



IARIW 2025

# IARIW 2025

Thursday, October 2 & Friday, October 3

## **Evaluating Alternative Approaches to Small Area Estimation of Poverty with Survey and Census Data**

David Newhouse (World Bank)  
Hai-Anh Dang (World Bank)  
Minh Do (World Bank)  
Melany Gualavisi (Amazon)  
Talip Kilic (World Bank)  
Partha Lahiri (University of Maryland College Park)  
Peter Lanjouw (VU Amsterdam)

Paper prepared for the IARIW–World Bank–UEB/VNU Conference on “Improving Well-being Measurement in Data-challenged Environments in Developing Countries for Better Evidence-based Policies” October 2-3, 2025

Session 2B Data and Methods

Time Slot: Thursday, October 2, 1:15- 3:15 PM

# Evaluating alternative approaches to small area estimation of poverty with survey and census data

Hai-Anh Dang (r) \*    Minh Do (r) \*    Partha Lahiri (r) †  
Melany Gualavisi (r) ‡    David Newhouse (r) \*    Talip Kilic (r) \*  
Peter Lanjouw (r) §    Roy Van der Weide (r) ¶

DRAFT prepared for the IARIW-UEB/VNU-WB conference in Hanoi, Vietnam.

*Please do not cite or distribute*

September 19, 2025

## Abstract

This paper uses five rounds of Mexican and Brazilian census extracts to evaluate the accuracy of different model specifications and estimation methods that use survey and census data to generate small area estimates of poverty. In the absence of measurement error and selection bias, EBP unit-level models perform slightly better than purely synthetic unit-level models in Mexico and are comparably accurate in Brazil. EBP specifications that omit household-level variables are comparably accurate with EBP unit-level models in Mexico and slightly less accurate in Brazil. Fay-Herriot area-level models and purely synthetic machine learning models are slightly less accurate than EBP specifications that omit household-level variables in Mexico and similarly accurate in Brazil. Models that omit household variables tend to be more robust to the use of old census data and classical measurement error in survey predictors. In the presence of selection bias and when using small samples, Fay-Herriot models become disproportionately less accurate, especially when variance smoothing is not applied. Rescaling sample weights is very important in the Mexican simulations due to the highly skewed distribution of population across areas. The usual practice of applying raw sample weights without rescaling in this case greatly reduces accuracy and distorts methodological comparisons. Overall, no one approach dominates across all contexts, but when sample weights are appropriately rescaled there is no downside to using more granular data for prediction.

**Keywords:** Small area estimation, poverty mapping

**JEL Codes:** C51, C52, I32

---

\*Development Economics Data Group, World Bank Group, Washington DC

†Joint Program on Survey Methodology and Department of Mathematics, University of Maryland College Park

‡Amazon.com, Seattle WA

§School of Business and Economics, Vrije University Amsterdam

¶Development Economics Research Group, World Bank Group, Washington DC

# 1 Introduction

Small area estimation (SAE) enables survey data to "borrow strength" from more geographically comprehensive auxiliary data such as census data. This enables the estimation of survey-based indicators such as poverty for highly disaggregated areas or subgroups for which there are no or insufficient household survey data to obtain reliable direct estimates. SAE can therefore be a critical input into the geographic targeting and evaluation of policies and programs.

The first known application of the SAE of poverty, combining survey and census data, was for small places in the US (Fay and Herriot, 1979) and employed Empirical Bayes estimation (Carter and Rolph (1974) and Efron and Morris (1977)). The US Census Bureau subsequently established the Small Area Income and Poverty Estimates program, which regularly produces small area estimates of income and poverty for school districts and counties. Later, an alternative method that combines survey and census data to obtain small area poverty estimates was developed by Elbers et al. (2003), henceforward referred to as the ELL approach. This method employs Monte-Carlo simulation techniques advocated by Berry et al. (1995). It has been used in over 60 countries worldwide, often with the support of the World Bank, as well as in several notable research applications (Demombynes and Özler (2005), Elbers et al. (2007), Andam et al. (2010), Crost et al. (2014), Enamorado et al. (2016), Bazzi (2017))

SAE models can broadly be divided into three groups depending on the level at which they are specified: *Area-level models* specified at the target area level, *unit-level models* specified at the household level (which represents the most disaggregated level), and *sub-area level models* which is specified at an aggregate level that is below the target area level, such as the village level. Any of these models could in principle be used to obtain purely synthetic predictions, as in Elbers et al. (2003), or alternatively to obtain Empirical

Best Predictors (EBP) that condition estimates of the random effect on survey sample data (Laird and Ware, 1982; Battese et al., 1988; Jiang and Lahiri, 2006; Molina and Rao, 2010).

When working with unit or sub-area-level models, estimates of poverty are first obtained at the household or sub-area level, and are then aggregated to the target area level.

While the literature on the small area estimation of poverty is well-established, there are comparatively few studies that evaluate the relative performance of different approaches across different settings, and existing evaluations can arrive at different conclusions (Corral et al., 2021; Das and Haslett, 2019). This is where our study aims to make a contribution. The paper is organized around five empirical questions.

1. First, how large are the merits of using EBP relative to methods that generate purely synthetic predictions across different empirically relevant settings?<sup>1</sup>
2. Second, is there utility in employing "unit-context models", in which the dependent variable is specified at the household level and all predictors are specified at a more aggregate level (either at the target area level or at a sub-area level)?<sup>2</sup> Drawing on results from Corral et al. (2021), Molina (2024) observes that while unit-context models and area-level models yield nearly equivalent results when estimating means, "estimates of poverty indicators using unit-context models can be significantly biased". This bias is attributed omitting household-level predictors. Yet in other settings, unit-context models predict poverty rates accurately (Masaki et al., 2022; Newhouse et al.,

---

<sup>1</sup>Das and Haslett (2019) and Elbers and Van der Weide (2025) find that the relative performance of EBP (assuming normally distributed errors) and ELL (allowing errors to be non-normally distributed) may vary across settings, notably with the magnitude of the area random effect and the degree of non-normality. Normal-EBP will outperform ELL when the random area effect is large and the degree of non-normality in errors is modest. Conversely, ELL does well when the random area error is small and errors exhibit a high degree of normality. Molina (2024) observes that "the gains in efficiency of EBP with respect to ELL may be remarkable when the nested error model assumptions hold and area effects are significant (equivalent to a poor explanatory power of auxiliary variables)." Finally, note that EBP will be reduced to purely synthetic estimates for target areas that are not covered by the survey sample, which may be a sizable share of target areas in many developing countries (Tzavidis et al., 2018).

<sup>2</sup>Unit-context ELL models were first proposed by Cuong (2012).

2025), particularly when predictors are incorporated at a highly disaggregated level such as at the village level (Newhouse, 2024; Haslett, 2024).

3. Third, how much is lost in aggregation? Do models using household level predictors generally outperform those using sub-area level predictors, and do models using sub-area predictors in turn outperform models using area-level predictors? Are any gains in precision marginal or meaningful, and what does this depend on? This question has become more important in recent years due to the widespread availability of geospatial data, which is typically available at the sub-area level, as predictors when recent census data is unavailable. (Van Der Weide et al., 2024; Newhouse, 2024)
4. Fourth, how robust are the different approaches to different types of “imperfect data”? This includes combining survey and census data from different years, using survey data subject to non-random selection bias and/or classical measurement error, and the use of small samples.
5. Fifth, and finally, how sensitive is comparative performance to details related to how model estimation is implemented? In particular, we focus on rescaling weights across areas, which is strongly recommended in the literature when estimating multilevel models (StataCorp, 2023; Carle, 2009; Pfeiffermann et al., 1998) but is often neglected in practice.<sup>3</sup> In addition, we also examine the sensitivity of comparative performance to the use of different methods of incorporating survey weights in the estimation of model parameters. Finally, we examine the role of using smoothed variance estimates when estimating the Fay-Herriot area-level models (Bell, 2008; You, 2022)

We employ design-based simulations based on five publicly available census extracts from Mexico and Brazil to explore each of the empirical questions listed above. These

---

<sup>3</sup>The importance of weight rescaling is also stressed in Parker et al. (2023), Rabe-Hesketh and Skrondal (2006), and Korn and Graubard (2003).

countries are well-suited for evaluating alternative approaches to the small area estimation of poverty, because their census data contain a measure of household labor income that allows us to infer small area poverty rates based on this data. It is very rare for census data to include such labor income data.

Our findings can be organized into five main insights. First, in the absence of measurement error and selection bias, unit- and PSU-level EBP models tend to produce the most accurate estimates. ELL estimates are found to be less accurate in Mexico and at par in Brazil. The advantage of EBP over ELL in the case of Mexico stems from the larger magnitude of the random area effect, which is smaller in the case of Brazil. When using Brazil's 2010 data, ELL is observed to be slightly more accurate than EBP. This confirms that comparative performance is context and country specific.

Second, there is utility in estimating unit-context models in some settings. When the sample data is collected without error and at the same time as the census auxiliary data, unit-context models that include both sub-area- and area-level predictors yield less accurate estimates in Mexico and Brazil, but the difference is very minor. When averaging across census rounds, rank correlation falls from 0.956 to 0.954 in Mexico and from 0.969 to 0.965 in Brazil when omitting household-level predictors. This translates into a tiny reduction in the poverty impact of a simulated targeted transfer program of approximately one basis point (one hundredths of a percentage point) in Mexico and thirteen basis points in Brazil. On the other hand, unit-context models appear to be more robust than unit-level models to the use of outdated census data, and can outperform unit-level models when there is classical measurement error in the survey predictors.

Third, little accuracy is lost when aggregating auxiliary data to the PSU level. Household-level EBP and sub-area EBP models perform comparably well, and unit-context models with PSU level predictors are only slightly less accurate. Machine learning models with PSU-level predictors tend to be slightly less accurate than unit-context models with PSU

level predictors in Mexico, and at par with unit-context models with PSU level predictors in Brazil. Machine learning models, despite offering greater flexibility with respect to non-linearities and interactions, do not condition estimates on the sample data. This is less important in Brazil due to the design of the simulated samples, which generates abnormally low sampling error at the PSU level. Among the two approaches that only use area-level auxiliary variables, Fay-Herriot models outperform unit-context models with area-level variables in Mexico, while the two approaches give comparable results in Brazil.

Fourth, approaches that use only area-level auxiliary data are less robust to the presence of selection bias and a smaller sample than approaches that use either household and PSU level variables or only PSU-level variables. Unlike area-level predictors, more disaggregated predictors at the PSU-level can partly correct for sample selection bias within areas. Meanwhile, when using smaller samples without selection bias, the additional variation utilized by incorporating more geographically disaggregated predictors becomes more valuable, leading to more accurate estimates. For both small samples and samples subject to selection bias, area-level models becomes much less accurate when variance smoothing is not applied. This highlights the benefits of both utilizing more disaggregated data when available and of variance smoothing when estimating area-level models.

Fifth, implementation details can be as or more important for accuracy than the choice of model and method. In particular, a seemingly minor detail related to rescaling sample weights prior to model estimation emerges as a crucial issue when estimating unit and unit-context models in Mexico.<sup>4</sup> This is because the distribution of population weights is highly skewed across municipalities in Mexico. Applying "raw" sample weights therefore leads to areas with large weights dominating the estimator, and sampling error in these areas harms the accuracy of model predictions. ELL and unit-context models, particularly

---

<sup>4</sup>To clarify, weight rescaling can be crucial when estimating a model specified at a level below the target area, such as the household or sub-area. Using "raw" sample weights remain appropriate when generating direct estimates from sample data.

those with only area-level predictors, are especially vulnerable to this source of inaccuracy when weights are not rescaled. Differences in the implementation of weight rescaling is therefore a major factor explaining the disagreement found in the literature on the relative performance of different methods and models such as EBP, ELL, unit-context, and area-level models.

The results indicate that, when weights are properly rescaled, estimating unit-level models using auxiliary data at the most geographically disaggregated level possible generally leads to as or more accurate predictions than using auxiliary data at more aggregate levels. While these benefits are minor in many of the settings we tested, using sub-area-level predictors can substantially improve accuracy when the sample is subject to selection bias and when using small samples. This is consistent with evidence from Burkina Faso, where a unit-context EBP model outperformed an area-level model with no variance smoothing by 11 correlation points when evaluated against unit-level EBP estimates ([Edochie et al., 2024b](#)).

The remainder of this paper is organized as follows. Section 2 briefly reviews the different methods and specifications for SAE models that we evaluate, which are described in more detail in Appendix A. Section 3 describes the simulation procedure, model selection, and evaluation metrics used for evaluation. This includes simulating estimation with old survey or census data, selection bias, classical measurement error in survey predictors, and varying sample sizes. Section 4 presents model diagnostics and the simulation results, and section 5 concludes.

## 2 Selected approaches to small area estimation

We consider seven distinct approaches to model specifications and estimation methods that can be used to generate small area poverty estimates using survey and census data. A *model*

*specification* in this context refers to the structure of the model, including the level at which the model is specified, the dependent variable, and the set of candidate predictor variables that are used. An *estimation method*, meanwhile, refers to the statistical algorithm that uses the model to transform the input data taken from the survey and census into small area poverty estimates.

Table 1 summarizes key differences in between the seven approaches, which Appendix A discusses in detail. The seven approaches are divided into three types of specifications, depending on the types of predictor variables that are used. Dividing the approaches into these three groups delineates differences in the nature of auxiliary data across the seven approaches, which is helpful for interpreting differences in performance.

The first group predicts household per capita income, utilizing predictor variables at all available levels, namely the household level, the PSU level, and the target area level, which is the municipality in both Mexico and Brazil.<sup>5</sup> This group includes two approaches: EBP and ELL. The primary difference between them is that EBP conditions the random effect on the sample data, while ELL estimation is purely synthetic. When estimating ELL, the estimated parameters are assumed to be fixed, to make all the methods comparable in this regard.<sup>6</sup> Other notable differences between the implementation of ELL and EBP are highlighted in Table 2.

The second group of specifications utilizes only predictors at the PSU and target area level. Throughout the paper we interchangeably use the terms PSU, village, sub-area, and cluster to refer to the primary sampling unit. This group contains three approaches, which we refer to as unit-context models with PSU and area level predictors (UC-PSU), sub-area models, and Boosted Regression Forests (BRF). In the unit-context models, the dependent variable is household income at the household level, which is linked with contex-

---

<sup>5</sup>We therefore subsequently use the terms "area" and "municipality" interchangeably.

<sup>6</sup>This differs from the usual approach to estimating ELL models, which incorporates uncertainty in the estimated model parameters when generating small area poverty estimates

tual predictors at the PSU and target area levels. For the sub-area and BRF approaches, the model is specified at the PSU level and the dependent variable is PSU-level poverty rates derived from the survey. Both the unit-context and subarea models are estimated using EBP. BRF, however, estimates one or more regression forests to predict cluster-level poverty rates. Each regression forest is based on two thousand decision trees generated using randomly generated subsets of the observations and data, as implemented in [Tibshirani et al. \(2018\)](#) and described in [Athey et al. \(2019\)](#). There are three main differences between BRF and the sub-area EBP model. The first is that BRF is purely synthetic and does not contain a random effect conditioned on the sample data. The second is that BRF is based on decision trees rather than linear models and therefore can flexibly accommodate non-linearities and interactions. Finally, BRF implements a boosting procedure that can estimate an additive sequence of regression forests. Further details on the implementation of BRF are provided in [Appendix A](#)

The final group of specifications utilizes predictors only at the target area level. This includes two approaches: Fay-Herriot area-level models, and unit-context models with only area-level predictors (UC-area). Fay-Herriot models are specified at the target area level and use the area-level poverty rate as the dependent variable. <sup>7</sup>

Unit-context models with area-level predictors are specified at the household level and predict household income using area-level predictors. The main difference between these is that the Fay-Herriot model allows for area-specific variance estimates, while the unit-context model assumes a single variance parameter for all areas. As a result, the Fay-Herriot model requires estimates of the variances of the poverty rate for each area as an input, which is estimated using the survey. Importantly, we implement a variance smoothing procedure when estimating the Fay-Herriot model, as recommended in the literature ([Bell,](#)

---

<sup>7</sup>We do not consider estimating a Fay-Herriot model at the sub-area level, because such a model would include a random effect specified at the sub-area level instead of the area-level, which would greatly reduce the benefit of the random effect.

2008; You, 2022) and described in Appendix A. Variance smoothing ensures that these estimated variances are strictly positive even when all sample households in an area are poor or non-poor. Section 4 shows how failing to smooth the variance reduces the accuracy of the Fay-Herriot estimates in samples that are smaller or subject to selection bias.

The unit-level EBP model, ELL model, and Fay-Herriot area-level model with area-level predictors are selected because they are or have been commonly used.<sup>8</sup> The other four approaches are included to examine whether there is scope to improve upon these most commonly used methods under certain conditions. For example, approaches that omit household level variables and/or PSU-level variables may be the only feasible option in cases where household level predictors are not available, and models that exclude household predictors may be more robust when using imperfect data such as old census data or survey data containing measurement error.

## 2.1 Incorporating sample weights in EBP models

Household survey data are typically collected using a two-stage sample with PSUs selected with probability proportional to population size, known as PPS sampling. When PSU population size is systematically correlated with household income and not included as a predictor variable, as is typically the case, failing to properly adjust for weights will lead to what statisticians call informative sampling bias, and what economists commonly refer to as endogenous sampling or sample selection bias. As a result, it is standard practice to adjust for sample weights when estimating descriptive statistics or model parameters (Solon et al. (2015)). The household sample weights are set equal to the product of the inverse probability of selection, which eliminates this source of bias in most cases. Unfortunately multilevel models, including EBP models, are a special case where applying standard sample

---

<sup>8</sup>It is also possible to estimate a Fay-Herriot model at the PSU level and aggregate the results to the municipal level. We do not consider this method because it does not include a random effect specified at the target area level.

weights does not eliminate this source of bias. (Carle, 2009; Pfeffermann and Sverchkov, 2009)

We consider two distinct issues related to the role of weights in model estimation. The first issue involves whether to rescale sample weights during model estimation. The literature generally recommends rescaling sample weights when estimating unit-level mixed effects models for two reasons (Pfeffermann et al., 1998). First, unlike in a standard OLS regression, finite population values in multi-level regressions are not independent of each other, meaning that the log likelihood cannot be represented as the sum of the weighted likelihoods for each observation. Second, in multilevel models, the overall weights do not carry sufficient information to correct for bias. Ideally, weighting would utilize the first stage selection probabilities, or the probability that each sampled PSU was selected. Unfortunately, this information is not typically included in household survey data files.

