh administrative units down to sub-district. HIES 2000 covers 63 districts and 295 sub-districts (Table 2). Table 3 indicates that there are at least three EAs per sub-district in the 5% census dataset but more than 75% of sub-districts have single EA in the survey dataset. In the census and survey simulation, only the 295 sub-districts sampled in HIES 2000 are considered for comparison purposes to ensure all small areas are sampled. The EAs are PSUs in HIES 2000 and hence the clusters in our simulation.

##### 4.1.1 Sampling design of Household Income and Expenditure Survey 2000

For HIES 2000, the 14 strata were based on the residential classification of PSUs: urban, rural and statistical metropolitan area. Two-stage sampling was used to select 442 PSUs from 14 strata by stratified random sampling at the first stage and then 10 or 20 HHs were drawn from the selected PSUs by systematic sampling. The number of HHs sampled in the selected PSUs ranges from 190 to 310 for rural and urban areas and 100 to 287 for statistical metropolitan areas. Twenty HHs were selected in rural and urban areas and 10 HHs from SMAs. A total of 7 440 HHs were sampled (only 12 were excluded as missing or incomplete). The HIES PSU classification differed from that of Census 2001, so detailed area matching at cluster level was required (Table S1).

##### 4.1.2 Sampling design for simulation study

In poverty analysis in developing countries, matching between census HHs and survey HHs is not feasible. Even matching at PSU/cluster (or sub-cluster) level can be very time consuming. Nevertheless, despite the benefit being slight in practice because survey size is usually several orders of magnitude smaller than the census, matching between survey and census units is maintained in the simulation. To ensure that all small areas contain sample, only 295 of 507 sub-districts are used, and these are partitioned into urban PSUs, rural PSUs and statistical metropolitan area PSUs. For each simulated sample, PSUs were selected randomly by geographic location, and HHs were selected as for HIES 2000. Some HIES 2000 PSUs were merged with neighbours so PSUs contained at least 40 HHs. Table S2 shows how the PSUs of HIES 2000 were selected over the country. Thus, even though contextual variables are used in the simulation, with this additional aggregation, they are somewhat less effective at describing specific effects for small areas than in BBS and UNWFP (2004).

Similarly to HIES 2000, 442 PSUs were selected with probability proportional to the number of HHs and then HHs were selected randomly from the selected PSUs. The sampling ratio for sub-districts ranges from about 0.28% to 5.15% with mean 0.98% and median 0.83%. Data on HH expenditure are collected in the real survey but not in the real census (which is the reason SAE is necessary). However, in the simulation, HH expenditure is simulated for all census HHs in the 295 sub-districts to allow direct comparison with the census for each of the SAE methods.

#### 4.1.3 Construction of simulated census data

Simulated censuses were constructed using a model fitted to the actual 2000 HIES data plus contextual variables at sub-district rather than (sub-)cluster or (sub-)PSU level from the 2001 5% census. The model is given in (7). This simulated census data should be distinguished from the pseudo-census data or its equivalent generated for every such census population to estimate MSE for the SAEs; this is required even if real, complete census data were available.

In the BBS and UNWFP (2004) study, 30 explanatory variables (including separate categories for categorical variables and two-way interactions) were used at HH or sub-district level (Table S3) in a robust regression model that accounts for complex survey design and gives consistent covariance matrix estimates. To simplify simulation, multilevel analysis, which includes hierarchical structure in the survey regression model, has instead been used to fit survey models for log-transformed per capita consumption expenditure to develop two-level and three-level survey-based models.

Summary results for the multilevel models are given in Table 4. Null models, those without auxiliary regression variables, indicate the significant contribution of PSUs ( $\sigma_u^2$ ) and sub-districts ( $\sigma_\eta^2$ ) in variation of response variable. When auxiliary variables are added to the null model, both between-cluster and between-area variations are reduced. Although the variance component at sub-district level is very small, it is found still significant after including 30 explanatory variables; 75% of sub-districts have only single cluster that complicates estimation. The small but significant small area variance component may reflect the use of contextual variables at sub-district rather than at the less aggregated (sub-)cluster level (due to limited area code information) and different survey model fitting techniques. Both marginal and conditional  $R^2$  values (Nakagawa & Schielzeth, 2013) are slightly higher in the three-level model. So a three-level model has been used to generate log-transformed per capita HH expenditure for each simulated census is

$$y_{ijk} = \mathbf{x}_{ijk}^T \boldsymbol{\beta}_{(3)} + \eta_i + u_{ij} + e_{ijk}; \boldsymbol{\beta}_{(3)} \sim MN \left( \hat{\boldsymbol{\beta}}_{(3)}, \hat{\mathbf{v}} \left( \hat{\boldsymbol{\beta}}_{(3)} \right) \right) \quad (7)$$

$$\eta_i \sim N(0, 0.0086), u_{ij} \sim N(0, 0.0186), e_{ijk} \sim N(0, 0.1135),$$

where ‘MN’ stands for multivariate normal and  $\boldsymbol{\beta}_{(3)}$  is the vector of regression parameters under the three-level model. The regression coefficients are given in Table S4.