There is, however, a third compelling reason to rescale sample weights. Rescaling sample weights increases the effective sample size of areas, especially when the distribution of sample weights across areas is highly skewed. Using raw sample weights in this case risks giving disproportionate weight to a few influential areas. The model may then fit sampling error in these highly influential areas, reducing the accuracy of both the estimated parameters and the resulting estimates of poverty. We show below that this is very significant empirically in the Mexican case.

Different ways have been proposed to rescale weights when only final sample weights are available. We adopt “Pfeffermann’s method 2” (Pfeffermann et al., 1998) which rescales the sum of the weights for each target area to equal the sample size for that area. This is a relatively simple approach that gives each target area weight according to its sample size, which helps correct for heteroscedasticity due to differing sample sizes across areas. Weighting each area equally, after accounting for heteroscedasticity due to sampling error, is also consistent with the standard approach to estimating the Fay-Herriot area-level model

(Fay and Herriot, 1979; Halbmeier et al., 2019). As we show below, estimating a unit-level model using the provided sample weights without rescaling, when the weights are highly skewed across areas, not only harms the accuracy of estimates, but also distorts comparisons across methods and specifications.

The second issue involves the choice of method for incorporating weights when estimating linear mixed models. We consider four weighting methods that have been proposed in the literature: The "conditional weighting method" implemented in the nlme and lme4 R packages (Bates et al., 2015), a "partial adjustment method" that partially adjusts unweighted estimates (Guadarrama et al., 2018), a weighted Generalized Least Squares (GLS) method (Huang and Hidirolou, 2003; Van der Weide, 2014), and a new approach that combines the partial adjustment method and the conditional weighting method that we refer to as "Hybrid weights". Appendix B describes these four methods in detail. In section 4 below, we show how predictive accuracy for different EBP models depend on the choice of method used to incorporate weights.

### 3 Design-based simulations

This section describes the structure of the data and design-based simulations used to evaluate the comparative performance of the methods and specifications described in the previous section. It covers the census data used, the process for model selection given a set of candidate variables, the process used to draw the survey samples, and the simulation of various imperfections in the survey or census data. These imperfections include the use of old census or survey data, selection bias in the survey, measurement error in predictors used in the survey data, and the use of smaller survey samples.

### 3.1 Census extract data

We obtained the 2010, 2015, and 2020 Mexican census microdata from the INEGI website, and the 2000 and 2010 Brazil census microdata from the IBGE website. For Mexico, the public-use 2010 and 2020 data are 10 percent census extracts. The 2015 data is from a 20 percent intercensus sample. In Brazil, the 2000 and 2010 data are 10 percent census extracts. Table 3 gives the number of municipalities, PSUs, and households in the census extract for each round, while table A1 shows the applicable poverty lines and resulting national poverty rates calculated in the census.<sup>9</sup> An important difference between the countries is the relative number of PSUs and municipalities in the census extracts. In Mexico there are approximately 70 times as many PSUs as municipalities in the census extracts, while in Brazil there are only approximately twice as many PSUs as municipalities. Because the Brazilian samples select up to ten PSUs per municipality, all population PSUs are selected in each sample in approximately 98 percent of the Brazilian municipalities. The Brazilian simulations therefore have an abnormally low amount of sampling error in the second stage, which limits the benefits of using PSU-level predictors relative to area-level predictors in the Brazilian case. The results therefore illustrate how the design of the survey can impact the relative performance of different estimation approaches.

### 3.2 Constructing synthetic survey samples

For both Mexico and Brazil, we follow the sampling strategy of the relevant household surveys, the MCS-ENIGH in Mexico and the PNAD Continua in Brazil. To execute the design-based simulations, we drew 100 samples from the census extracts, using a three stage sampling design that first selected municipalities, then PSUs within selected municipalities, and finally households within selected PSUs.

---

<sup>9</sup>In each case we use a single poverty line, calculated as the weighted average of national moderate poverty lines

For Mexico, we first selected 900 municipalities out of a total of approximately 2,460 to be included in the sample. This matches the official survey, the MCS-ENIGH, and also ensures a sufficiently large sample to estimate results both for in and out of sample municipalities. In all samples, we included approximately 370 municipalities with the largest population, and used probability proportional to size (PPS) sampling to select the remaining municipalities. The second step sampled PSUs within selected municipalities using simple random sampling. A maximum of 8 PSUs per municipality were selected, which yields samples of about 4 percent of all PSUs. In the third stage, we selected 5 households per PSU in highly urban areas and 20 households per PSU in other urban and rural areas, again using simple random sampling. The difference in the number of households selected in highly urban and other areas introduces informative samples within municipalities. As shown in Table 3, this results in a main sample of approximately 79,000 households in 2010, 106,000 households in 2015, and 74,000 in 2020. Sample weights are constructed as the product of the inverse probability of selection in each stage.

For Brazil, we also began by sampling 900 municipalities. Just under 200 municipalities with the largest population counts were selected with probability one. The remaining municipalities were sampled with probability proportional to their population size. The second stage sampled up to 10 PSUs from each municipality using simple random sampling. As noted above, in both 2000 and 2010 approximately 98 percent of the municipalities contain ten or fewer PSUs, meaning that all census PSUs in these municipalities are selected in each simulated sample. Finally, the third stage sampled 28 households per PSU in both 2000 and 2010 using simple random sampling. As shown in Table 3, this resulted in a total sample of approximately 73,000 households in 2000 and 92,000 in 2010. The Brazilian samples are not informative within municipalities because both the second and third stage were selected using simple random sampling while the number of households selected in the third stage is constant for all PSUs.

### 3.3 Candidate variables and model selection

In the Mexican case, we consider the following candidate predictors:

1. Household size and its square
2. Head's years of education
3. Household Assets: Stove, Shower, Radio, Television, Refrigerator, Washing Machine, Car, Cell phone, internet access.
4. Housing characteristics: Improved wall, improved floor, improved water, improved roof, electricity, flush toilet, sewage, type of cooking fuel
5. Whether the household owns their home and the number of household members per room

The set of candidate predictors is similar in Brazil. Improved water, roof, and flush toilets were not available in the Brazilian data, while motorcycle ownership was included in Brazil but not Mexico. The square of household size was included at the household level but not at the PSU or municipal level. This slightly favors models that use household-level variables over those that rely solely on aggregate variables, but not by enough to substantially change the results.<sup>10</sup> Overall, the census in both countries provide a rich set of variables with which to predict household per capita income.

For each approach that we evaluate, the predictor variables are selected using the LASSO.<sup>11</sup> In particular, we rely on a variant of LASSO implemented in Stata that minimizes the Bayesian Information Criteria (BIC). In addition, we utilize the postselect option,

---

<sup>10</sup>In the baseline specifications, removing the square of household size as a candidate predictor only decreases rank correlation in the unit-level models by 0.0006 in the 2010 Mexican data

<sup>11</sup>We do not implement LASSO for estimating Boosted Regression Models because they are based on decision trees do not require prior model selection

which does not shrink the coefficients when calculating the BIC. This option is appropriate when using LASSO for selecting variables to be used in subsequent estimation, as is the case here where we are implementing "post-lasso" estimation (Belloni et al., 2014).

### 3.4 Simulating imperfect data

The samples described in subsection 3.2 assume a best case scenario with regard to data availability and quality. The simulated survey and census were collected at the same time, the sample weights correctly reflect the inverse probability of selection, the survey data is collected without error, and the sample size approximately reflects the household surveys used for poverty measurement in each country. We consider modifications to the input data that relax these assumptions to be more consistent with common real-world settings.

#### 3.4.1 Survey and census data from different years

In many applications, the available survey and census data are collected in different years, which can lead to bias in small area estimates. For example, if the survey is collected in year  $t$  and the census is collected in a prior year  $t - l$ , simulated welfare will be based on  $X_{t-l}\hat{\beta}_t$ , then this may bias predictions of  $y_t$  to the extent that  $\beta_t$  varies across time.<sup>12</sup> The same applies to combining old survey data with new census data. The use of old survey data may introduce an additional bias in estimates of the area random effects when using EBP estimation.

To investigate how robust different specifications are to old survey or census data, we draw the survey from one year's census and use the census from a different year. We simulate both the case where the census is older than the survey and the case where the survey is older than the census. In each case, the benchmark for evaluation is poverty

---

<sup>12</sup>This argument also carries over to variance parameters; if these are not time-invariant, then a misalignment in survey and census years may introduce a bias in poverty estimates.

calculated from the full census from the most recent year.

For example, in Mexico there are three rounds of census extracts available, allowing us to simulate old census data in three ways: The first is to use simulated 2020 survey data and 2015 census data as inputs. The second is to use simulated 2020 survey data and 2010 census data as inputs. In both of these cases, we evaluate accuracy using the full census extract from 2020, which we assume is the year of interest. Finally, we can use simulated 2015 survey data and 2010 census data as inputs, and evaluate against the 2015 census data. In Brazil, where we only utilize two rounds of census data, we simulate age bias in the census by using the 2010 survey and 2000 census, and evaluate against the 2010 census.

Similarly, we can use a similar approach to investigate the use of old survey data, by drawing the survey from an older census. For example, in Brazil we simulate age bias in the survey by using a 2000 survey and 2010 census, and evaluate predictions against the 2010 census. In all cases, we drop any household variable for which the confidence interval in the survey mean does not contain the census mean, following standard practice when generating SAE estimates. This helps limit the bias in local poverty estimates due to a misalignment of survey and census years.

We face an important and unavoidable limitation when using old survey or census data: In both countries, the PSU identifiers are not consistently defined across years due to confidentiality restrictions, so it is not possible to match PSU identifiers across years. Therefore, we cannot estimate how age bias affects the performance of either unit-context models with PSU-level aggregates, or sub-area models of PSU-level poverty rates. This is unfortunate because in actual applications it is plausible that national statistics offices can link PSU identifiers between survey and census data collected in different years. Nonetheless, comparing the performance of unit-context models that use only area-level aggregates with unit-level models that use household level variables can provide insight into the robustness of each specification to bias caused by the use of old survey or census data.

### 3.4.2 Selection bias

Survey data can be prone to selection bias. Selection bias in this context implies that the probability of selecting a PSU is determined partly by PSU characteristics that are correlated with average welfare or poverty, which are not accounted for when calculating sample weights. We simulate an extreme version of selection bias by making the selection probabilities of each PSU in the second stage of the three-stage sampling design depend on a household asset index, in a way that excludes all PSUs in the bottom third of the asset index distribution from the sample. The procedure is implemented as follows:

1. Calculate  $\pi_{ic} = \frac{pop_{ic}}{\Sigma_{pop,i}}$  as the share of the municipal population in PSU  $c$ , where  $pop_{ic}$  is the population of cluster  $c$  located in municipality  $i$  and  $\Sigma_{pop,i}$  is the total population of municipality  $i$ .  $\pi_{ic}$  is therefore the probability of PSU  $c$  being selected in the second stage, conditional on municipality  $i$  having been selected in the first stage, when no selection bias is present.
2. For each PSU, calculate the elements of the household asset index by taking the weighted mean, weighting by household size, of 18 household characteristics. Ten of the 18 characteristics are asset ownership dummies: cell phone, computer, car, washing machine, refrigerator, telephone, radio, shower, stove, and owning the home of residence. The remaining eight are dwelling characteristics: household members per room, whether the household cooks with fuel, uses a flush toilet, uses a shared toilet, has access to internet, and whether the house has an improved floor, roof, and walls. Next, calculate the first principal component of the 18 mean household welfare characteristics proxies from step 2. Finally, calculate  $a_c$ , the unweighted percentile of the first principal component, for all PSUs in the census extract.
3. Adjust the sampling probabilities as follows:  $\pi'_{ic} = \pi_{ic} * \left[ \min \left( \left( \frac{a_c}{35} \right)^{11}, 1 \right) \right]^{11}$ . As shown in Figure A, this excludes all PSUs in the bottom third of the asset index

from the sample, while leaving the relative probability of selection unchanged for all households above the 35th percentile.

4. Draw the second stage of the sample, giving each PSU probability  $\pi'_{ic}$  chance of selection.
5. Draw the third stage of the sample, giving each household within selected PSUs equal chance of being selected.

All subsequent analysis is conducted using sampling weights that assume the second stage selection probability is  $\frac{1}{\pi_{ic}}$ , derived prior to the adjustment in step 3. In this scenario, the household weights no longer equal the inverse probability of the households being sampled, which introduces selection bias. Selection bias at the cluster level can occur in practice if the sample of PSUs is not entirely random, for example because less accessible (and poorer) PSUs are less likely to be included than their population share would indicate, or because there are PSUs affected by conflict that are impossible to safely travel to.

This is only one illustrative model of selection bias. Below, we examine the correlation between direct estimates under selection bias and truth, to ensure that the degree to which the simulated selection bias degrades accuracy is reasonable. More research on how model-based estimates are affected by different patterns of selection bias in survey data would be useful. Nonetheless, this illustrative model provide a useful first insight into the robustness of different methods and specifications to a specific form of selection bias in the sample.

### 3.4.3 Classical measurement error

Survey data is subject to measurement error, due to the method of data collection, the respondent, or the questionnaire (Biemer et al. (2013)). Measurement error can be caused by reporting bias, errors in recollection, and errors in enumeration. We take a first step towards allowing for measurement error by simulating classical measurement error in pre-

dictor variables obtained from the survey. Specifically, for continuous predictor variables, we add a normally distributed error term distributed with mean 0 and variance equal to half the cross-sectional variance in the survey variable, estimated from the sample. Thus, we simulate a reliability coefficient of 0.5. For dummy variables, we simulate an incorrect report with 10 percent probability, replacing 0 with 1 and vice-versa in those cases. In Brazil, educational attainment is measured as an ordinal categorical variable. In this case, for 5 percent of households we add one if possible, and for 5 percent of households we subtract one if possible. These are also intended to be illustrative of how one particular type of measurement affects the accuracy of estimates. Further research would be useful to examine the impact of other types of measurement error in the survey, including non-classical measurement error ([Bound et al. \(2001\)](#)), as well as measurement error in the census.

#### **3.4.4 Smaller samples**

Finally, we explore drawing three types of smaller survey samples to examine how this affects the relative accuracy of different approaches. We refer to these samples as “large”, “medium”, and “small” and report their corresponding sample sizes in [Table 3](#). The main set of samples for Mexico, as described in [Section 3.2](#), includes up to 8 PSUs per municipality in the second stage, while the third stage includes 20 households in each rural areas and 5 households in each highly urban area. This implies a relatively large sample of approximately 78,000 households in 2020, 106,000 households in 2015, and 74,000 households in 2020. In all three of the smaller Mexican samples we reduce the second stage to 5 PSUs in each sampled municipality. The three variants thus differ in the size of the third stage of the samples. The “large” sample includes 16 households per PSU in rural areas and 4 households per PSU in urban areas, the “medium” sample includes 7 households in rural PSUs and 2 in urban PSUs, and the “small” sample includes 3 households in rural areas and 1 in urban areas. Thus, as seen in [Table 3](#), the total size of the variants are approximately

55%, 30%, and 15% of the size of the main samples. For Brazil, the main set of samples described in section 3.2, all samples included up to 10 PSUs per municipality. This was not changed for the three types of smaller samples. Instead, the third stage was reduced from 28 households in the third stage to 16 for the large sample, 8 for the medium sample, and 4 households per PSU for the small sample. Therefore, the total size of the variants are approximately 60%, 30%, and 15% of the size of the main samples.

### 3.5 Evaluation metrics

We focus on two comparative performance metrics: The rank correlation between the estimates and the truth, and the mean absolute error (MAE) of the estimates. The rank correlation provides a measure of how accurately different methods rank target areas from poorest to least poor. This measure only evaluates the accuracy of rankings and not the estimated poverty levels themselves. MAE is therefore a useful supplemental measure of accuracy.

If we define  $\hat{P}$  as the vector of area estimates to be evaluated and  $P^*$  as the benchmark “truth” derived from the census data, and  $R(\hat{P})$  and  $R(P^*)$  measure the ranking across areas of the predicted and true poverty rates respectively, then the rank correlation is defined as:

$$RC = \frac{Cov\left(R(\hat{P}), R(P^*)\right)}{\sigma_{R(\hat{P})}\sigma_{R(P^*)}}, \quad (1)$$

where  $\sigma_{R(\hat{P}_i)}$  and  $\sigma_{R(P_i^*)}$  are the standard deviations of  $R(\hat{P}_i)$  and  $R(P_i^*)$ . The MAE is defined as:  $\frac{1}{A} \sum_{i=1}^A \left| \hat{P}_i - P_i^* \right|$ , where  $A$  is the number of municipalities in the population, and  $\hat{P}_i$  and  $P_i^*$  are the estimated and true poverty rates, respectively, for municipality  $i$ . Therefore, for both evaluation indicators, each municipality is given equal weight.

Finally, we report the results of a simulated transfer program that provides a fixed per capita income transfer to households equal to 10 percent of the poverty line, following

Merfeld et al. (2025). In each case, households are ranked according to the estimated poverty rate of their area, estimated using different approaches. The simulated program provides transfers to households whose estimated municipal poverty rate is above the  $X^{th}$  percentile, where X is equal to one hundred minus the national poverty rate. This ensures that a share of the population approximately equal to the national poverty rate benefit from the program in each case.<sup>13</sup> We then recalculate the poverty gap following the simulated transfer, for each of the 100 samples and average across them. We report the average poverty gap rather than the headcount rate when simulating impacts on poverty, because targeting areas based on their poverty headcount rates maximizes the impact of a transfer program on the poverty gap (Besley and Kanbur, 1991; Kanbur, 1986).<sup>14</sup> In addition, we report the share of poverty reduction achieved by each method relative to that achieved by using an ebp unit level model, which is the approach that generally produces the most accurate estimates. This indicates the extent to which the poverty impact of the simulated transfer program, in terms of its effect on the poverty gap, is reduced when targeting based on headcount poverty estimates generated by other approaches.

### 3.6 Direct estimates

The direct estimates derived solely from the survey data are useful as a benchmark. One would expect small area estimates to be more accurate than direct estimates due to the utilization of additional information from the census data. Direct estimates of the poverty headcount rate for each area can be obtained following Horvitz and Thompson (1952) and

---

<sup>13</sup>The share of the population receiving simulated transfers varies slightly across the methods used to targeting, due to differences in the cumulative population distribution when ranking municipalities according to different sets of estimates.

<sup>14</sup>In contrast, targeting areas based on the share of households that are "barely poor", and would become non-poor as a result of the transfer, would maximize the impact of a transfer program on headcount poverty.

Foster et al. (1984) as:

$$\hat{P}_i^{dir} = \frac{\sum_{j=1}^{n_i} w_{ij} I(Y_{ij} < Z)}{\sum_{j=1}^{n_i} w_{ij}}, \quad (2)$$

Where  $\hat{P}_i^{dir}$  is the estimated direct poverty estimate for area  $i$ ,  $n_i$  is the number of sample households in area  $i$ ,  $w_{ij}$  is the product of household size and the household sample weight for household  $j$  in area  $i$ , and  $I(Y_{ij} < Z)$  is an indicator function for whether household per capital income  $y_{ij}$  falls below the poverty line  $Z$ .

Besides providing a useful benchmark for evaluation, the variance of the direct estimates are essential inputs when estimating the Fay-Herriot model. We follow [Molina and Marhuenda \(2015\)](#) by using the Horvitz-Thompson approximation to estimate the variance of the direct poverty estimates. The formula for the Horvitz-Thompson variance approximation, using the sum of the sample weights as an approximation of population size for each area, is given in [appendix A](#). This estimator assumes that the probability of selection of each unit into the sample is independent of the probability that any other unit is selected. This assumption does not hold in standard two stage samples, because two households from the same PSU are more likely to be sampled than would be the case if each household were selected independently from the population. However, the estimator appears to be quite robust to violations of this assumption, as coverage rates are generally reasonable in design-based simulations ([Masaki et al. \(2022\)](#), [Newhouse et al. \(2025\)](#), [Edochie et al. \(2024b\)](#)) and can be far more accurate than standard cluster-robust variance estimates that do not account for intercluster correlation within target areas. We then follow best practice by smoothing these direct estimates of variance, as described in [appendix A](#). This ensures that all variance estimates are strictly positive, which greatly improves accuracy in some cases.

## 4 Results

This section presents the results of the design-based simulations and is divided into four subsections. The first subsection presents various diagnostics associated with the considered models and methods. The second discusses the comparative performance of models and methods when there are no data imperfections. The third analyzes how comparative performance changes when the sample data is subject to various imperfections. The final subsection examines how rescaling survey weights, as well as different methods for incorporating weights, impact comparative performance.

### 4.1 Model diagnostics

Table 4 gives basic diagnostics for three EBP specifications: Unit-level models, PSU-context models, and area-level models. It shows that the model fit, in terms of explaining inter-area variation, is much greater for Brazil than for Mexico. This is seen clearly in the Intraclass Correlation Coefficient (ICC), defined as  $ICC = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$ . Across the three Mexican surveys, the average ICC ranges from 5 percent for the unit-context model with PSU covariates to 7.3 percent for the unit-context model with area-level covariates. The analogous range for Brazil is only 0.9 percent to 1.4 percent. As a result, we expect the benefit of using EBP relative to ELL to be greater in Mexico than Brazil.

With respect to the normality of the residuals, we see greater normality in Mexico than Brazil. In particular, the kurtosis of the area effect in Brazil is 14 to 15, much greater than the value of 3 associated with a normal distribution. In addition, kurtosis is high in the 2020 Mexican census, at 20 for the household error and 9 for the area effect residual (results not shown). This should also raise the relative performance of ELL relative to EBP in Brazil, since ELL does not assume the error terms are normal.

Table 3 reports the marginal and conditional  $R^2$  both across households and across

target areas. Marginal  $R^2$  refers to the share of the variance in welfare explained by the predictor variables  $x_{ij}$ , while conditional  $R^2$  refers to the share of the the variation explained by  $x_{ij}$  and the predicted random effects  $\hat{b}_i$ . Area  $R^2$  is an important indicator because the model is ultimately used to generate area-level estimates.

A striking finding is the high area  $R^2$  values in Brazil relative to Mexico. This reflects the success of the model in explaining variation in welfare across municipalities, in part because of the abnormally low amount of sampling error in the second stage of the Brazillian samples. As expected, household-level  $R^2$  falls greatly when moving from unit-level models to contextual models. However, moving from unit to contextual models leads to relatively small declines in area-level  $R^2$  values, suggesting that unit-context models may perform well for generating area-level estimates.

Finally, given the potential importance of bias due to informative sampling when weights are not rescaled, Table A2 examines several measures related to the informative nature of the sample. The first row gives the average skewness of the weights across areas. The results show that the distribution of weights across areas is positively skewed in all cases, with the most skewness in Brazil, followed by Mexico 2010, Mexico 2015, and Mexico 2020. In part because of this skewness, using population weights without rescaling reduces the effective sample size, in terms of the number of areas. The second row gives the effective sample size, defined as

$$ESS = \frac{\left(\sum_{i=1}^M w_i\right)^2}{\sum_{i=1}^M w_i^2} \quad (3)$$

where  $w_i = \sum_{j \in i} w_{ij}$  is defined as the sum of the sample weights for all sample households in area  $i$ . Applying sample weights greatly reduces the effective number of areas in the sample, for example from 900 to 85 in Mexico 2000 and from 900 to 144 and 147 in Brazil's 2000 and 2010 round respectively. The third row of Table A2 gives the result of an informal test of informative sampling recommended by Pfeffermann and Sverchkov (2009). Specifically, the

table reports the average, across one hundred samples, of  $|\hat{\gamma}|$  estimated from the following weighted OLS regression:

$$w_{ij} = x'_{ij}\beta + y_{ij}\gamma + \epsilon \quad (4)$$

where  $w_{ij}$  is the sample weight for household  $i$  in area  $j$ , and  $x_{ij}$  and  $y_{ij}$  are defined as in equation 6. The average absolute value of the coefficients is exceptionally high in the Mexico 2010 round, at approximately 90,000, but also exceeds 2000 in the other two Mexican rounds. In Brazil, it is much lower, at approximately 300. The fourth row shows a measure of informative sampling within areas, obtained from estimating  $\hat{\gamma}$  from a demeaned version of equation 4:

$$(w_{ij} - \bar{w}_i) = (x_{ij} - \bar{x}_i)' \beta + (y_{ij} - \bar{y}_i) \gamma_{wa} + \epsilon \quad (5)$$

where  $\gamma_{wa}$  indicates within-area. We report the average absolute value of  $\gamma_{wa}$  across samples. Within areas, the Mexico 2010 round continues to stand out as very highly informative, with a measure over 70000. The Mexico 2015 and 2020 rounds are also clearly informative, with a measure of at approximately 2500, but this measure falls to approximately 100 in Brazil. This reflects the differences in the design of the second stage of the sampling design, which is explicitly informative in Mexico and uses simple random sampling in Brazil.

The bottom panel of Table A2 shows comparable diagnostics when the weights are rescaled. After rescaling the weights, skewness across areas falls dramatically, with a resulting increase in the effective sample size of areas relative to using unscaled weights. The most striking difference is in the measure of informativeness, which falls to essentially zero both in the full sample and within areas. When examining the within-area measure of informativeness, this is entirely due to the difference in the weights applied to the regression. Using rescaled weights increases the effective sample size across areas by giving areas more equal weight during model estimation. This in turn makes model estimation less prone to

sampling error in particular populous areas that are given disproportionate weight when using raw sample weights. Thus, rescaling the sample weights makes the estimates far less susceptible to inaccuracy due to sampling error.

## 4.2 Results with "perfect" data

*Finding 1.1: EBP estimates are slightly more accurate than ELL estimates in Mexico but equally accurate in Brazil.*

This can be seen in Table 5, which shows average rank correlation and mean absolute error across the three rounds for Mexico and two rounds for Brazil (Table A3 confirms that the same patterns hold for each round). The left three columns show results for Mexico while the right three show results for Brazil. Results are shown separately for all municipalities, in-sample municipalities, and out of sample municipalities. Each row represents a different method and model specifications. Sample weights are rescaled to sum to the sample size in each area, and the EBP models are estimated using hybrid weights, as described in Appendix B.3.

The comparisons between EBP and ELL unit-level models can be seen in the top two rows of the top and bottom panels. In Mexico, EBP unit-level models on average have a rank correlation with the true values of 0.956, as compared with 0.948 for ELL. Similarly, when looking at mean absolute error, the average for unit-level EBP models is 5.3 percentage points, as opposed to 5.9 percentage points for ELL. As expected, the advantage of EBP is greater for in-sample areas, though it also retains an accuracy advantage in out-of-sample areas.<sup>15</sup> Figure 1 shows the full distribution of errors across methods and rounds. The bottom panel reports the combined distribution for each country. When comparing the top two bars, the distribution of errors is slightly greater for ELL than EBP. Finally,

---

<sup>15</sup>This may indicate that differences in the estimation procedure for ELL such as the correction for heteroscedasticity reduces accuracy in this case.

in the targeting simulations described in Section 3.5 and reported in Table 6, the impact of the program on the poverty gap is 99.83 percent the size of the impact when using EBP estimates, indicating a 17 basis point reduction in the poverty impact when targeting based on ELL estimates instead of EBP estimates.

The results from Brazil are quite different from those from Mexico, partly due to the small intra-cluster coefficient, which is 0.014 on average in Brazil and 0.063 in Mexico (Table 4). In Brazil, unit-level EBP and ELL estimates are equally accurate, with an overall rank correlations of 0.969. When judging by mean absolute error, ELL is very slightly more accurate on average than EBP, with an average MAE of 3.9 percentage points for Brazil as opposed to 4.0 percentage points for EBP. There is no meaningful difference between in-sample and out-of-sample areas. In Figure 1, there is no distinguishable difference in the distribution of errors. Table 6 indicates that a simulated transfer program that targets based on ELL has the same impact on poverty in Brazil as one that targets based on EBP.

*Finding 1.2: EBP specifications that omit household-level variables are comparably accurate with EBP unit-level models in Mexico and slightly less accurate in Brazil.*

This can also be seen in Table 5, Figure 1, and Table 6. As noted above, table 5 indicates that the Mexico EBP unit-level models on average have a rank correlation of 0.956 with the true poverty rates. Sub-area models, which use PSU-level poverty estimates as the dependent variable, have the next highest average correlation at 0.955. This is followed closely by unit-context models with PSU-level predictors, at 0.954. Differences in the full distribution in Figure 1 are barely noticeable, although sub-area model estimates appear to be slightly less susceptible to large inaccuracies. The minor differences across approaches translate to negligible differences in poverty impacts when simulating the transfer program. Poverty reduction fall 4 basis points when using the sub-area model instead of the ebp unit-level model, and 1 basis point when using the unit-context model (Table 6). In Brazil,

sub-area models are still only slightly less accurate than unit-level models, with accuracy falling from 0.969 to 0.967. PSU unit-context models are in turn slightly lower than sub-area models, with an average rank correlation of 0.965. In both countries, similar patterns are seen in mean absolute error. Table A3 shows that the patterns are similar for each year within countries. When looking at poverty impacts in Table 6, unit-context and sub-area models reduce the poverty gap by 13 and 15 basis points less than ebp unit-level models, respectively.

*Finding 1.3: Fay-Herriot area-level models and PSU-level machine learning models are slightly less accurate than EBP specifications that omit household variables.* This can also be seen in Tables 5, 6, and Figure 1. Both Fay-Herriot and PSU-level machine learning models are slightly less accurate than sub-area and PSU unit-context models. In Table 5, BRF and Fay-Herriot models have a rank correlation of 0.948 in Mexico on average. This is slightly below the PSU unit-context and sub-area models, which are at 0.955 and 0.954. Figure 1 shows little discernible difference between Fay-Herriot, BRF, ELL models, and PSU unit-context models, though each tends to be a tiny bit less accurate than EBP unit and sub-area model. Table 6, targeting based on the Fay-Herriot estimates is comparable but slightly less effective than using the sub-area model and PSU unit-context models, as the poverty gap is reduced by 10 basis points relative to the ebp benchmark (as opposed to 1 and 4 basis points). In Brazil, meanwhile, the Fay-Herriot model has an average rank correlation of 0.965, which is equal to the PSU unit-context model (0.965) and only slightly below the sub-area model (0.967). Targeting based on the Fay-Herriot model reduces the poverty gap by 34 basis points less than the unit-level models. This is 21 basis points less than the comparable figure for ebp unit-context models, which is 13 basis points.

## 4.3 Data Imperfections

### 4.3.1 Old survey or census data

*Finding 2.1: Models that use only area-level predictors are more robust to the use of old survey or census data than models that use household-level predictors.*

Tables 9 and 10 show results in Mexico and Brazil when using old survey or census data. As noted above, above, PSU level aggregates are unfortunately unavailable for this exercise. We therefore only consider specifications that use household and area-level variables or specifications that use area-level aggregates only. In general, the models using area level predictors are more robust than the unit-level models when using old census data, largely negating the advantage of the latter when there is no bias. For example, when examining the third column of numbers in table 9, in terms of rank correlation the unit-context model is only slightly less accurate than the EBP unit-level model (0.925 vs 0.935) and the Fay-Herriot model (0.939) is more accurate. The same pattern is apparent in the second column of Table 10, where the area-context model is equally accurate (0.951) and the Fay-Herriot model slightly more accurate (0.953) than the unit-level EBP model (0.951).

This pattern also appears when using old survey data. For example, looking at the 4th column of Table 10, the area-level unit context model (0.938) is slightly more accurate than the unit-level model (0.937), while the rank correlation of the Fay-Herriot model (0.944) is about 0.7 percentage point higher than the unit-level EBP model. Unfortunately we cannot observe the performance of unit-context models with PSU level variables when age bias is present. But Table 5 shows that unit-context models with PSU level variables are significantly more accurate than those with only area level variables. It therefore seems very likely that EBP unit-context models with PSU-level aggregates would outperform unit-level models when there is a 5 year old mismatch between the survey or census data in Mexico, and a 10 year mismatch in Brazil.

*Finding 2.2: In the presence of selection bias, Fay-Herriot models become disproportionately less accurate.*

Table 11 shows the results for the simulations in which we introduce selection bias into the design-based simulations, as described in section 3.4.2. As expected, the direct estimates are biased and therefore less accurate. In Mexico, the effect of selection bias on the accuracy of the direct estimates is large, as the rank correlation between the direct estimates and the full census falls nearly ten correlation points from 0.881 in 0.783. The latter level of correlation between direct estimates and census values is similar to what has been observed in evaluations of non-monetary welfare indicators, for example in Tanzania in Masaki et al. (2022). The effect of selection bias on the accuracy of direct estimates is smaller in Brazil, partly because all population clusters are sampled in approximately 98 percent of municipalities. Nonetheless, the rank correlation of the direct estimates falls 6.2 percentage points when introducing selection bias in Brazil, from 0.943 to 0.881.

The most striking finding when introducing selection bias is the poor performance in Mexico of the unit-context model with area-level variables. The average rank correlation reported in Table 11 is only 0.8 in Mexico for the unit-context model with area-level predictors, only a bit better than the direct estimates. However, of the remaining approaches, the Fay-Herriot model is the least accurate by a significant margin, with an average rank correlation of 0.929. In contrast, the models that incorporate PSU level predictors achieve a higher average rank correlation, of at least 0.942 (ELL). The EBP unit-context model with PSU level variables has an average rank correlation of 0.950, just a tick below the EBP unit-level model of 0.952. The sub-area model, by a slight margin, is the most accurate on average, with an average rank correlation of 0.953. The results for MAE tell a similar story, with the area unit-context model by far the least accurate, followed by the Fay-Herriot model. The sub-area model again is most accurate, followed closely by the EBP unit and the PSU unit-context model, which are themselves only 0.1 pp apart.

The results for Brazil differ greatly, as introducing selection bias has smaller impacts on the estimates. However, in this case the Fay-Herriot estimates again decline disproportionately. Fay-Herriot models yield the lowest rank correlations, at 0.956, and the second lowest mean absolute errors to unit-context models with area-level predictors.

Overall, the results highlight the large advantage of using methods that can incorporate PSU level variables when selection bias is present at the PSU level, as in the simulations using Mexican data. Because the model uses household and/or PSU-level predictors, it can capture systematic differences between sampled and non-sampled PSUs. The Fay-Herriot and area-level context models, on the other hand, cannot incorporate PSU-level information in the predictors, with negative consequences for accuracy when significant selection bias is present.

*Finding 2.3: The accuracy of Fay-Herriot estimates declines disproportionately as the sample size is reduced.*

So far, the results have only considered one type of sample design, as described in Section 3.2. Although this design was intended to follow the main household surveys used in Mexico and Brazil, many other countries collect smaller household surveys to measure poverty. Smaller sample sizes may degrade the performance of some types of statistical methods and models more than others. We therefore repeated the simulations using three different types of samples: Large, Medium, and Small, as described in Section 3.4.4. These three types of samples vary the number of households sampled in the third stage. All are smaller than the main samples described in Section 3.2 used for all other analysis.

The results are shown in Figure 3 for Mexico and 4 for Brazil. The figures show the average rank correlation of the estimates produced by each method and the truth derived from the census, across the three Mexican rounds (2010, 2015, and 2020) and the two Brazilian rounds (2000 and 2010). Since 100 samples were drawn for each round, the reported averages are taken across 300 samples in Mexico (3 rounds times 100 samples per

round) and 200 samples in Brazil. We do not show results for unit-context models with area-level predictors, which consistently perform poorly in other settings. As expected, performance for all methods declines monotonically as the sample shrinks.

The results indicate a disproportionate decline in the accuracy of the Fay-Herriot estimates as the sample size is reduced. In Mexico, when evaluating using rank correlation, the average correlation for the Fay-Herriot estimates is 0.911 in the smallest sample, moderately below the next lowest of 0.928. The pattern is similar for mean absolute error. The Fay-Herriot estimates have the largest MAE at 7.4 pp in the smallest sample, moderately larger than ELL at 6.7 pp. This pattern is also seen in Brazil, where for the small sample the Fay-Herriot model estimates have an average correlation of 0.953, which is similar with BRF (0.954) and sub-area models (0.955). This is also reflected in the MAE. While the poor performance of the Fay-Herriot model is most apparent in the smallest sample, the Fay-Herriot estimates are the least accurate in all the smaller sample variants, except in the large sample in Brazil where it overtakes BRF. Although Figures 3 and 4 only show the averages across rounds, these same patterns also hold for each individual round.