#### 4.2 Simulation Process

In the simulation study,  $S=500$  censuses are simulated with expenditure for every HH (rather than only the sampled HHs) based on the superpopulation model (7). ‘True’ FGT measures were then calculated for each of the 500 simulated censuses. Then a sample was drawn from each population (i.e. from each simulated census) using the HIES 2000 design FGT measures were then estimated for two-level ELL (ELL.2L), three-level ELL (ELL.3L), EBP and MQ using the survey model applied to each simulated census dataset (without using the original simulated expenditure). RB and RRMSE are calculated based on these 500 artificial populations. The Spearman rank correlation between the ‘true’ FGT measures from the simulated

Table 4. Two-level and three-level regression models fitted by restricted maximum likelihood (REML).

Model	<i>df</i>	Marginal $R^2$	Conditional $R^2$	P value	$\hat{\sigma}_\varepsilon^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_\eta^2$
2-L: Null	3	—	0.4102		0.2178	0.1515	—
3-L: Null	4	—	0.3970	< 0.01	0.2177	0.0770	0.0663
2-L: Full	33	0.5941	0.6733		0.1135	0.0275	—
3-L: Full	34	0.5962	0.6744	< 0.01	0.1135	0.0186	0.0086

census and the estimated FGT measures from the SAE methods (using the upper poverty line from BBS & UNWFP, 2004) has been calculated at small area level for each  $S$ .

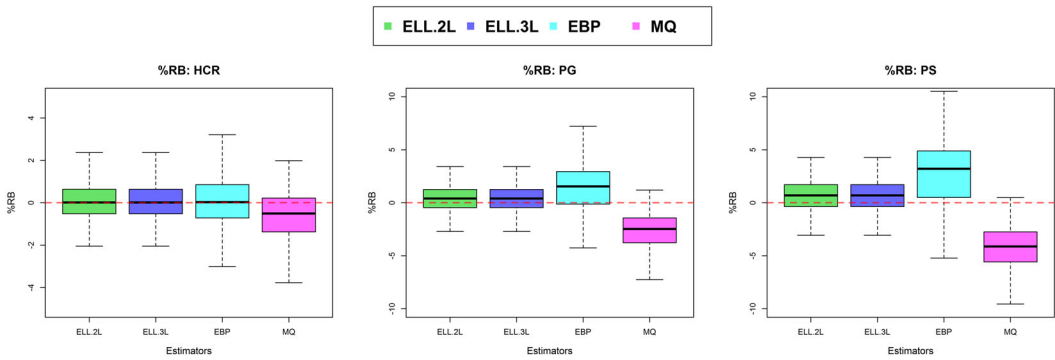
In ELL,  $L = 500$  pseudo-censuses were generated for each simulated population. Estimates of root MSE (ERMSE) of estimated FGT measures were calculated in parallel with FGT measure estimation. ERMSEs were calculated separately in both EBP and MQ methods via bootstrap procedures; to reduce the computation burden, ERMSEs were calculated using  $L = 100$  pseudo-censuses. To maintain comparison, ERMSEs for ELL were also re-calculated using only  $L = 100$  pseudo-censuses. For EBP, for each simulated population, FGT measures were calculated by generating  $L = 100$  out-of-sample vectors based on the responses of the same sample units, while in the MQ method, for each bootstrap population,  $R = 5$  samples were drawn via SRSWOR from each area, and then FGT measures were estimated via  $L = 100$  out-of-sample vectors from each sample. More samples may have reduced bias. The performance of the RMSE estimators are shown by comparing the area-specific estimated RMSE with the simulated true RMSE based on 100 of the  $S = 500$  simulated populations and by calculating coverage rates (CRs) of nominal 95% confidence interval for the RMSE estimators.