*Finding 2.4: Variance smoothing for Fay-Herriot models is important in small or selected samples.*

The Fay-Herriot models considered so far follow good practice by smoothing the variance estimates prior to estimation, as described in appendix A. However, variance smoothing is not automatically applied in available software packages and is therefore not always implemented by practitioners. Table 12 compares the accuracy of Fay-Herriot model estimates when smoothing and not smoothing the direct estimates of the smoothing. When using the main sample, using the raw variance estimates has no discernible impact on rank correlations, and modestly increases mean absolute error in Mexico. For example, not smoothing increases mean absolute error from 5.9 to 6.4 pp in 2010 and from 5.9 to 6.7 pp in 2020. However, when using the smallest sample, the impact of variance smoothing is large, par-

ticularly on mean absolute error. When variance smoothing is not implemented in the smallest sample, mean absolute error ranges from 11.5 pp (in Mexico 2015) to 23.7 pp (in Brazil 2010). Similarly, when selection bias is present, accuracy deteriorates greatly in Mexico when using Fay-Herriot models without variance smoothing, with rank correlations ranging from 0.795 (in 2015) to 0.865 (in 2010).

This pattern highlights the importance of variance smoothing in boundary cases where all sample households in an area are either non-poor or poor. In these cases, the estimated variance of the direct estimates is zero, meaning that a Fay-Herriot model without variance smoothing estimates poverty in those areas to be either zero or one hundred percent. Model parameters are estimated under the assumption that these boundary estimates are certain, introducing bias into model estimates. Samples are particularly susceptible to this problem when they are smaller, making boundary estimates in the sample more likely to occur. Similarly, the manner in which we implemented selection bias systematically excluded poor households from the simulated samples, increasing the numbers of areas with no poor households in the samples subject to selection bias. Variance smoothing addresses this issue by ensuring that the assumed variance of the direct estimates when estimating the Fay-Herriot model is strictly positive.

*Finding 2.5: Neither unit-level or unit-context models are dominant when simulating classical measurement error in survey data.*

The results shown so far are based on simulations that assume that the sample and census questions are each measured in the same way and that the sample is collected without error. In practical settings, sample and census data that measure the same concept can vary due to minor differences in wording, differences in respondents, or reporting or processing error. This subsection takes a small step towards relaxing the assumption that the survey and census data are identical for sampled households, by considering a stylized way in which measurement error in survey data can affect the estimates. In particular,

we simulate classical measurement error in the simulated survey in several covariates. In Mexico, we simulate error in three continuous variables: Individual age, completed years of education, and the number of rooms in the house. For all three variables, we add a measurement error term, randomly drawn from a normal distribution with mean zero and standard deviation equal to 10 percent of the standard deviation of the variable in the sample. Any resulting negative values are set to zero. For binary variables, we simulate measurement error by replacing the value with its complement in a randomly selected 2 percent of cases. In Mexico, this is applied to individuals' sex and literacy status, as well as household asset ownership and housing characteristics. In Brazil, the education variable is measured categorically. For the lowest and highest categories, we shift the category by an amount equal to one in a randomly selected two percent of cases. For the middle categories, we shift the variable up one in a randomly selected one percent of cases and down one in another randomly selected one percent of cases.

In this simple scenario, we do not introduce any error into contextual variables, since these are drawn from the census which is assumed to be collected without error. In addition, we do not introduce any error into the dependent variable, although we expect that introducing additional random noise would have similar effects as reducing the size of the sample. Therefore, only the ELL and EBP unit level models are affected by the simulated measurement error, since these are the only specifications that use predictors from the survey. Because of this particular set up, the only impact that measurement error has is to attenuate the coefficient on household level variables in the estimated model towards zero.

As seen in Figures 5 and 6, the introduction of classical measurement error generally has minor impacts, but in some cases affects the relative rankings of different methods. In Mexico, introducing measurement error only reduces correlation by 0.001 and the EBP unit-level model, which even with error outperforms all other models and methods. When it comes to mean absolute error, however, there is a larger impact for the EBP unit-level

model, as the introduction of error leads to a 0.8 pp increase in MAE. This is enough to make the EBP model with error 0.5 pp less accurate than the EBP PSU unit-context model. For Brazil, the simulated measurement error has a stronger impact on rank correlation, reducing correlation for the EBP unit-level model by 0.6 pp. This puts the correlation of the EBP unit-level model with error a bit below the unit-context model. When it comes to mean absolute error in Brazil, however, the EBP unit-level model with error does slightly better than the unit-context model on average, by about 0.1 pp. In short, neither the unit-level nor the unit-context model are dominant under the small amount of classical measurement error introduced in this set of simulations. Allowing for survey measurement error in household predictors that is correlated with the true values of the predictors may tip the balance further in favor of the unit-context model. In addition, allowing for measurement error in household income would harm EBP estimates more than synthetic estimates like ELL and BRF, since EBP estimates would be conditioned on sample data measured with error. On the other hand, if the survey is subject to less measurement error than the census, this could benefit unit-level EBP models, which are less vulnerable to bias in the census data than purely synthetic methods that do not condition on the sample data.

#### 4.4 Adjusting for survey weights

All of the results for unit-level models reported in Table 5 above use rescaled weights, where the weights are normalized in each municipality to sum to the sample size.<sup>16</sup> This avoids a large reduction in the effective sample size across areas, and is a common recommended method for rescaling weights when estimating linear mixed models (StataCorp, 2023). However, practitioners typically use the raw survey sample weights as given, without rescaling. In addition, there are several different methods for incorporating weights when

---

<sup>16</sup>Each area is given equal weight when estimating Fay-Herriot models, making weight rescaling unnecessary

estimating EBP models. This section examines the impact of using the raw weights without rescaling, and of using different weighting methods, on the accuracy of estimates produced using different models and methods. We organize the results into five main findings:

*Finding 3.1: In Mexico, rescaling weights is as or more important than the choice among the set of considered models and methods.*

Table 7 shows that when estimating unit-level models, the failure to rescale weights greatly reduces accuracy in Mexico. In particular, the average rank correlation between the estimates and the "truth" falls from 0.956 when using the rescaled weights to 0.919 when not rescaling the weights, and the mean absolute error increases 30 percent from 5.3 pp to 6.9 pp. Therefore, the unit-level model without weight rescaling is less accurate on average than all other tested methods and models in Mexico that use rescaled weights, with the exception of unit-context models that only use area-level predictors, when using mean absolute error as a measure of accuracy.

Table A5 breaks out accuracy separately by round, and separately for in and out of sample municipalities. The decline in accuracy in Mexico when weights are not rescaled is largest for the 2010 round, followed by 2020 and 2015. The negative impacts of not rescaling are much larger for non-sampled municipalities than municipalities. In all three Mexican rounds, the negative impact of using raw weights on accuracy exceeds the negative impact of using unit-context models, by a large amount in non-sampled areas. The impacts of not rescaling the weights are much smaller in Brazil. Nonetheless, failing to rescale the weights in Brazil moderately affects the accuracy of the ELL estimates, reducing the rank correlation by a point and increasing mean absolute error by 0.4 pp.

Figure 7 sheds light on how sampling error in highly weighted areas in Mexico degrades the accuracy of estimates from unit-level models when weights are not rescaled. The vertical axis indicates municipal-level sampling error, defined as the difference between the estimated municipal average of log per capita income derived from each synthetic sample

and the and the "true" mean obtained from the census. The horizontal axis indicates the log population size of the area. The three lines show a locally smoothed estimate of mean sampling error by log population size, separately according to the accuracy of the poverty estimates derived from that simulated sample. Accuracy is measured by the rank correlation between the poverty estimates obtained using that sample and the "true" poverty values in the census. Thus, when a line passes through -0.2 on the vertical axis, this indicates that the sample underestimated income by approximately 20 percent on average for municipalities with a particular population. The bars in the background indicate the weighted distribution of population size both when weights and rescaled and when they are not.

In Mexico, a clear relationship emerges between the accuracy of the estimates when using raw weights and the extent of sampling error in large municipalities, especially Mexico City which is the largest municipality. In 2000 and 2010, in simulated samples that yield the least accurate estimates, there is large negative sampling error between 0.2 and 0.3 in the most populous municipalities. This is consistent with the informative sample design within areas in Mexico, which tends to omit small, wealthy, PSUs within populous areas like Mexico City from the simulated samples in the second stage of the sample. In 2015, meanwhile, where the accuracy penalty from using raw weights is smaller, there is moderate positive sampling error from 0.05 to 0.1 in the most populous areas. The difference in sampling error in large municipalities between 2015 and the other two rounds in Mexico could be explained by an differences in sampling strategy used by INEGI. When conducting the 2015 intercensus, clusters were sampled with probability proportional to size, whereas the 2010 and 2020 census extracts sampled clusters within municipalities using simple random sampling. Thus, the 2015 intercensus frame was less likely to include small, wealthy PSUs within populous municipalities that caused negative sample error in the 2010 and 2020 rounds when they were omitted from the synthetic samples. Nonetheless, even the

more moderate positive sampling error among the most populous areas in 2015 leads to a large degradation of accuracy in estimates using unit-level models, amounting to an average of 2.3 correlation points for out of sample areas (Table A5).

Sampling error in populous areas is important in Mexico when using raw weights, because of the disproportionate weight given to these municipalities as indicated by the height of the bars on the right of Figure 7. Sampling error among the least populated areas has negligible impact on accuracy because of the small amount of weight these municipalities receive, whether weights are rescaled or not. However, using raw weights makes estimates from linear models less accurate by fitting the model to sampling error in the most populous Mexican areas such as Mexico City. Rescaling redistributes the weights to moderately populated areas, where sampling error is averaged across more municipalities and is minor in all cases. In Brazil, where the second stage of the synthetic samples typically included all population PSUs, and used simple random sampling when necessary, sampling error is minor in both rounds.

Table A6 in the annex shows that using raw weights in 2010 and 2020 substantially increases the variance of the estimated random effect. Ignoring weights altogether, on the other hand, has minor impacts on the accuracy of unit-level model estimates in Mexico, and has a much smaller impact on the variance of the random effect in table A6. This is consistent with the importance of the effective sample size across areas, which is the same when rescaling the weights as it is when ignoring weights, in determining the accuracy of the estimates in Mexico.

Finally, Figure 8 shows a scatterplot for each round. Each point represents a different simulated sample. The horizontal axis indicates the effective sample size across areas while the vertical axis shows the accuracy of EBP unit-level estimates, as measured by the rank correlation with the true census values. The Xes on the left of each graph are estimated using raw sample weights without rescaling, while the dots on the right are estimated using

rescaled weights. In all cases, the Xes are far to the left of the dots, reflecting the reduction in the effective sample size across areas when using raw sample weights. In Mexico, the Xes consistently lie below the dots, consistent with the reduction in accuracy when not rescaling the weights reported in Table 7. Furthermore, the estimates using raw weights in Mexico show much more variability in accuracy across samples than the estimates using rescaled weights. Finally, in Mexico, when weights are not rescaled there is a strong positive relationship across samples between the effective sample size across areas and the accuracy of the estimates derived from that sample. In Brazil, where sampling error is minor, there is no discernible relationship. In sum, using raw weights provided with the survey without rescaling reduces the effective sample size across areas, which in Mexico is associated with markedly less accurate and reliable estimates.

*Finding 3.2: Not rescaling sample weights distorts methodological comparisons in Mexico*

This is particularly noticeable in comparisons between EBP unit-level and unit-context models, and in comparisons between EBP unit-level models, sub-area models, and Boosted Regression Forests. In Mexico, the average rank correlation of unit-context models with PSU variables falls from 0.954 when normalizing the weights to 0.906 when using the raw weights. For unit-context models with area-level variables, the average rank correlation falls further, from 0.934 when normalizing the weights to 0.865 when using the raw weights. (column 1 in Table 7). Thus, not rescaling the weights greatly increases the penalty to accuracy when estimating unit-context models, relative to unit-level models.

Conversely, Boosted Regression Forests (BRF) and sub-area models are much more robust than EBP estimates to not rescaling weights. In table 7, the rank correlation of boosted regression forests falls only 0.3 correlation points, from 94.8 to 94.5, when using raw population weights instead of rescaled weights. This is far smaller than all the other

methods, where the reduction ranges from 2.2 correlation points (sub-area model) to 6.9 correlation points (unit-context model with area-level predictors). This is because the BRF model does not impose a linear functional form, making it less vulnerable to bias stemming from sampling error in the most populous and highly weighted areas. Furthermore, sub-area models are also more robust to using raw sample weights, perhaps because mean poverty is less subject to sampling error in the most populous areas than log per capita income in this case.

#### 4.4.1 Alternative EBP weighting methods

All of the results reported above use the hybrid weighting method to estimate EBP models. Table 8 shows how results change for these models when using three other weighting methods: The partial adjustment method, the nlme method, and the GLS method.

*Finding 3.3: Using the hybrid and GLS weighting methods generally give the most accurate estimates*

This can be seen in table 8. In all cases, differences in accuracy between the hybrid and GLS weighting methods are negligible. While the hybrid and GLS methods generally give the most accurate estimates, there are exceptions. For example, the partial adjustment method estimates have slightly higher rank correlations in Mexico when estimating unit level and unit-context models with PSU-level covariates (0.957 vs 0.955). However, these differences are minor, and the partial adjustment method is also associated with higher mean absolute errors (0.55 vs 0.53 percentage points). In Brazil, the choice of weighting method has minor impacts on the accuracy of the estimates.

*Finding 3.4: Use of the partial weighting method greatly increases mean absolute error when estimating unit-context models in Mexico*

This is also apparent in table 8. When estimating unit-context models with PSU-level aggregates, using partial adjustment weights instead of hybrid weights raises mean absolute error 5.5 pp from 7.2 pp. When estimating unit-context models with area-level aggregates, mean absolute error rises to 10.4 pp. This pattern is not seen in Brazil. These results can be explained by the informative nature of the Mexican sample. In Mexico, the average estimated variance of  $\sigma_v^2$  is 0.034 when using the hybrid weights (results not shown). However, it is only 0.018 when using the partial adjustment weights, which is mechanically equal to the variance component estimates when using no weights. This is a large enough difference to cause substantial upward bias in the estimated poverty rates when using the partial adjustment weighting method. In Brazil, the sample is not informative within municipalities and the magnitude of the estimated area effect is small, as it is only 0.006 when using the hybrid weights and 0.004 when using the partial adjustment method. In the Brazillian context, using the partial adjustment method only leads to a slight upward bias in estimated poverty rates.

## 5 Conclusions

This paper evaluates the comparative performance of different methods of small area poverty estimation under different conditions using unique micro-census data from Brazil and Mexico. What makes these data unique is that they contain a measure of household labor income, which is highly unusual for population census data, allowing us to establish a measure of true poverty rates at the small area level. We adopt design-based simulations in an effort to replicate a variety of empirically relevant conditions.

The main findings are four-fold. First, between estimates derived from unit-level models, EBP is modestly more accurate than ELL estimates in the case of Mexico, while the two approaches are at par (i.e., yield similar levels of accuracy) in the case of Brazil. A key

factor in determining the relative performance of EBP versus ELL, and a major reason why EBP outperforms ELL in Mexico but not in Brazil, is the intraclass correlation coefficient (ICC). The ICC, which is also known as the location effect, is defined as the ratio of the variance of the random effect to the variance of the total error. In Mexico, the ICC is about 6 percent, while in Brazil it is about 1.5 percent. This observation is consistent with [Elbers and van der Weide \(2014\)](#), which finds that ELL is typically at par with EBP when the ICC is below 2.5 percent.

Second, the comparative performance of unit-context models are found to be contingent on a variety of factors. We make an important distinction between unit-context models that include both PSU and target-area level predictors and models that only include area-level predictors. In both cases, the dependent variable is log household income per capita. On average, rank correlations fall by 0.2 percentage points (pp) in Mexico and by 0.4 pp in Brazil when household level predictors are omitted (i.e., when using unit-context models instead of unit-level models). This translates into a minor reduction in the effectiveness of a simulated transfer program in reducing the poverty gap, amounting to 0.01 pp in Mexico and 0.13 pp in Brazil. Dropping sub-area predictors and solely relying on area-level variables as predictors leads to a more notable decline in correlation of about 2 pp in the case of Mexico, which translates to a 0.4 pp decline in the effectiveness of the simulated transfer program. In the presence of measurement error, however, there are instances where unit-context models with sub-area predictors yield more accurate estimates than unit-level models. Unit-context models and area-level models are also found to be more robust to a misalignment of survey and census years, for example when the census pre-dates the household survey by several years.

Third, estimates obtained with models that only include area-level predictors tend to be less accurate than estimates obtained with models featuring sub-area-level and/or household-level predictors. This is not always the case, however, as area-level and unit-

context models are at par when using Brazil’s 2010 data. While the differences in accuracy are generally minor, there are instances where the gains of working with more granular data are substantial. Fay-Herriot models (and unit-context models with area-level predictors) fare more poorly in the presence of sample selection bias at the PSU level, with rank correlations a bit more than 2 points below methods that use PSU level variables. This is because the area-level models can not utilize auxiliary data that has already been aggregated to the area level to correct for systematic differences between sampled and non-sampled PSUs. A similar accuracy penalty for area-level models is observed when the sample size is small due to fewer households being interviewed in each primary sampling unit.

In addition, the loss in accuracy when estimating area-level models in the presence of sample selection and when using small samples rises greatly when variance smoothing is not applied. This is because a larger number of areas in these cases that have zero variance in the poverty estimates derived from the sample, due to all sample households in a target area being poor or non-poor. This highlights the importance of variance smoothing for area-level models when sample poverty rates are either zero or one hundred percent in particular areas.

Finally, in the Mexican case, the treatment of survey weights is a key factor affecting both overall performance and the comparative performance of different approaches to the small area estimation of poverty. For example, the accuracy of estimates from unit-level and unit-context models is greatly reduced in Mexico when weighting areas using sample weights, without rescaling. The use of raw weights without rescaling reduces the effective sample size of areas, magnifying the impact of sampling error within populous areas. Using raw sample weights in this case is found to reduce estimates of accuracy more for unit-context models and unit-level ELL models. Nonetheless, estimates of accuracy for unit-level EBP models, are also negatively affected when not rescaling weights. On the other hand, boosted regression trees, which adopt a more flexible machine learning approach to

estimation, are barely affected. This corroborates several studies that have emphasized the importance of weight rescaling when estimating unit level models (i.e. [Parker et al. \(2023\)](#); [Carle \(2009\)](#); [Rabe-Hesketh and Skrondal \(2006\)](#); [Pfeffermann et al. \(1998\)](#)).

In summary, neither the unit-level EBP model, the unit-level ELL model, the sub-area EBP model, or the unit-context EBP model is found to dominate across all empirically relevant settings considered. We therefore conclude that the comparative performance of the different approaches to the small area estimation of poverty is context-specific. There are two general observations that stand out, however. The first is the importance of rescaling sample weights when estimating unit, unit-context, or sub-area models in cases like Mexico, where the distribution of sample weights across target areas is highly skewed and the sample within target areas is informative. Second, when sample weights are appropriately rescaled and accounted for, there is no downside and potentially significant upside to using more granular predictors.