## 5 Simulation Results

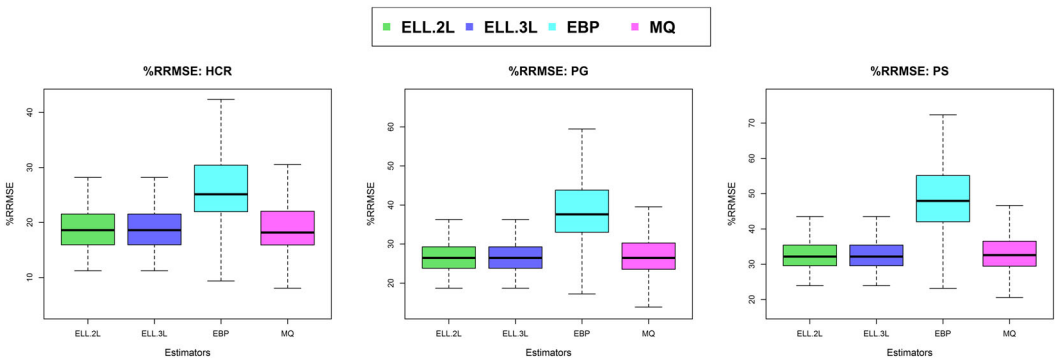
The distributions of area-specific RB and RRMSE are shown as percentages in Figure 1 and Figure 2, respectively. Despite ‘over-aggregated’ contextual variables, and a completely parametric bootstrap, the figures indicate that both ELL.2L and ELL.3L are performing better than EBP and MQ in terms of both RB and RRMSE for all the FGT indicators, perhaps not surprisingly because an ELL3L model was used to generate the data.

The EBP estimator is likely to overestimate and the MQ estimator to underestimate. These trends are obvious when FGT measure shifts from poverty incidence to PS. No significant differences are observed between the two ELL estimators in either RB or RRMSE. Both ELL estimators provide lower and more stable RBs and RRMSEs than EBP and MQ. The EBP estimator provides higher RB and RRMSE with large variation. Although MQ estimator provides downwardly biased RB, RRMSEs are lower than those of EBP estimator.

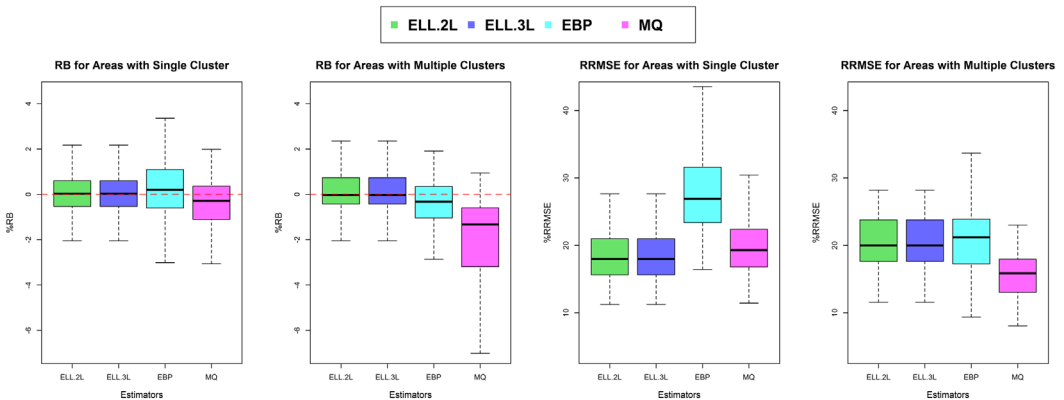
The number of sampled PSUs per area is expected to have substantial influence, because about 75% sampled areas have single sampled cluster. To examine this, RB and RRMSE of four estimators are plotted by dividing the areas into those with single and multiple clusters sampled per small area. Figure 3 shows that ELL estimators perform in the similar manner, but EBP and MQ estimators behave differently. The EBP estimator shows stable distributions of RB and RRMSE for small areas with multiple clusters when compared with those with single cluster. The MQ estimator behaves similarly to the ELL estimators for the single-sampled cluster small areas but interestingly shows higher negative RBs and smaller RRMSE for the areas with multiple clusters. So estimates from EBP and MQ are markedly influenced by the number of sampled clusters per small area. The additional advantage ELL has in the simulation reflects the design of HIES2000 that has many small areas with only one sampled cluster.