## References

- Andam, K. S., Ferraro, P. J., Sims, K. R., Healy, A., and Holland, M. B. (2010). Protected areas reduced poverty in costa rica and thailand. *Proceedings of the national academy of sciences*, 107(22):9996–10001.
- Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.
- Baltagi, B. H. (2009). *Econometric Analysis of Panel Data: A Companion to Econometric Analysis of Panel Data*. John Wiley & Sons Incorporated.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.
- Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.
- Bazzi, S. (2017). Wealth heterogeneity and the income elasticity of migration. *American Economic Journal: Applied Economics*, 9(2):219–255.
- Bell, W. R. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, volume 327, page 334. Citeseer.
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.
- Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.
- Besley, T. and Kanbur, R. (1991). The principles of targeting. In *Current issues in development economics*, pages 69–90. Springer.
- Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13:1063–1095.
- Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.
- Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (2013). *Measurement errors in surveys*, volume 548. John Wiley & Sons.
- Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics*, volume 5, pages 3705–3843. Elsevier.
- Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC medical research methodology*, 9:1–13.
- Carter, G. M. and Rolph, J. E. (1974). Empirical bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, 69(348):880–885.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2022). Xgboost: extreme gradient boosting. *R package version 1.6.0.1*.
- Corral, P., Himelein, K., McGee, K., and Molina, I. (2021). A map of the poor or a poor map? *Mathematics*, 9(21):2780.
- Corral, P., Molina, I., Cojocar, A., and Segovia, S. (2022). *Guidelines to small area estimation for poverty mapping*. World Bank Washington.

- Crost, B., Felter, J., and Johnston, P. (2014). Aid under fire: Development projects and civil conflict. *American Economic Review*, 104(6):1833–1856.
- Cuong, N. V. (2012). A method to update poverty maps. *The Journal of Development Studies*, 48(12):1844–1863.
- Das, S. and Chambers, R. (2017). Robust mean-squared error estimation for poverty estimates based on the method of elbers, lanjouw and lanjouw. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(4):1137–1161.
- Das, S. and Haslett, S. (2019). A comparison of methods for poverty estimation in developing countries. *International Statistical Review*, 87(2):368–392.
- Demombynes, G. and Özler, B. (2005). Crime and local inequality in south africa. *Journal of development Economics*, 76(2):265–292.
- Edochie, I., Newhouse, D., Schmid, T., and Wurz, N. (2024a). Povmap: Extension to the emdi package for small area estimation. *available at Comprehensive R Archive Network*.
- Edochie, I., Newhouse, D., Tzavidis, N., Schmid, T., Foster, E., Luna Hernandez, A., Ouedraogo, A., Sanoh, A., and Savadogo, A. (2024b). Small area estimation of poverty in four west african countries by integrating survey and geospatial data. *Journal of Official Statistics*.
- Efron, B. and Morris, C. (1977). Stein’s paradox in statistics. *Scientific American*, 236(5):119–127.
- Elbers, C., Fujii, T., Lanjouw, P., Özler, B., and Yin, W. (2007). Poverty alleviation through geographic targeting: How much does disaggregation help? *Journal of Development Economics*, 83(1):198–213.
- Elbers, C., Lanjouw, J., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.
- Elbers, C. and van der Weide, R. (2014). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. *World Bank Policy Research Working Paper*, (6962).
- Enamorado, T., López-Calva, L. F., Rodríguez-Castelán, C., and Winkler, H. (2016). Income inequality and violent crime: Evidence from mexico’s drug war. *Journal of development economics*, 120:128–143.
- Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277.
- Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52(3):761–766.
- González-Manteiga, W., Lombardía, M., Molina, I., Morales, D., and Santamaría, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate fay-herriot model. *Computational Statistics and Data Analysis*, 52(12):5242–5252.
- Guadarrama, M., Molina, I., and Rao, J. (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics & Data Analysis*, 121:20–40.
- Halbmeier, C., Kreutzmann, A.-K., Schmid, T., and Schröder, C. (2019). The fayherriot

- command for estimating small-area indicators. *The Stata Journal*, 19(3):626–644.
- Haslett, S. (2024). Discussion. *Calcutta Statistical Association Bulletin*, 76(1):39–51.
- Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2):226–252.
- Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.
- Huang, R. and Hidirolou, M. (2003). Design consistent estimators for a mixed linear model on survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association (2003)*, pages 1897–1904.
- Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15:1–96.
- Kanbur, S. (1986). *Budgetary rules for poverty alleviation*. IIES.
- Korn, E. L. and Graubard, B. I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):175–190.
- Kreutzmann, A.-K. (2018). *fayherriot: An R package for estimating empirical best linear unbiased predictors based on the Fay-Herriot model*. Available on github:.
- Lahiri, P. and Salvati, N. (2023). A nested error regression model with high-dimensional parameter for small area estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):212–239.
- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.
- Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101(4):882–902.
- Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., and Engstrom, R. (2022). Small area estimation of non-monetary poverty with geospatial data. *Statistical Journal of the IAOS*, 38(3):1035–1051.
- Merfeld, J. D., Dang, H.-A., and Newhouse, D. L. (2025). Improving estimates of mean welfare and uncertainty in developing countries. Policy Research Working Paper Series 10348, The World Bank.
- Molina, I. (2024). Frontiers in small area estimation research.
- Molina, I. and Marhuenda, Y. (2015). R package sae: Methodology. *The R Journal*, 7(1):81–98.
- Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38(3):369–385.
- Newhouse, D. (2024). Small area estimation of poverty and wealth using geospatial data: What have we learned so far? *Calcutta Statistical Association Bulletin*, 76(1):7–32.
- Newhouse, D., Merfeld, J. D., Ramakrishnan, A. P., Swartz, T., and Lahiri, P. (2025). Small area estimation of monetary poverty in mexico using satellite imagery and machine learning. *Oxford Bulletin of Economist and Statistics*. <https://doi.org/10.1111/obes.12678>.
- Nguyen, M., Corral Rodas, P. A., Azevedo, J. P., and Zhao, Q. (2018). sae: A stata package for unit level small area estimation. *World Bank Policy Research Working Paper*, (8630).

- Parker, P. A., Janicki, R., and Holan, S. H. (2023). A comprehensive overview of unit-level modeling of survey data for small area estimation under informative sampling. *Journal of Survey Statistics and Methodology*, 11(4):829–857.
- Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 60(1):23–40.
- Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In *Handbook of statistics*, volume 29, pages 455–487. Elsevier.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2015). nlme: Linear and nonlinear mixed effects models. R package version 3.1-122.
- Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(4):805–827.
- Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society A*, 183(1):121–148.
- Solon, G., Haider, S. J., and Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human resources*, 50(2):301–316.
- StataCorp, L. (2023). Multilevel mixed-effects reference manual, release 18. *StataCorp LLP, College Station, TX.*, pages 515–517.
- Swamy, P. A. V. B. and Arora, S. S. (1972). The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica: journal of the Econometric Society*, pages 261–275.
- Tarozzi, A. and Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics*, 91(4):773–792.
- Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., Wright, M., and Tibshirani, M. J. (2018). Package ‘grf’. *Comprehensive R Archive Network*.
- Torabi, M. and Rao, J. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, 127:36–55.
- Tzavidis, N., Zhang, L.-C., Luna Hernandez, A., Schmid, T., and Rojas-Perilla, N. (2018). From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society A*, 181(4):927–979.
- Van der Weide, R. (2014). Gls estimation and empirical bayes prediction for linear mixed models with heteroskedasticity and sampling weights: a background study for the povmap project. *World Bank Policy Research Working Paper*, (7028).
- Van Der Weide, R., Blankespoor, B., Elbers, C., and Lanjouw, P. (2024). How accurate is a poverty map based on remote sensing data? an application to malawi. *Journal of Development Economics*, 171:103352.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.
- Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of statistical software*, 77:1–17.
- You, Y. (2022). Small area estimation using fay-herriot area level model with sampling variance smoothing and modeling. *Survey Methodology*, 47(2):361–371.

- You, Y. and Rao, J. (2002). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. *The Canadian Journal of Statistics*, 30(3):431–439.
- Zhao, Q. (2006). User manual for povmap. *World Bank*. [http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Zhao\\_ManualPovMap.pdf](http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Zhao_ManualPovMap.pdf).

# Figures

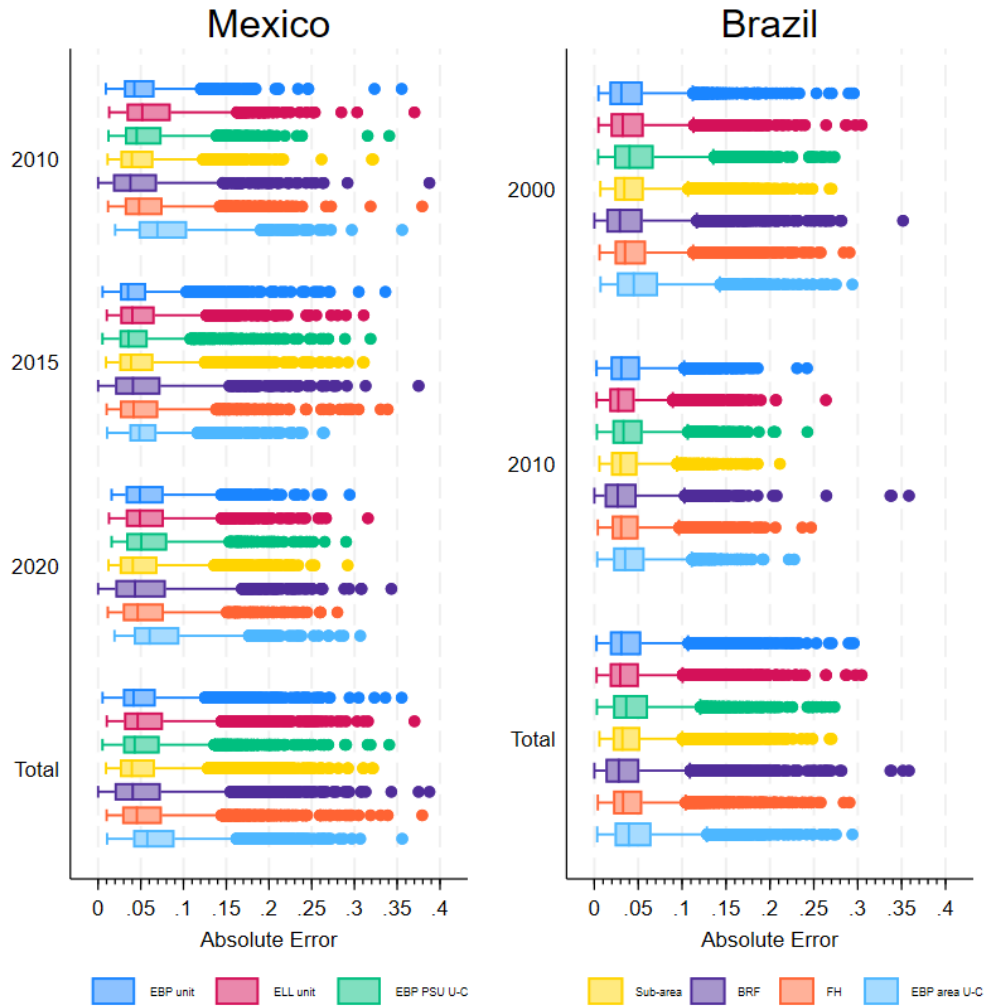


Figure 1: Distribution of absolute error by method, country, and round

Notes: Figure shows box plots of the absolute value of errors in poverty estimates across municipalities, by country and round.

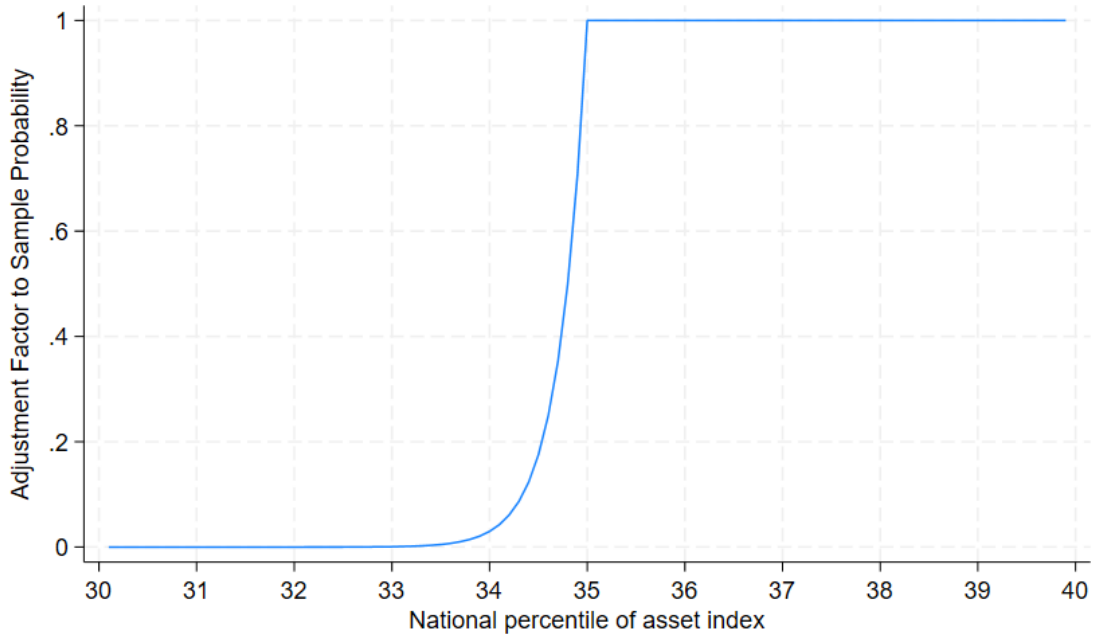
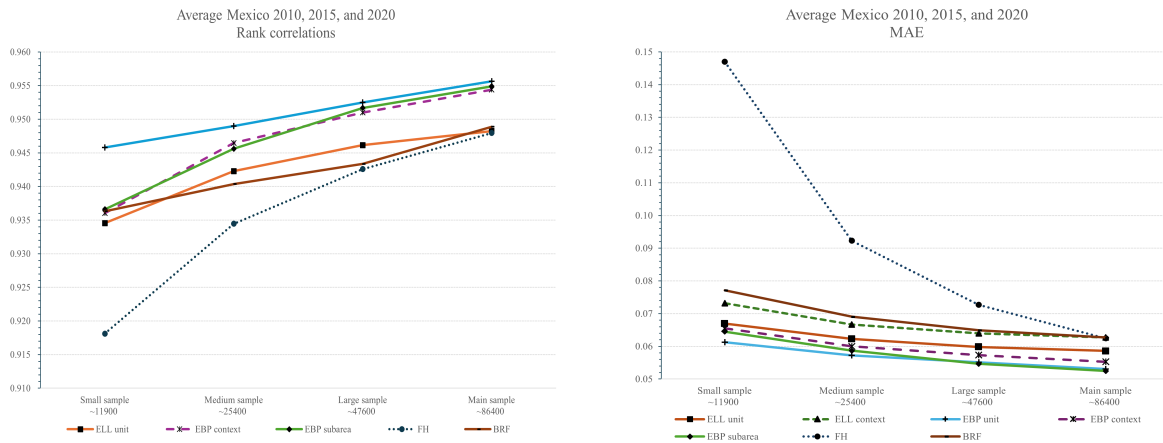


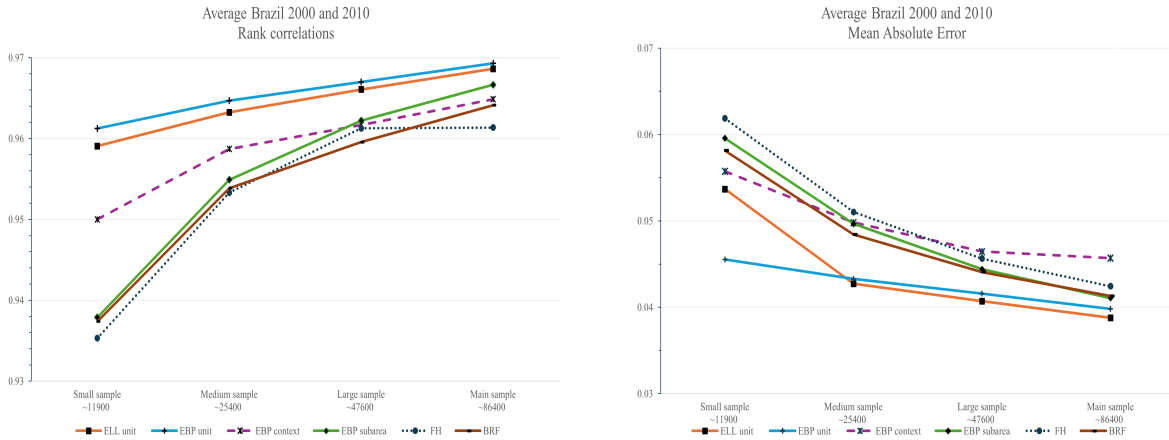
Figure 2: Adjustment factor used to implement selection bias

Figure 3: Performance of different methods and models by sample size: Mexico



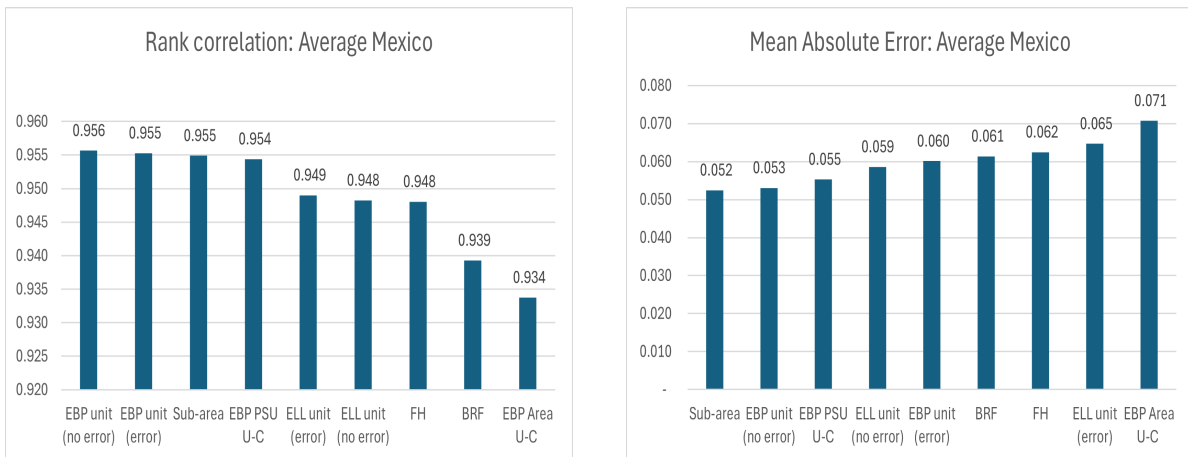
Notes: Figure shows average rank correlations and Mean Absolute Error by method across 300 simulations in Mexico (3 rounds and 100 simulations per round)