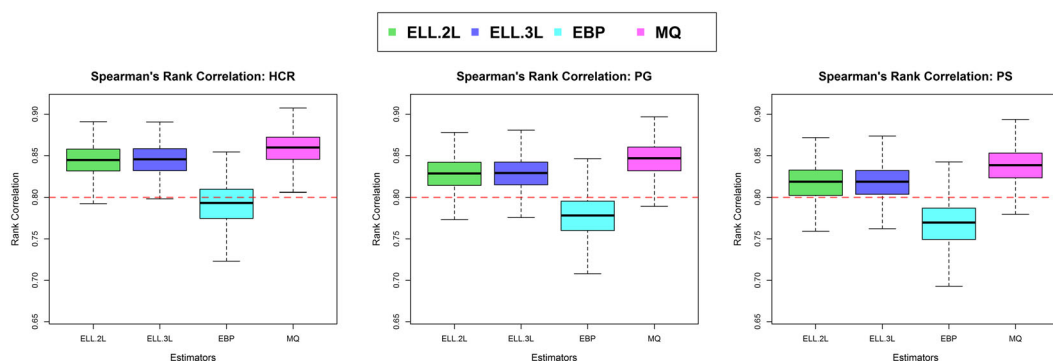
**Figure 1.** Distribution of relative bias (%RB) of ELL.2L, ELL.3L, EBP and MQ estimators of Foster, Greer and Thorbecke poverty indicators over 500 simulations. EBP, empirical best prediction; ELL, World Bank method; HCR, head count ratio; MQ, M-quantile; PG, poverty gap; PS, poverty severity. [Colour figure can be viewed at [wileyonlinelibrary.com](#)]



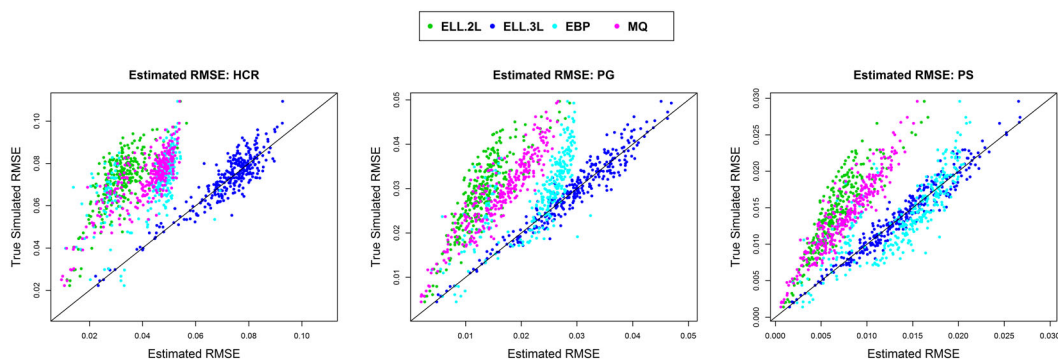
**Figure 2.** Distribution of relative RMSE (%RRMSE) of ELL.2L, ELL.3L, EBP and MQ estimators of Foster, Greer and Thorbecke poverty indicators over 500 simulations. EBP, empirical best prediction; ELL, World Bank method; HCR, head count ratio; MQ, M-quantile; PG, poverty gap; PS, poverty severity; RMSE, root mean square error. [Colour figure can be viewed at [wileyonlinelibrary.com](#)]



**Figure 3.** Distribution of %RB and %RRMSE of ELL.2L, ELL.3L, EBP and MQ estimators for HCR over 500 simulations by areas with single and multiple sampled clusters. EBP, empirical best prediction; ELL, World Bank method; HCR, head count ratio; MQ, M-quantile; RB, relative bias; RRMSE, relative root mean squared error. [Colour figure can be viewed at [wileyonlinelibrary.com](#)]



**Figure 4.** Distribution of rank correlations between the true Foster, Greer and Thorbecke measures and their estimates by ELL.2L, ELL.3L, EBP and MQ estimators over 500 simulations. EBP, empirical best prediction; ELL, World Bank method; HCR, head count ratio; MQ, M-quantile; PG, poverty gap; PS, poverty severity. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**Figure 5.** Average of estimated RMSE of estimated Foster, Greer and Thorbecke measures by ELL.2L, ELL.3L, EBP and MQ estimators against the true simulated RMSE over 100 simulations. EBP, empirical best prediction; ELL, World Bank method; HCR, head count ratio; MQ, M-quantile; PG, poverty gap; PS, poverty severity; RMSE, root mean square error. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

The distribution of Spearman rank correlations between true and estimated FGT are plotted in Figure 4 to examine whether the bias issue changes small area ranking of their poverty estimates. The MQ estimator provides good performance in ranking sub-districts (small areas) according to poverty, even though it has significant downward RB. Performance of the MQ estimator also remains the same for all three FGT measures. ELL and EBP estimators show a slightly downward trend with increasing degree of the FGT measures. The EBP estimator shows the poorest correlations for all FGT measures.