Figure 4: Performance of different methods and models by sample size: Brazil



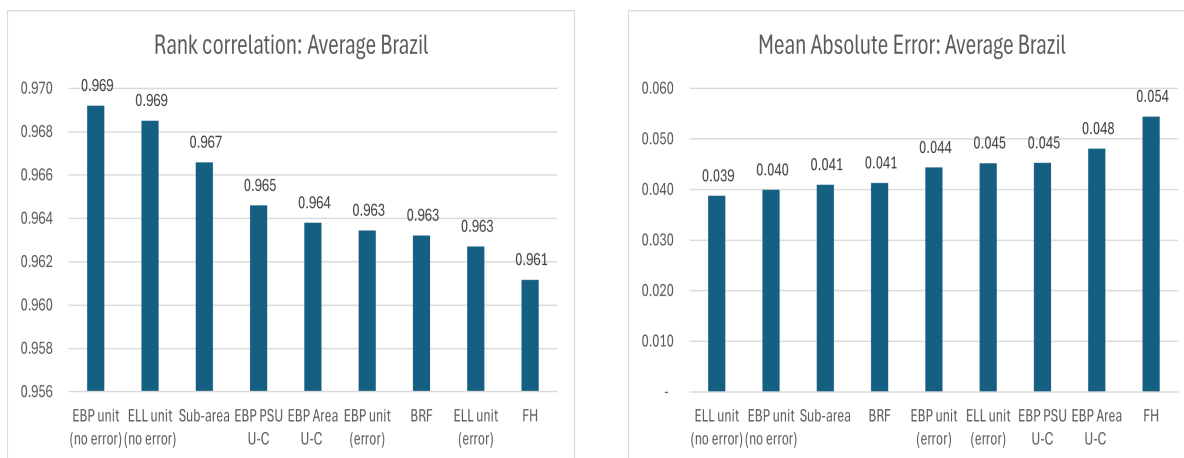
Notes: Figure shows average rank correlations and Mean Absolute Error by method across 200 simulations in Brazil (2 rounds and 100 simulations per round)

Figure 5: Performance of different methods and models with classical measurement error in sample household variables: Mexico



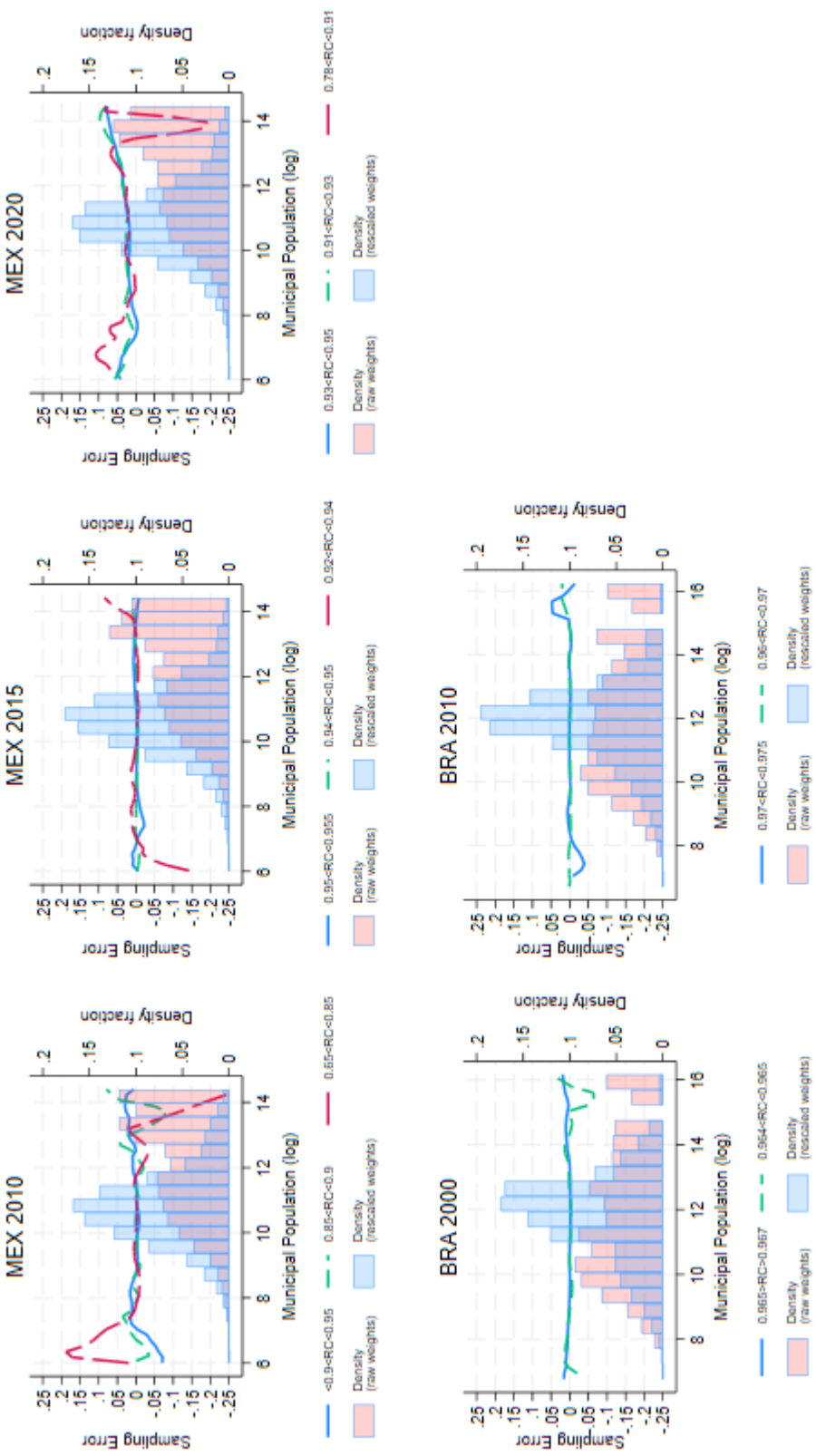
Notes: Figure shows average rank correlations and Mean Absolute Error by method across 300 simulations in Mexico (3 rounds and 100 simulations per round). EBP and ELL unit-level models are estimated both with and without simulated measurement error in household variables added to sample variables. Continuous variables are assumed to have a reliability quotient of 90 percent, and dummy variables are assumed to have a 2 percent chance of incorrect classification.

Figure 6: Performance of different methods and models with classical measurement error in sample household variables: Brazil



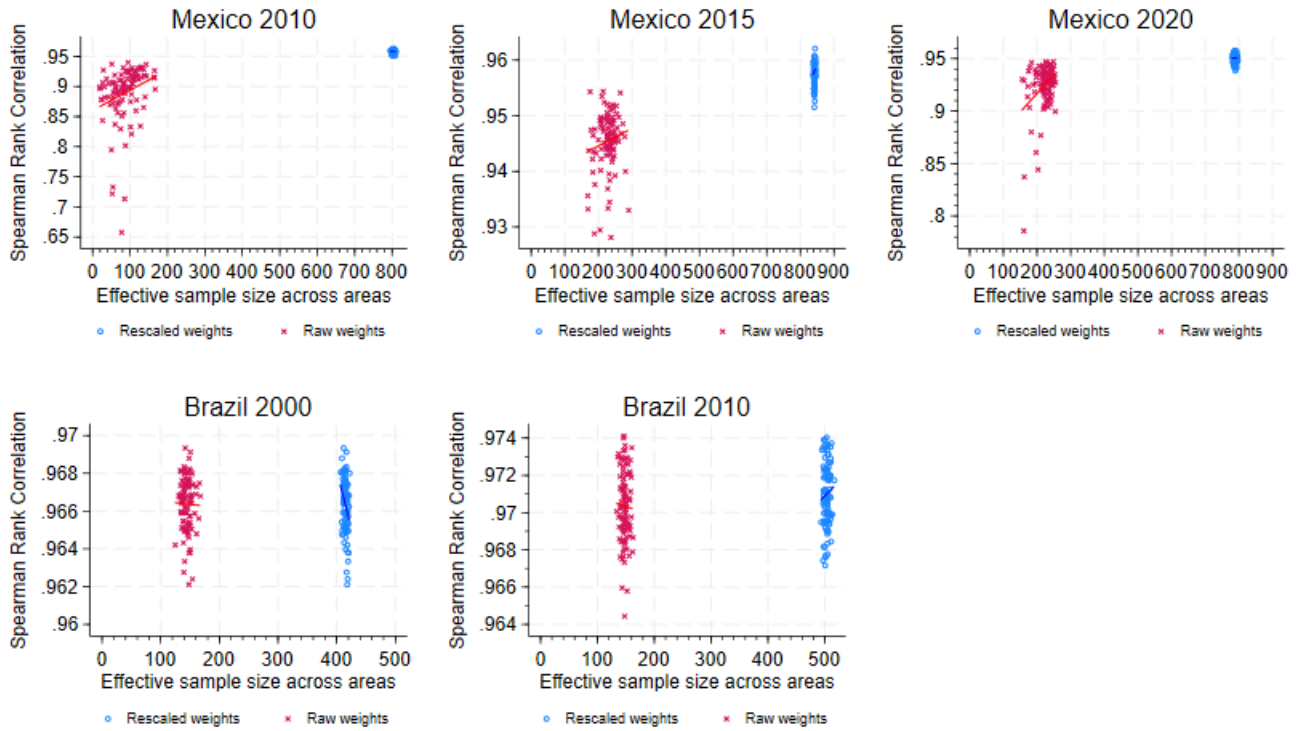
Notes: Figure shows average rank correlations and Mean Absolute Error by method across 200 simulations in Brazil (2 rounds and 100 simulations per round). EBP and ELL unit-level models are estimated both with and without simulated measurement error in household variables added to sample variables. Continuous variables are assumed to have a reliability quotient of 90 percent, and dummy variables are assumed to have a 2 percent chance of incorrect classification.)

Figure 7: Sampling error vs population size across areas and samples, by accuracy of unit-level model estimates and round



Notes: Figure shows locally smoothed estimate of sampling error in mean log per capita income (PCI) versus log population size, across municipalities and simulations. Sampling error is the difference between estimated mean log PCI from the simulated survey and mean log PCI in the census for each municipality. Smoothed estimates are obtained using local polynomial regression weighted by raw sample weights. The solid blue line uses simulated samples that yielded the most accurate poverty estimates when using raw weights, as measured by rank correlation (RC) with the census. The dotted green line represents the least accurate estimates. The red and blue bars show the distribution of population size across municipalities when using raw and rescaled weights, respectively, on the right axis. The figure demonstrates how high levels of sampling error in populous Mexican municipalities generate less accurate estimates when using raw weights.

Figure 8: Accuracy vs effective sample size across areas and samples, by country, round, and weighting method



Notes: Figure shows rank correlation of estimates from EBP unit-level model with household level variables, plotted against the effective sample size across areas. The effective sample size is defined as the sum of sample weights squared, across target areas, divided by the sum of the squared weights. Each point represents a different simulated sample. Points denoted by x represent estimates generated using a model with rescaled weights. Points denoted by a circle represent estimates generated using a model with raw unadjusted sample weights. Solid lines indicate the best linear fit.

# Tables

Table 1: Differences between models

Models	Level	Dep var	Candidate predictors			Functional Form
			HH	PSU	Mun	
Unit-level	HH	Log PCI	X	X	X	Linear
Sub-area context	HH	Log PCI		X	X	Linear
Area context	HH	Log PCI			X	Linear
Sub-area model	PSU	Poverty rate		X	X	Linear
Boosted Regression Forest	PSU	Poverty rate		X	X	Tree-based
Fay-Herriot	Mun	Poverty rate			X	Linear

Table 2: Differences between EBP and ELL methods

	EBP	ELL
Area effect conditioned on sample data	Yes	No
Heteroscedasticity correction ("Alpha model")	No	Yes
Area effect and idiosyncratic error term	Assumed normal	Non-parametric

Notes: See notes to table 5. Average accuracy statistics are shown for cases where the sample weights are rescaled to sum to the sample size for each area (Rescaled), are taken as is with areas effectively weighted by population (Raw), or are not used (None).

Table 3: Size of census extract and simulated samples

	Mexico			Brazil		
	2010	2015	2020	2000	2010	2010
Municipalities	Census extract	2,456	2,457	2,469	5,507	5,565
	Main sample	900	900	900	900	900
	Smaller samples	900	900	900	900	900
PSUs	Census extract	177,957	273,773	252,558	9,336	10,184
	Main sample	7,100	7,200	7,199	2,616	3,271
	Smaller samples	4,452	4,500	4,494	2,628	3,356
Households	Census extract	2,903,640	5,854,392	4,016,627	5,304,711	6,192,332
	Main sample	78,708	106,023	74,361	73,248	91,588
	Large sample	43,297	57,618	41,940	42,048	53,696
	Medium sample	24,564	28,025	23,487	21,024	26,848
	Small sample	12,046	12,306	11,454	10,512	13,424

Table 4: Average diagnostic indicators by EBP specification and country

	Average Mexico			Average Brazil		
	EBP unit	EBP PSU U-C	EBP area U-C	EBP unit	EBP PSU U-C	EBP area U-C
Area effect	Variance	0.034	0.034	0.055	0.006	0.009
	Skewness	-0.383	-0.334	-0.168	-0.202	0.167
	Kurtosis	7.153	6.621	5.165	13.915	19.396
Household error	Variance	0.520	0.636	0.697	0.410	0.906
	Skewness	-0.675	-0.277	-0.134	-0.202	0.167
	Kurtosis	11.721	8.213	7.201	13.915	19.396
ICC	Marginal	0.063	0.050	0.073	0.014	0.009
Unit R2	Cond	0.422	0.239	0.137	0.629	0.190
	Marginal	0.442	0.258	0.179	0.632	0.196
Area R2	Cond	0.817	0.800	0.709	0.941	0.901
		0.919	0.899	0.874	0.960	0.932

Table 5: Country average accuracy results by method, model, measure, and sample status

Accuracy Measure	Auxiliary Data available	Method	Average Mexico (2010, 2015, 2020)		Average Brazil (2000, 2010)	
			Overall	In-sample	Out-of-sample	Overall
RC	None	Direct	0.881		0.943	
	All	EBP unit	0.956	0.968	0.936	0.984
		ELL unit	0.948	0.956	0.928	0.983
	PSU and area	EBP PSU U-C	0.954	0.968	0.934	0.979
		Sub-area	0.955	0.970	0.933	0.983
Area only		BRF	0.948	0.954	0.930	0.981
		FH	0.948	0.959	0.927	0.981
		EBP Area U-C	0.934	0.930	0.917	0.980
	None	Direct	0.079		0.055	
MAE	All	EBP unit	0.053	0.041	0.060	0.028
		ELL unit	0.059	0.048	0.065	0.029
	PSU and area	EBP PSU U-C	0.055	0.043	0.062	0.034
		Sub-area	0.053	0.040	0.060	0.028
		BRF	0.052	0.049	0.054	0.029
Area only		FH	0.057	0.047	0.063	0.030
		EBP Area U-C	0.071	0.061	0.076	0.037

Notes: Table shows average rank correlation (RC) and Mean Absolute Error (MAE) in design-based simulations between estimated and true values across three Mexican censuses (2010, 2015, 2020) and two Brazilian censuses (2000 and 2010). Accuracy statistics for each census are averaged across 100 simulated samples per year and then averaged across years. Statistics for each simulated sample are calculated across municipalities, giving each municipality equal weight. Averages are shown separately for all municipalities, municipalities in the sample, and municipalities not in the sample. Each row represents a different model and method. EBP estimates are generated using the Empirical Best Predictor method (Molina and Rao (2010)) using hybrid weights, ELL estimates are generated following the ELL method Elbers et al. (2003). See Table 1 for details. The second column distinguishes models according to the availability of different types of auxiliary predictor variables. Direct indicates survey-based direct estimates. Unit indicates unit-level models that use candidate predictors at the household, Primary Sampling Unit (PSU) and municipal level. PSU U-C indicates unit-level models, which are unit-level models estimated only using PSU and municipal aggregates as candidate predictors. Area U-C indicates unit-context models using only municipal aggregates as candidate predictors. Sub-area estimates are generated using EBP models of PSU-level poverty rates predicted using PSU and municipal-level aggregates. Similarly, BRF predicts PSU-level poverty rates using PSU and municipal level predictors, but using Boosted Regression Forest models. FH indicates area-level models estimated following Fay and Herriot (1979). See Table 2 for further details.

Table 6: Results of poverty targeting simulation

	Average Mexico	Brazil
Poverty gap		
No program	0.25073	0.20312
EBP unit	0.22221	0.17917
ELL unit	0.22226	0.17917
EBP PSU U-C	0.22221	0.17920
Sub-area	0.22222	0.17921
BRF	0.22230	0.17922
FH	0.22224	0.17926
EBP Area U-C	0.22232	0.17922
Decrease in poverty gap (relative to EBP unit)		
EBP unit	100%	100%
ELL unit	99.83%	100%
EBP PSU U-C	99.99%	99.87%
Sub-area	99.96%	99.85%
BRF	99.69%	99.82%
FH	99.90%	99.66%
EBP Area U-C	99.60%	99.80%

Notes: Table shows headcount and poverty gap rates under a hypothetical program that transfers a fixed amount per capita to the a share of the population equal to the poverty rate in each round. The simulated beneficiaries are determined by ranking municipalities according to the estimates of headcount poverty provided by different methods and specifications. The bottom panel shows the reduction in poverty when targeting based on headcount estimates for each method and specification, as a share of the reduction in poverty achieved when targeting based on estimates produced by EBP unit-level models

Table 7: Country average accuracy results by method, model, measure, and weight rescaling

Weights	Average Mexico (2010, 2015, 2020)			Average Brazil (2000, 2010)		
	Rescaled	Raw	None	Rescaled	Raw	None
RC	0.956	0.919	0.957	0.969	0.968	0.970
EBP unit	0.948	0.868	0.952	0.973	0.964	0.968
ELL unit	0.954	0.906	0.955	0.964	0.963	0.962
EBP PSU U-C	0.955	0.933	0.954	0.967	0.966	0.968
Sub-area	0.948	0.945	0.948	0.963	0.963	0.963
BRF	0.934	0.865	0.923	0.963	0.962	0.960
EBP area U-C	0.053	0.069	0.054	0.040	0.041	0.041
EBP unit	0.059	0.087	0.058	0.039	0.043	0.043
ELL unit	0.055	0.073	0.072	0.046	0.049	0.049
EBP PSU U-C	0.053	0.064	0.053	0.041	0.040	0.040
Sub-area	0.055	0.057	0.056	0.041	0.042	0.042
BRF	0.071	0.088	0.104	0.048	0.046	0.046
EBP area U-C						

Table 8: Country average accuracy results by method, model, measure, and weighting approach

Rank correlation	Average Mexico (2010, 2015, 2020)			Average Brazil (2000, 2010)		
	Unit	U-C	Subarea	Unit	U-C	Subarea
Conditional	0.953	0.951	0.928	0.969	0.964	0.967
Partial adjustment	0.957	0.955	0.925	0.969	0.964	0.966
GLS	0.955	0.954	0.945	0.969	0.965	0.961
Hybrid	0.956	0.954	0.947	0.969	0.964	0.963
Mean Absolute Error						
Conditional	0.0546	0.057	0.074	0.040	0.045	0.041
Partial adjustment	0.0544	0.072	0.104	0.040	0.045	0.041
GLS	0.0532	0.055	0.071	0.040	0.045	0.041
Hybrid	0.0531	0.055	0.071	0.040	0.046	0.041

Table 9: Old survey or census data: Mexico

Survey year	2020	2015	2010	2015	2010	2015	2010	2015	2010	2015	2010	
Census year	2020	2015	2010	2015	2010	2015	2010	2015	2010	2015	2010	
Target year	2020	2020	2020	2015	2015	2015	2020	2020	2020	2015	2015	2020
Available												
Indicator	auxiliary data		Method									
RC	All	EBP unit	0.951	0.936	0.935	0.958	0.948	0.946	0.958	0.941	0.928	0.928
		ELL unit	0.944	0.939	0.926	0.951	0.939	0.933	0.950	0.932	0.916	0.916
	Area only	EBP U-C	0.932	0.925	0.925	0.945	0.939	0.937	0.924	0.912	0.893	0.893
		FH	0.947	0.939	0.938	0.948	0.944	0.941	0.949	0.931	0.925	0.925
MAE	All	EBP unit	0.059	0.068	0.068	0.047	0.054	0.084	0.053	0.063	0.174	0.174
		ELL unit	0.060	0.068	0.068	0.052	0.057	0.137	0.064	0.064	0.168	0.168
	Area only	EBP U-C area	0.073	0.075	0.074	0.057	0.059	0.095	0.082	0.078	0.190	0.190
		FH	0.059	0.064	0.062	0.055	0.055	0.081	0.059	0.069	0.073	0.073

Notes: Table shows age bias results for Mexico. Values represent average rank correlation (RC) and mean absolute error (MAE) across 100 simulations. Each column represent different years of simulated sample and census data used to generate estimates. Each row represent different models and methods (see Tables 1 and 2). Estimates are evaluated against "truth" calculated from the target year census. Accuracy statistics for each simulated sample are calculated across municipalities, giving each municipality equal weight.



Table 11: Accuracy metrics by model and method with PSU-level selection bias

Indicator	Auxiliary Data available	Method	Average Mexico		Average Brazil	
			In-sample	Out-of-sample	In-sample	Out-of-sample
RC	None	Direct	0.783		0.909	
	All	EBP unit	0.967	0.929	0.977	0.960
		ELL unit	0.954	0.917	0.977	0.959
	PSU and area	EBP PSU U-C	0.966	0.927	0.975	0.961
		Sub-area	0.969	0.931	0.980	0.962
		BRF	0.956	0.924	0.977	0.957
	Area only	FH	0.949	0.898	0.976	0.952
		EBP Area U-C	0.859	0.715	0.970	0.956
		Direct	0.109		0.068	
		EBP unit	0.048	0.073	0.035	0.048
MAE	All	ELL unit	0.054	0.080	0.036	0.046
	PSU and area	EBP PSU U-C	0.049	0.074	0.035	0.048
		Sub-area	0.043	0.064	0.030	0.044
		BRF	0.054	0.080	0.030	0.045
	Area only	FH	0.062	0.088	0.036	0.054
		EBP Area U-C	0.100	0.157	0.051	0.060
		Overall				
		Overall				
		Overall				
		Overall				

Notes: Table shows average rank correlation (RC) and Mean Absolute Error (MAE) in design-based simulations between estimated and true values across three Mexican censuses (2010, 2015, 2020) and two Brazilian censuses (2000 and 2010). Selection bias is simulated in the sample and not accounted for when constructing sample weights, as described in section 5.5. Accuracy statistics for each census are averaged across 100 simulated samples per year and then averaged across years. See notes to table 3 for additional details.

Table 12: Accuracy of Fay-Herriot model estimates by country, round, type of sample, measure, and implementation of variance smoothing

RC	Main sample		Small Sample		Selection bias	
	Smoothed	Raw	Smoothed	Raw	Smoothed	Raw
Mexico 2010	0.949	0.949	0.918	0.913	0.930	0.867
Mexico 2015	0.948	0.948	0.917	0.912	0.927	0.795
Mexico 2020	0.947	0.947	0.920	0.908	0.930	0.804
Brazil 2000	0.962	0.961	0.931	0.876	0.961	0.960
Brazil 2010	0.968	0.968	0.940	0.761	0.951	0.961
Mean Absolute Error						
Mexico 2010	0.059	0.064	0.076	0.150	0.075	0.131
Mexico 2015	0.055	0.056	0.071	0.115	0.075	0.137
Mexico 2020	0.059	0.067	0.074	0.175	0.083	0.156
Brazil 2000	0.046	0.046	0.066	0.205	0.048	0.059
Brazil 2010	0.039	0.039	0.058	0.237	0.054	0.066

Notes: Table shows average rank correlation (RC) and Mean Absolute Error (MAE) between estimated and true values for smoothed and raw Fay-Herriot model estimates. Results are taken from design-based simulations across three Mexican censuses (2010, 2015, 2020) and two Brazilian censuses (2000 and 2010). Variance smoothing is conducted following [You \(2022\)](#) as described in annex [A.3.1](#).

Table A1: Poverty lines and national rates by country and year

	Poverty line	National Poverty rate
MEX 2010	962.8	46.0
MEX 2015	1236.5	52.1
MEX 2020	1611.8	50.3
BRA 2000	105.9	46.3
BRA 2010	205.4	31.5

Table A2: Informative sampling diagnostics

	MEX 2010	MEX 2015	MEX 2020	BRA 2000	BRA 2010
Population weight					
Average skewness across areas	12	6	5	17	17
Effective sample size across areas	85	226	220	144	147
Average informative sample test coefficient	89416	2634	2381	294	308
Average within-area informative sample test coef	71196	1850	1621	168	180
Average skewness across areas	-0.3	-1.3	0.1	1.4	1.0
Effective sample size across areas	802	841	788	414	503
Informative sample test coefficient	0.2	0.0	0.0	0.0	0.0
within-area informative sample test coef	0.1	0.0	0.0	0.0	0.0

Table A3: Accuracy results by method, model, round, and measure

	Direct estimates		EBP		ELL		Other	
	Unit	U-C (PSU)	U-C (Area)	Subarea	Unit	FH	BRF	
<b>Rank correlation</b>								
Mexico 2010	0.849	0.958	0.957	0.924	0.959	0.949	0.957	
Mexico 2015	0.923	0.958	0.957	0.945	0.953	0.948	0.946	
Mexico 2020	0.873	0.951	0.949	0.932	0.953	0.947	0.940	
Brazil 2000	0.946	0.966	0.964	0.963	0.963	0.962	0.960	
Brazil 2010	0.939	0.971	0.965	0.963	0.970	0.968	0.966	
<b>Mean Absolute Error</b>								
Mexico 2010	0.092	0.053	0.057	0.082	0.052	0.059	0.052	
Mexico 2015	0.066	0.047	0.047	0.057	0.052	0.055	0.054	
Mexico 2020	0.079	0.059	0.061	0.073	0.054	0.059	0.058	
Brazil 2000	0.059	0.041	0.049	0.053	0.045	0.046	0.044	
Brazil 2010	0.052	0.039	0.041	0.043	0.037	0.039	0.038	

Table A4: Accuracy results for EBP estimates by model, measure, weight rescaling and census year

		Rank correlation		Mean Absolute Error	
		Rescaled	Raw	Rescaled	Raw
MEX 2010	EBP Unit	0.958	0.889	0.053	0.078
	ELL unit	0.951	0.796	0.064	0.114
	U-C PSU	0.957	0.854	0.057	0.091
	Subarea	0.959	0.915	0.051	0.077
	BRF	0.957	0.954	0.052	0.056
MEX 2015	U-C area	0.924	0.788	0.082	0.112
	EBP Unit	0.958	0.945	0.047	0.057
	ELL unit	0.950	0.921	0.052	0.065
	U-C PSU	0.957	0.943	0.047	0.054
	Subarea	0.953	0.942	0.052	0.057
MEX 2020	BRF	0.946	0.944	0.054	0.055
	U-C area	0.945	0.914	0.057	0.068
	Unit	0.951	0.924	0.059	0.072
	ELL unit	0.943	0.887	0.060	0.082
	U-C PSU	0.949	0.920	0.061	0.074
BRA 2000	Subarea	0.953	0.943	0.054	0.059
	BRF	0.941	0.938	0.058	0.060
	U-C area	0.932	0.892	0.073	0.085
	Unit	0.966	0.966	0.041	0.042
	ELL unit	0.970	0.965	0.042	0.043
BRA 2010	U-C PSU	0.962	0.962	0.050	0.050
	Subarea	0.963	0.962	0.045	0.044
	BRF	0.960	0.960	0.044	0.044
	U-C area	0.963	0.961	0.053	0.055
	Unit	0.971	0.970	0.039	0.040
	ELL unit	0.976	0.963	0.035	0.044
	U-C PSU	0.965	0.963	0.041	0.047
	Subarea	0.970	0.970	0.037	0.036
	BRF	0.966	0.966	0.038	0.039
	U-C area	0.965	0.962	0.043	0.038

Table A5: Accuracy results for EBP estimates by model, measure, weight rescaling, census year, and sample status

Measure	Sample status	Weight rescaling	Rank correlation				Mean Absolute Error			
			In-sample		Out-of-sample		In-sample		Out-of-sample	
			Rescaled	Raw	Rescaled	Raw	Rescaled	Raw	Rescaled	Raw
MEX 2010	EBP Unit	0.975	0.950	0.936	0.828	0.042	0.052	0.059	0.093	
	ELL unit	0.962	0.828	0.926	0.736	0.052	0.096	0.072	0.125	
	U-C PSU	0.973	0.928	0.934	0.780	0.065	0.063	0.088	0.107	
	Subarea	0.973	0.932	0.935	0.877	0.041	0.064	0.057	0.084	
	BRF	0.958	0.951	0.943	0.940	0.049	0.057	0.054	0.056	
	U-C area	0.912	0.884	0.906	0.690	0.073	0.080	0.087	0.131	
MEX 2015	EBP Unit	0.978	0.975	0.938	0.915	0.035	0.036	0.053	0.063	
	ELL unit	0.962	0.932	0.931	0.894	0.043	0.057	0.057	0.070	
	U-C PSU	0.978	0.974	0.936	0.912	0.034	0.038	0.054	0.064	
	Subarea	0.976	0.969	0.926	0.910	0.037	0.041	0.061	0.067	
	BRF	0.962	0.957	0.923	0.923	0.043	0.046	0.060	0.061	
	U-C area	0.955	0.950	0.926	0.874	0.049	0.051	0.062	0.077	
MEX 2020	Unit	0.952	0.942	0.935	0.895	0.048	0.057	0.066	0.081	
	ELL unit	0.943	0.887	0.929	0.859	0.050	0.071	0.066	0.088	
	U-C PSU	0.952	0.939	0.933	0.888	0.049	0.058	0.068	0.084	
	Subarea	0.960	0.949	0.936	0.926	0.042	0.047	0.061	0.066	
	BRF	0.944	0.938	0.925	0.923	0.047	0.050	0.064	0.066	
	U-C area	0.923	0.914	0.918	0.853	0.061	0.067	0.080	0.095	
BRA 2000	Unit	0.987	0.987	0.961	0.961	0.029	0.029	0.045	0.045	
	ELL unit	0.984	0.982	0.962	0.960	0.031	0.033	0.045	0.045	
	U-C PSU	0.981	0.981	0.957	0.957	0.039	0.039	0.053	0.053	
	Subarea	0.984	0.983	0.957	0.957	0.030	0.031	0.047	0.047	
	BRF	0.981	0.980	0.955	0.954	0.031	0.032	0.047	0.047	
	U-C area	0.980	0.979	0.958	0.956	0.042	0.043	0.056	0.057	
BRA 2010	Unit	0.982	0.981	0.969	0.968	0.027	0.028	0.041	0.043	
	ELL unit	0.981	0.980	0.968	0.968	0.026	0.026	0.037	0.037	
	U-C PSU	0.975	0.973	0.964	0.961	0.029	0.031	0.043	0.046	
	Subarea	0.983	0.980	0.968	0.967	0.025	0.028	0.040	0.040	
	BRF	0.980	0.978	0.964	0.963	0.027	0.028	0.040	0.041	
	U-C area	0.969	0.967	0.964	0.962	0.031	0.034	0.045	0.049	

Table A6: Variance of random effect and Intra-Cluster Coefficient (ICC) by weight rescaling

		Variance of random effect			ICC		
		Rescaled weights	Raw weights	No weights	Rescaled weights	Raw weights	No weights
MEX 2010	EBP unit	0.035	0.081	0.017	0.071	0.166	0.033
	U_C PSU	0.038	0.102	0.015	0.058	0.145	0.022
	U-C area	0.069	0.157	0.044	0.092	0.198	0.055
MEX 2015	EBP unit	0.024	0.025	0.017	0.054	0.059	0.038
	U_C PSU	0.023	0.024	0.016	0.040	0.044	0.026
	U-C area	0.039	0.046	0.033	0.059	0.069	0.046
MEX 2020	EBP unit	0.042	0.055	0.024	0.063	0.062	0.035
	U_C PSU	0.041	0.055	0.023	0.052	0.054	0.028
	U-C area	0.058	0.072	0.040	0.069	0.067	0.044
BRA 2000	EBP unit	0.015	0.015	0.004	0.034	0.033	0.010
	U_C PSU	0.020	0.020	0.005	0.023	0.023	0.005
	U-C area	0.010	0.021	0.009	0.010	0.020	0.009
BRA 2010	EBP unit	0.005	0.011	0.003	0.012	0.024	0.007
	U_C PSU	0.006	0.016	0.006	0.008	0.018	0.008
	U-C area	0.007	0.016	0.006	0.008	0.017	0.006

Table A7: Accuracy results for EBP estimates by model, measure, weighting approach, and census year

		Rank Correlation			Mean Absolute Error				
		Conditional	Guadarrama	GLS	Hybrid	Conditional	Guadarrama	GLS	Hybrid
MEX 2010	Unit	0.955	0.960	0.958	0.958	0.055	0.055	0.053	0.053
	U-C PSU	0.952	0.958	0.956	0.957	0.060	0.080	0.057	0.057
	U-C area	0.914	0.907	0.924	0.924	0.087	0.127	0.081	0.082
MEX 2015	Subarea	0.955	0.955	0.959	0.959	0.055	0.051	0.051	0.051
	Unit	0.957	0.959	0.958	0.958	0.048	0.047	0.047	0.047
	U-C PSU	0.956	0.958	0.957	0.957	0.048	0.062	0.047	0.047
MEX 2020	U-C area	0.942	0.942	0.945	0.945	0.059	0.086	0.057	0.057
	Subarea	0.952	0.953	0.953	0.953	0.053	0.052	0.052	0.052
	Unit	0.946	0.951	0.949	0.951	0.061	0.062	0.060	0.059
BRA 2000	U-C PSU	0.944	0.949	0.948	0.949	0.063	0.074	0.062	0.061
	U-C area	0.924	0.926	0.930	0.932	0.076	0.099	0.073	0.073
	Subarea	0.951	0.953	0.953	0.953	0.056	0.054	0.054	0.054
BRA 2010	Unit	0.967	0.967	0.968	0.967	0.041	0.041	0.041	0.041
	U-C PSU	0.963	0.962	0.964	0.962	0.050	0.050	0.049	0.050
	U-C area	0.962	0.962	0.963	0.962	0.053	0.054	0.053	0.053
BRA 2015	Subarea	0.962	0.962	0.957	0.962	0.045	0.045	0.063	0.045
	Unit	0.971	0.971	0.971	0.971	0.039	0.038	0.039	0.038
	U-C PSU	0.966	0.966	0.965	0.966	0.041	0.041	0.041	0.041
BRA 2020	U-C area	0.966	0.966	0.965	0.966	0.042	0.042	0.043	0.042
	Subarea	0.971	0.971	0.964	0.965	0.037	0.038	0.043	0.038

# A Annex A: Description of methods

## A.1 Models with household-level predictors

Within the set of specifications that utilize all available predictors including household characteristics, we evaluate two methods: The Empirical Best Predictor (EBP) method under the assumption of normally distributed errors, and the ELL method.

### A.1.1 Unit-level Empirical Best Predictor (EBP)

We first consider the following log-linear mixed model:

$$\ln y_{ij} = x'_{ij}\beta + u_i + \epsilon_{ij} \quad (6)$$

where  $i$  represents the municipality (i.e., target area),  $j$  represents the household, and  $\ln y_{ij}$  is the log per capita household income<sup>17</sup>;  $x'_{ij}$  is a  $p + 1$  row vector containing the constant and  $p$  predictor variables, which can contain household, PSU, and municipal level variables in this specification;  $u_i$  and  $\epsilon_{ij}$  are independent normal random variables with mean 0 and variances  $\sigma_u^2$  and  $\sigma_\epsilon^2$ , respectively.  $\beta$ ,  $\sigma_u^2$ , and  $\sigma_\epsilon^2$  are model parameters estimated using Maximum Likelihood Estimation.