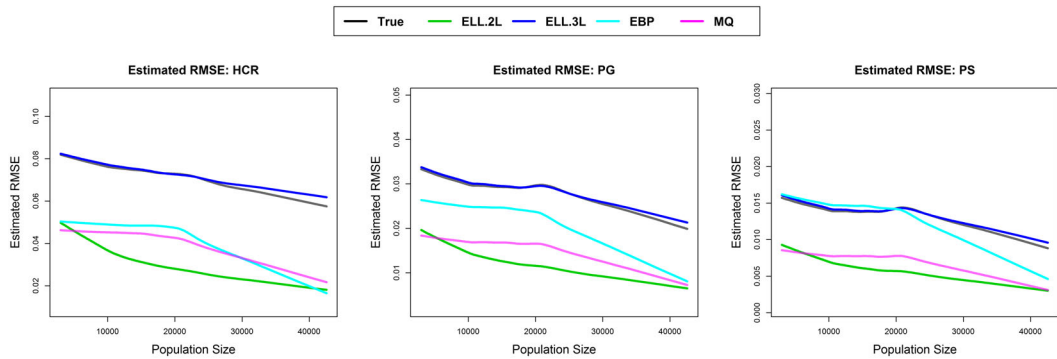
Figures 1–4 represent a situation that is not feasible to measure outside a simulation study, namely, assessment of RB and RRMSE over a number of different ‘equivalent’ populations generated from a superpopulation. In real SAE problems, there is only one population. The parallel in the simulation is to estimate RMSE separately for each population, then average. So to gauge how well the various SAE methods estimate their accuracy, the area-specific averages of estimated RMSEs over the  $S = 100$  simulations are plotted against the true simulated RMSE of FGT indicators in Figure 5. The mean and median of RB and RRMSE of the estimated RMSE over areas are also given in Table S5.

Estimated RMSE of the ELL.3L estimator tracks the true simulated RMSE best, perhaps unsurprisingly because it was used to generate the simulated censuses. The ELL.2L estimator underestimates the true RMSE, because the simulated census data contain significant small area

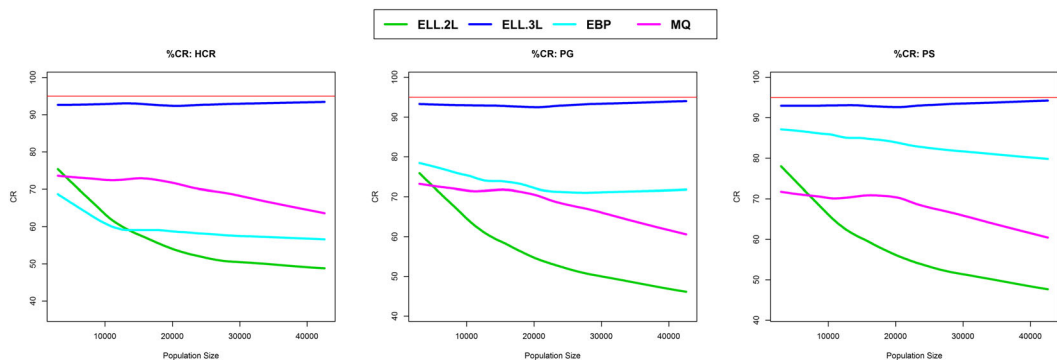
level variance estimated RMSEs for EBP and MQ also fail to track the true RMSE. With the increasing degree of FGT measures from poverty incidence to severity, estimated RMSEs of EBP shift towards the true RMSE while an opposite trend is observed for MQ. The mean and median RBs of the estimated RMSE shown in Table S5 clearly shows that ELL.2L and MQ severely underestimated the true RMSE for head count ratio and PG, while highly overestimated for PS with higher RRMSE. EBP shows a decreasing trend of RB with the degree of FGT indicators.

Average estimated RMSEs are plotted as a smoothed line against small area population size in Figure 6. All estimated RMSE estimators show downward trend with increasing population size, but performance varies for areas with large population depending on whether there are multiple clusters in the small area. Both the EBP and MQ estimators show stable ERMSE for the areas with smaller population size and steadily declining ERMSE for the larger areas. For EBP, the difference between the true and estimated RMSEs gradually decreases with the degree of FGT indicator but always shows lower ERMSE for the larger areas. The ELL estimators show a similar trend in ERMSE for all FGT indicators.

The influences of the estimated RMSEs on CR are shown in Figure 7 via the smooth trend of CRs with increasing population size. The CRs decrease with the population size for all



**Figure 6.** Average of estimated RMSE of estimated Foster, Greer and Thorbecke measures by ELL.2L, ELL.3L, EBP and MQ estimators over 100 simulations against population size of each area. EBP, empirical best prediction; ELL, World Bank method; HCR, head count ratio; MQ, M-quantile; PG, poverty gap; PS, poverty severity; RMSE, root mean square error. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**Figure 7.** Coverage rate (%CR) of ELL.2L, ELL.3L, EBP and MQ estimators of Foster, Greer and Thorbecke measures over 100 simulations against population size of each area. EBP, empirical best prediction; ELL, World Bank method; HCR, head count ratio; MQ, M-quantile; PG, poverty gap; PS, poverty severity. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

estimators except the ELL.3L estimator similar to the trend of ERMSEs. The ELL.2L estimator shows a downward curve of CRs, essentially because the simulated census data use a three-level not a two-level model. The EBP and MQ estimators also show under-CRs but perform better than the ELL.2L estimator. The performance of EBP estimator improves with the degree of FGT indicators, while there is no change for both the ELL.2L and MQ estimators. The lower estimated RMSE and lower CR are due to ignoring the between-area variability in the ELL.2L estimator and the failure of the EBP and MQ estimators to capture cluster level variation.

## 6 Concluding Remarks

To the extent possible, in the simulation study, the census and survey datasets generated and used reflect data structures in developing countries where both cluster-homogeneity and area-homogeneity assumptions can be violated, and the use of unit record census data with contextual variables can provide good area-specific adjustments even for non-sampled areas. We included substantial small area level and cluster level variation in the simulation. We showed that if only a census subsample is available, the contextual variables compromised, and the census is best described by a three-level model with area and cluster effects, then: (i) the standard ELL estimator (ELL.2L) can fail to capture significant small area level variability; (ii) the EBP estimator ignores cluster level variation and merges the cluster level variation with both individual level and area level variation; and (iii) MQ when used as a two-level model fails to capture cluster-specific random effects by not distinguishing small area-specific random effects from cluster-specific effects.

Nevertheless, the simulation results also show that in terms of the small area estimates themselves, the standard ELL estimators can perform better than the more complex EBP and MQ methods in terms of RB and RRMSE when most small domains have single cluster in the sample and a negligible between-area variation exists in the population after using contextual variables at cluster (or sub-cluster) level for both surveyed and non-surveyed small areas. As expected, when the ratio of small area to cluster variance is relatively high and contextual variables are compromised by aggregation to small area level, the basic ELL.2L estimator underestimates the true RMSE of FGT estimates and hence shows a higher under-CR. In these circumstances, the EBP and MQ estimators also show under-coverage but are still better than the ELL.2L estimator.

As an option for future study, provided census and survey cluster matching were feasible, it may be interesting to consider the performances of MQ and EBP in simulation with clusters as small areas, followed by aggregation of estimates to actual small areas. Marhuenda *et al.* (2017) consider such a 'bottom-up build' technique but instead use domains and subdomains. When the majority of the model error is at cluster level, adopting this strategy would increase performance of both EBP and MQ, especially if both were used in three-level form. This option further strengthens the argument for more extensive data availability and/or modelling.

Reasons for the partial failure of EBP and MQ methods in the simulation are (i) higher between-cluster than between-area variation; (ii) in EBP and MQ as two-level models, both individual-specific and area-specific random errors are generated from distributions that do not fully correspond with a three-level linear model; (iii) most small areas have only a single cluster in the survey data that misguides prediction of the distribution function in EBP and MQ. (The poverty mapping study of Minas Gerais state of Brazil by Souza *et al.*, 2015 is another example where 80% of target small areas were not in the survey dataset).

In terms of computational burden and not needing matching of survey and census HHs, the ELL method is easier and faster than both EBP and MQ. Where most of the target administrative



units are not available in the survey dataset, a separate calculation procedure is required in both the EBP and MQ methods. For the non-sampled small areas (none of which were included in the simulation) due to EBP's and MQ's estimation procedures for FGTs and their RMSEs prediction power will be lower than for the sampled areas

The main criticism of basic ELL method is the underestimation of RMSE when area-homogeneity assumption is violated. However, a three-level model-based ELL method can be applied when the area variability is found significant in the regression model. From the simulation study, it is clear that the three-level ELL method performs better than the traditional two-level ELL method for FGT estimates and their estimated RMSE, because in the simulation, the over-aggregated contextual variables have not been able to 'soak up' all small area-specific variation that has been incorporated into the simulated census data. Conversely, ELL.3L can be sensitive because it may overestimate the true RMSE if there are no non-zero mean small area-specific effects. For data like the simulated Bangladesh census unit record data, considering a three-level model with an appropriate variance component estimation method is critical, even if using it is thwarted by few small areas having more than one sampled cluster. Even when good contextual variables are available, both two-level and three-level ELL models used at a finer level are sensitive to how well the regression model is fitted to the survey data. ELL is routinely used to fit models in which the small areas are fewer than when EBP and MQ are typically applied, but even for fewer small areas, EBP and MQ are sensitive to significant cluster level variation when used as two-level models

Because actual census data are richer than model-based census data from cross-tabulations, in ELL, the regression model tends to have larger predictive power but will usually require more auxiliary variables, good contextual variables and considerable time and effort when model fitting and testing. The better between-cluster and between-area variations are controlled, the better the expected performance. For ELL, explanatory variables, especially census information available at cluster or preferably sub-cluster level for both sampled and non-sampled units, should be included in the regression model. Even high  $R^2$  via inclusion of more explanatory variables in the model may not guarantee capture of area variability, so great care and considerable expertise is needed in the application of ELL methodology to avoid underestimated MSE and poor CRs. On the other hand, although EBP and MQ can produce excellent SAE results, especially where there are not a large number of small areas, without contextual variables, they can be unsuitable when there is little or no sample in some of those small areas or where they are applied as two-level rather than three-level models, and there is a significant cluster level area component.

## References

- Alderman, H., Babita, M., Demombynes, G., Makhatha, N. & Özler, B. (2002). How low can you go? Combining census and survey data for mapping poverty in South Africa. *J. Afr. Econom.*, **11**(2), 169–200.
- BBS. (2003). *Report of the Household Income and Expenditure Survey 2000*. Dhaka: Bangladesh Bureau of Statistics.
- BBS & UNWFP. (2004). *Local Estimation of Poverty and Malnutrition in Bangladesh*. Dhaka: The Bangladesh Bureau of Statistics and The United Nations World Food Programme.
- Battese, G.E., Harter, R.M. & Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Am. Stat. Assoc.*, **83**(401), 28–36.
- Betti, G., Neri, L. & Salvati, N. (2007). Comparing M-quantile and ELL methods for poverty mapping. In *IASS Satellite Conference on Small Area Estimation*, pp. 15–16: Pisa.
- Breckling, J. & Chambers, R. (1988). M-quantiles. *Biometrika*, **75**(3), 761–71.
- Chambers, R., Chandra, H. & Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Surv. Methodol.*, **37**, 153–170.
- Chambers, R. L. & Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**(3), 597–604.

- Chambers, R. & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**(2), 255–268.
- Das, S. & Chambers, R. (2017). Robust mean-squared error estimation for poverty estimates based on the method of Elbers, Lanjouw and Lanjouw. *J. R. Stat. Soc. Ser. A (Stat. Soc.)*, **180**(4), 1137–1161.
- Diallo, M.S. & Rao, J. (2014). Small area estimation of complex parameters under unit-level models with skew-normal errors. In *Proceedings of the Survey Research Methods Section*, pp. 970–984. Boston, Massachusetts: American Statistical Association.
- Elbers, C., Lanjouw, J. O. & Lanjouw, P. (2002). *Micro-level Estimation of Welfare*, Vol. 2911 Washington D.C.: Development Research Group, the World Bank.
- Elbers, C., Lanjouw, J.O. & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, **71**(1), 355–364.
- Elbers, C., Lanjouw, J.O., Lanjouw, P. & Leite, P.G. (2004). Poverty and inequality in Brazil: new estimates from combined PPV-PNAD. In *Inequality and Economic Development in Brazil*, pp. 81–104. Washington, D.C.: World Bank.
- Ferretti, C. & Molina, I. (2012). Fast EB method for estimating complex poverty indicators in large populations. *J. Indian Soc. Agric. Stat.*, **66**, 105–120.
- Foster, J., Greer, J. & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society*, **52**, 761–766.
- Fujii, T. (2004). *Commune-Level Estimation of Poverty Measures and its Application in Cambodia*, Vol. 2004/48 Helsinki: UNU-WIDER, United Nations University (UNU).
- Giusti, C., Marchetti, S., Pratesi, M. & Salvati, N. (2012). Robust small area estimation and oversampling in the estimation of poverty indicators. *Surv. Res. Methods*, **6**(3), 155–163.
- Giusti, C., Marchetti, S., Pratesi, M. & Tzavidis, N. (2011). Small area methodologies for poverty estimation: an application to Italian data. In *International Statistical Institute: Proceedings of the 58th World Statistical Congress*, Dublin, pp. 6234–6237.
- González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *J. Stat. Comput. Simul.*, **78**(4), 443–462.
- Haslett, S. (2012). Practical guidelines for design and analysis of sample surveys for small area estimation. *J. Indian Soc. Agric. Stat.*, **66**(1), 203–212.
- Haslett, S. (2013). Small area estimation of poverty using the ELL/PovMap method, and its alternatives. In *Poverty and Social Exclusion: New Methods of Analysis*, Eds. Betti, G. & Lemmi, A., pp. 224–245. London, UK: Routledge.
- Haslett, S. (2016). Small area estimation using both survey and census unit record data: links, alternatives, and the central roles of regression and contextual variables. In *Analysis of Poverty Data by Small Area Estimation*, Ed. Pratesi, M., Chapter 18: Wiley. ISBN:978-1-118-81501-4.
- Haslett, S. & Jones, G. (2005). *Estimation of Local Poverty in the Philippines*. Philippines National Statistics Co-ordination Board / World Bank, November 2005.
- Haslett, S. & Jones, G. (2006). *Small Area Estimation of Poverty, Caloric Intake and Malnutrition in Nepal*. Kathmandu: Nepal Central Bureau of Statistics/World Food Programme, United Nations/World Bank.
- Haslett, S., Jones, G. & Isidro, M. (2014). *Small-area Estimation of Child Undernutrition in Bangladesh*. Dhaka: Bangladesh Bureau of Statistics, United Nations World Food Programme and International Fund for Agricultural Development. ISBN 978-984-33-9085-1.
- Haslett, S., Jones, G., Noble, A. & Ballas, D. (2010). More for less? Using existing statistical modeling to combine existing data sources to produce sounder, more detailed, and less expensive. *Official Statistics Report Series*, 5. Retrieved from <http://www.statisphere.govt.nz/official-statistics-research/series/2010/page1.aspx>.
- Haslett, S., Jones, G. & Sefton, A. (2013). *Small-area Estimation of Poverty and Malnutrition in Cambodia*. Phnom Penh: National Institute of Statistics, Ministry of Planning, Royal Government of Cambodia, and the United Nations World Food Programme. ISBN 9789996375507.
- Healy, A.J., Hitsuchon, S. & Vajragupta, Y. (2003). *Spatially Disaggregated Estimates of Poverty and Inequality in Thailand*. Massachusetts Institute of Technology and Thailand Development Research Institute.
- Marchetti, S., Beresewicz, M., Salvati, N.S. & Wawrowski, L. (2018). The use of a three-level M-quantile model to map poverty at local administrative unit 1 in Poland. *J. R. Stat. Soc.: Ser. A (Stat. Soc.)*, **181**(3), 1077–1104.
- Marchetti, S., Tzavidis, N. & Pratesi, M. (2012). Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. *Comput. Stat. Data Anal.*, **56**(10), 2889–2902.
- Marhuenda, Y., Molina, I., Morales, D. & Rao, J. (2017). Poverty mapping in small areas under a twofold nested error regression model. *J. R. Stat. Soc.: Ser. A (Stat. Soc.)*, **180**(3), 1111–1136.
- Molina, I. & Rao, J.N. (2010). Small area estimation of poverty indicators. *Can. J. Stat.*, **38**(3), 369–385.
- Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R<sup>2</sup> from generalised linear mixed-effects models. *Methods Ecol. Evol.*, **4**(2), 133–142.

- Rao, J.N.K. & Molina, I. (2015). *Small Area Estimation*. Hoboken, New Jersey: John Wiley and Sons.
- Souza, D., Brendolin, N., Quintaes, V., Pessoa, D. & Onel, S. (2015). Comparing small area estimation methods for poverty indicators in the municipalities of Brazilian State of Minas Gerais. In *ISI Satellite Meeting on Small Area Estimation*, pp. 56. Santiago, Chile.
- Tranmer, M. & Steel, D.G. (2001). Ignoring a level in a multilevel model: evidence from UK census data. *Environ. Plann. A*, **33**(4), 941–948.
- Tzavidis, N. & Chambers, R. (2007). Robust prediction of small area means and distributions. Centre for Statistical and Survey Methodology Working Paper 2007-08, University of Wollongong. Wollongong, NSW.
- Tzavidis, N., Marchetti, S. & Chambers, R. (2010). Robust estimation of small-area means and quantiles. *Aust. N. Z. J. Stat.*, *52*:2, 167–186.
- Tzavidis, N., Marchetti, S. & Donbavand, S. (2013). Outlier robust semi-parametric small area methods for poverty estimation. In *Poverty and Social Exclusion: New Methods of Analysis*, Eds. Betti, G. & Lemmi, A., pp. 283. London, UK: Routledge.
- Tzavidis, N., Salvati, N., Pratesi, M. & Chambers, R. (2008). M-quantile models with application to poverty mapping. *Stat. Methods Appl.*, **17**(3), 393–411.
- Van der Weide, R. (2014). GLS estimation and empirical Bayes prediction for linear mixed models with heteroskedasticity and sampling weights. World Bank Policy Research Working Paper 7028, The World Bank. Washington, DC.
- Zhao, Q. & Lanjouw, P. (2009). *Using PovMap2-A User's Guide*. Washington D.C.: The World Bank.

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

[Received May 2017, accepted February 2019]