The Best Predictor (BP) of the expected log per capita income for each household under model (1) is given by:

$$\ln \hat{y}_{ij}^{bp} = x'_{ij}\beta + b_i, \quad (7)$$

where  $b_i = \gamma_i (\bar{y}_{ij} - \bar{x}'_{ij}\beta)$ ,  $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \delta_i \sigma_\epsilon^2}$ , and  $\delta_i = \frac{\sum_{j=1}^{n_i} w_{ij}^2}{(\sum_{h=1}^{n_i} w_{ih})^2}$ , and where  $w_{ij}$  represents the sample weights for household  $j$  in municipality  $i$ . We normalize  $w_{ij}$  to sum to  $n_i$ , the number of sample households in each municipality  $i$ , following rescaling method 2 proposed in Pfeffermann et al. (1998). Below in section 4, we report how results change when the weights are not normalized in this way. We obtain the Empirical Best Predictor (EBP) of  $\ln \hat{y}_{ij}^{EBP}$  from equation 7 by replacing the model parameters  $\beta, \sigma_u^2$ , and  $\sigma_\epsilon^2$  with their estimators  $\hat{\beta}$ ,  $\hat{\sigma}_u^2$ , and  $\hat{\sigma}_\epsilon^2$  respectively.

The EBP is a shrinkage estimator, where  $\hat{\gamma}_i$  represents the extent of shrinkage towards the direct estimate relative to the synthetic model estimate  $x'_{ij}\hat{\beta}$ . At one extreme, when the synthetic model fully explains across-municipality variation in  $y_{ij}$ ,  $\sigma_u^2$  and the shrinkage parameter  $\gamma_i$  will equal 0, and the Best Predictor will equal the synthetic model estimate  $x'_{ij}\hat{\beta}$ . Conversely at the other extreme, if all sample households have the same per capita income in an area, the direct estimates fully explain the variation in  $y_{ij}$ . In that case,  $\sigma_\epsilon^2$

---

<sup>17</sup>Other transformations of per capita income besides the natural logarithm are available in the povmap and EMDI R packages (Edochie et al. (2024a), Kreutzmann (2018)), and can improve on the log in terms of achieving normality (Rojas-Perilla et al. (2020), Tzavidis et al. (2018)). We only consider the log transformation in this evaluation, however, primarily because it remains popular and has always been used when implementing the ELL method.

will equal 0,  $\gamma_i$  will equal 1, and the Best Predictor will equal the direct estimate. As  $\sigma_\epsilon^2$  increases, indicating less accurate direct estimates,  $\gamma_i$  declines, giving more weight to the synthetic model predictions relative to the direct estimates. More weight is also given to the synthetic model estimates as the effective sample size declines, either due to a reduction in the sample size or an increase in the heterogeneity in the sample weights, all else equal. This is reflected in an increase in  $\delta_i$  and a corresponding decline in  $\gamma_i$ . In non-sampled areas, there is no survey data to condition the random effect on. As a result, there is no shrinkage and  $\gamma_i = b_i = 0$  for these areas.

The best predictor of the poverty rate for each area can be written as:

$$E \left[ P_i^{bp} \right] \approx \frac{1}{\sum_{j=1}^{N_i} w_{ij}^p} \sum_{j=1}^{N_i} w_{ij}^p \Phi \left( \frac{\ln(Z) - x'_{ij} \beta - b_i}{\sqrt{(1 - \gamma_i) \sigma_u^2 + \sigma_\epsilon^2}} \right), \quad (8)$$

where  $P_i^{ebp}$  is the empirical best predictor of the poverty rate in area  $i$ ,  $N_i$  is the number of census households in area  $i$ ,  $\Phi$  is defined as the standard normal cumulative distribution function,  $Z$  represents the poverty line, and  $w_{ij}^p$  represents the population weight used for averaging across households, which in this case is equal to household size. It is important to distinguish  $w_{ij}^p$ , the population weights, from sample weights  $w_{ij}$ . The former represents population weights used when aggregating from units to target areas to generate area-level estimates, while the latter represent sample weights typically used to adjust for differences in households' sample selection probabilities.

To estimate municipal poverty rates, we take repeated Monte-Carlo draws of  $u_i$  and  $\epsilon_{ij}$  in order to simulate household welfare multiple times. Because it is an empirical best linear unbiased prediction, we substitute  $\hat{\beta}, \hat{b}_i, \hat{\sigma}_u^2$  and  $\hat{\sigma}_\epsilon^2$  for  $\beta, b_i, \sigma_u^2$ , and  $\sigma_\epsilon^2$ , assuming that these are fixed across simulations. Therefore,  $u_i$  and  $\epsilon_{ij}$  are assumed to be normal with mean zero and variance  $\hat{\sigma}_u^2$  and  $\hat{\sigma}_\epsilon^2$  respectively. The employed Monte-Carlo approach with 100 simulations is similar to the Monte-Carlo simulation approach adopted by ELL. For each simulation, all households are classified as poor or not poor depending on whether their simulated welfare exceeds the poverty line. Poverty rates are then averaged across the 100 simulations and all households in each area, as follows:

$$\hat{P}_i^{ebp} = \frac{1}{\sum_{j=1}^{N_i} w_{ij}^p} \sum_{j=1}^{N_i} \frac{w_{ij}^p}{100} \sum_{l=1}^{100} I \left( x'_{ij} \hat{\beta} + \hat{b}_i + u_i^l + \epsilon_{ij}^l < \ln(Z) \right) \quad (9)$$

where  $u_i^l$  is the  $l^{th}$  draw from a normal distribution with mean 0 and variance  $(1 - \hat{\gamma}_i) \hat{\sigma}_u^2$ ,  $\epsilon_{ij}^l$  is the  $l^{th}$  draw from a normal distribution with mean 0 and variance  $\hat{\sigma}_\epsilon^2$ , and  $w_{ij}^p$  is the population weight specified for household  $h$ , and  $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \delta_i \hat{\sigma}_\epsilon^2}$ . As mentioned above,  $\gamma_i = b_i = 0$  for target areas not included in the sample. This Monte-Carlo approach can also be used to generate other statistics derived from the distribution of welfare, such as the poverty gap and severity, inequality measures, and the percentiles of the area welfare distribution.

While we do not evaluate uncertainty estimates, the procedure for estimating uncer-

tainty is quite different for EBP than for ELL. In particular, because the estimated random effect is conditioned on the sample data, it is standard when estimating EBP models to utilize a parametric bootstrap approach to estimating uncertainty. This entails using the estimated model to repeatedly generate one hundred simulated welfare values for the entire population. The model is then re-estimated using the sample households, with their new simulated welfare values. The resulting EBP predictions are then compared with the "truth" derived from the simulated welfare values from the population to estimate the Mean Squared Error (MSE) of the estimates for each target area. Finally, it is standard when estimating uncertainty to use the square root of the MSE as an estimate of the standard error, under the assumption that the point estimates of poverty are unbiased. Further details can be found in [González-Manteiga et al. \(2008\)](#) and [Tzavidis et al. \(2018\)](#).

### A.1.2 The ELL method

The ELL method is an alternative unit-level model that is based on the following estimating equation:

$$\ln y_{ij} = x'_{ij}\beta + u_i + \varepsilon_{ij}, \quad (10)$$

where  $u_i$  and  $\varepsilon_{ij}$  are independent random variables with mean 0 and variance  $\sigma_u^2$  and  $\sigma_{\varepsilon,ij}^2$  respectively. Thus the error term is assumed to be heteroscedastic, and equation 10 is therefore estimated using Generalized Least Squares. The variance of  $\sigma_{\varepsilon,i}^2$  is estimated using OLS, following [Zhao \(2006\)](#)

$$\ln \left( \frac{e_{ij}^2}{A - e_{ij}^2} \right) = Z'_{ij}\alpha + r_{ij} \quad (11)$$

where  $e_{ij} = \ln y_{ij} - x'_{ij}\hat{\beta}$ ,  $Z_{ij}$  is a vector of explanatory variables, and  $A = 1.05 * \max(e_{ij}^2)$ . We select predictor variables  $Z_{ij}$  for the alpha model using the Least Absolute Shrinkage and Selection Operator (LASSO), applied to the same candidate variables used for selecting predictors for the EBP and ELL "beta" models used to predict log per capita income.<sup>18</sup> No further distributional assumptions are made, i.e., the errors need not be normally distributed. Model parameters  $\hat{\beta}$ ,  $\hat{\sigma}_u^2$ , and  $\hat{\sigma}_\varepsilon^2$  are estimated following the method described in [Swamy and Arora \(1972\)](#).

Poverty rates are estimated by drawing  $u_i^l$  and  $\varepsilon_{ij}^l$  100 times from their estimated distributions, and calculating area poverty estimates as follows:

$$\hat{P}_i^{ELL} = \frac{1}{\sum_{j=1}^{N_i} w_{ij}^p} \sum_{j=1}^{N_i} \frac{w_{ij}^p}{100} \sum_{l=1}^{100} I \left( x'_{ij}\hat{\beta} + u_i^l + \varepsilon_{ij}^l < \ln(Z) \right) \quad (12)$$

Our implementation of ELL differs from the original implementation in [Elbers et al. \(2003\)](#), to make the ELL and EBP estimates more comparable. In particular, we assume that the

---

<sup>18</sup>One can also use predicted income from the "beta" model as a predictor in the "alpha model", but we took the simpler approach of using the same candidate predictor variables.

estimated model parameters  $\hat{\beta}, \hat{\sigma}_u^2$ , and  $\hat{\sigma}_\epsilon^2$  are fixed across simulations. In addition,  $u_i$  is specified at the target area level  $i$  instead of the PSU level  $c$ . While we don't consider estimates of uncertainty in this analysis, specifying the random effect at the area level rather than the PSU level prevents the underestimation of standard errors, as documented by [Tarozzi and Deaton \(2009\)](#) and [Das and Chambers \(2017\)](#).<sup>19</sup>

The ELL method differs in theory and implementation from the EBP method in three key ways, as listed in Table 1.

1. ELL, unlike EBP, is based on what econometricians commonly refer to as a random effect model ([Baltagi \(2009\)](#), [Wooldridge \(2010\)](#)) rather than a mixed effects model. The random effects model is a restricted version of the mixed effects model used for EBP, with  $b_i = 0 \forall i$ . Therefore the random effect in this model, unlike the mixed effects model, is not conditioned on the sample data and is not a shrinkage estimator. As a result, the ELL and EBP estimators are very similar for out-of-sample areas ([Tzavidis et al. \(2018\)](#)). Because ELL is purely synthetic, unlike EBP it is not design-consistent for sampled areas, in the sense that ELL estimates do not converge to the population "truth" as the sample becomes arbitrarily large.

2. The ELL implementation includes a heteroscedasticity correction, based on the "alpha model" of the variance of the residuals described above, following [Elbers et al. \(2003\)](#) and [Zhao \(2006\)](#).

3. When estimating ELL, we allow for non-normal error terms. This is done by drawing from the empirical distribution of both the estimated area random effects and the household error terms. In particular, for each target area we decompose the residual into its weighted mean and the deviation from that weighted mean, as follows:

$$e_{ij} = \left( y_{ij} - x'_{ij} \hat{\beta} \right) = \bar{e}_i + \tilde{e}_{ij}, \text{ where:}$$

$$\bar{e}_i = \frac{\sum_{j=1}^{m_i} w_{ij} (y_{ij} - x'_{ij} \hat{\beta})}{\sum_{j=1}^{m_i} w_{ij}}, \text{ and}$$

$$\tilde{e}_{ij} = \left( y_{ij} - x'_{ij} \hat{\beta} \right) - \bar{e}_i$$

For each simulation and area, a value of  $u_i$  is drawn from the empirical distribution of  $\bar{e}_i$  with replacement. Similarly, for each simulation and household, a value of  $\epsilon_{ij}$  is drawn from the empirical distribution of  $\tilde{e}_{ij}$  with replacement. Thus, the simulated values of the random effect  $u$  and the idiosyncratic error term  $\epsilon_{ij}$  are drawn from their empirical distributions, avoiding the assumption of normality.<sup>20</sup>

---

<sup>19</sup>[Tarozzi and Deaton \(2009\)](#) report underestimated standard errors when evaluating the traditional ELL specification with the random effect specified at the PSU level, but attributes that to unmodeled heterogeneity in the model. [Das and Chambers \(2017\)](#) correctly identify the cause as a failure to account for the positive correlation in welfare between PSUs within municipalities when including random effects at the PSU level.

<sup>20</sup>Statistical tests often reject tests for normality in practical settings, although the extent to which this affects the accuracy of the estimates is context-specific.

## A.2 Models with sub-area level predictors

### A.2.1 Unit-context EBP model with sub-area level predictors

The unit-context EBP model with PSU and area-level predictors is nearly identical to the EBP model with household variables described in section A.1.1, with one important difference: All household variables are omitted from the set of predictors. Therefore the model estimates:

$$\ln Y_{ij} = x'_{ij}\beta + u_i + \epsilon_{ij} \quad (13)$$

where all parameters are defined as section A.1.1, except that  $x'_{ij}$  now consists only of PSU and municipal level predictor variables and the constant.

### A.2.2 Sub-area EBP model

The sub-area model is mechanically similar to a unit-level model, except it is specified at the sub-area level, with sub-area level predictions then aggregated to the area level. The predictor variables may include both sub-area (e.g., PSU level) and area-level predictors. Sub-area estimates of poverty in this case are aggregated to the target area level (i.e., municipal level). We refer to this as a “sub-area model” because the specification is similar to the sub-area model presented in Torabi and Rao (2014), Merfeld et al. (2025), and Van Der Weide et al. (2024).<sup>21</sup> The unit of analysis is the PSU and the model is specified as follows:

$$\hat{P}_{ic}^{dir} = x'_{ic}\beta + u_i + \epsilon_{ic}, \quad (14)$$

where  $\hat{P}_{ic}^{dir}$  is the direct estimate of poverty for PSU  $c$  from the survey, and  $x_{ic}$  is a vector of PSU and area aggregate predictors. The error term  $\epsilon_{ic}$  can further be decomposed into model error and sampling error as follows:

$$\epsilon_{ic} = \psi_{ic} + \omega_{ic} \quad (15)$$

where  $\psi_{ic} \sim N(0, \sigma_{\psi}^2)$  represents model error and  $\omega_{ic} \sim N(0, \sigma_{p,ic}^2)$  represents the sampling error associated with the direct estimate of poverty derived from the sample,  $\hat{P}_{ic}$ . Since  $\sigma_{\psi}^2$  and  $\sigma_{p,ic}^2$  cannot be separately identified, we ignore the heteroscedasticity in  $\epsilon_{ic}$  due to sampling error.

Poverty estimates for each PSU are aggregated to municipalities by taking a weighted average across PSUs, using the PSUs’ population sizes as weights.<sup>22</sup> The statistical method is otherwise identical to the household EBP model presented above, and the random effect  $u_i$  is conditioned on the sample data. Because this model predicts mean poverty directly using a linear model, it is less reliant on accurate estimates of the variance components. In addition, unlike models predicting log per capita income, there is no need to exclude

<sup>21</sup>The model proposed in Torabi and Rao (2014) allows for area-specific variance components, while the model in Van Der Weide et al. (2024) allows for spatially correlated random effects

<sup>22</sup>No transformation is used in this specification.

sample households with zero per capita income when estimating the model.

### A.2.3 Boosted Regression Forests

An alternative method that only uses PSU and area-level models is the Boosted Regression Forests (BRF), implemented in the GRF package for R (Tibshirani et al. (2018)) and described in the online documentation to that package as well as in Athey et al. (2019). BRF is a machine-learning prediction method that is very similar to Extreme Gradient Boosting, commonly known as XGboost (Chen et al. (2022)).<sup>23</sup> BRF estimates a series of regression forests that successively predict the residuals from the previous round. Each regression forest consists of a set of decision trees that the algorithm grows on randomly selected subsets of the data, as described in Biau and Scornet (2016), and Wright and Ziegler (2017). The assumed data generating process for each forest is described in Biau (2012).

We implement BRF to estimate a tree-based model to predict direct estimates of poverty at the PSU level:

$$\hat{P}_{ic}^{dir} = g(x'_{ic}) + \epsilon_{ic}, \quad (16)$$

where  $x_{ic}$  contains selected aggregate predictors variables at the PSU and municipal level, selected using LASSO, taken from the census and merged with the survey. No assumptions are made regarding  $\epsilon_{ic}$ . As with EBP, we use rescaled weights in the estimation, as discussed in section ?? below. Furthermore, we do not estimate "honest trees", in order to maximize predictive performance in small samples (Tibshirani et al., 2018). Finally, we tune all parameters using cross-validation, using the tuning option provided in the GRF package.

After estimating the model, we apply it to the the census auxiliary data to obtain predictions in each PSU:  $\hat{P}_{ic}^{BRF} = \hat{g}(x'_{ic})$  where  $\hat{g}$  denotes the estimated model,  $c$  denotes the PSU, and  $x'_{ic}$  is the vector of selected PSU and area-level predictors taken from the census. As in the sub-area model, predictions are aggregated to the area-level by taking the population weighted-average of PSU-level predictions for each municipality. Therefore,  $\hat{P}_i^{BRF} = \frac{\sum_{c=1}^{C_i} w_{ic}^p \hat{P}_{ic}^{BRF}}{\sum_{c=1}^{C_i} w_{ic}^p}$ , where  $C_i$  is the number of PSUs in area  $i$  present in the census, and  $w_{ic}^p = \sum_{h=1}^{M_{c,i}} w_{ij}^p$  is the number of people residing in PSU  $c$  in area  $i$  according to the census.

## A.3 Models with target area-level predictors

### A.3.1 The Fay-Herriot method

The Fay-Herriot model is an area-level model that assumes the following data generating process:

$$P_i = x'_i \beta + u_i \quad (17)$$

---

<sup>23</sup>BRF offers the "honesty" option to use one subsample of the data to grow trees and another to generate predictions at the leaves of trees. We decline to use this option due to the small size of the training data

where  $P_i$  is the poverty rate of area  $i$ ,  $x_i$  is a vector of predictor specified at the area level, and  $u_i \sim N(0, \sigma_u^2)$  is an error term.

Since  $P_i$  is unobserved, the model is estimated by assuming that

$$\hat{P}_i^{dir} = P_i + \epsilon_i \quad (18)$$

where  $\hat{P}_i$  is the direct estimate of poverty for area  $i$ , and  $\epsilon_i \sim N(0, \sigma_{\epsilon,i}^2)$  represents sampling error in  $\hat{P}_i^{dir}$ . Combining equations 17 and 18 yields:

$$\hat{P}_i^{dir} = x_i' \beta + u_i + \epsilon_i, \quad (19)$$

Thus, the random effect in this model  $u_i$  has a constant variance while the variance of  $\epsilon_i$  is area-specific. Under this model, the best linear unbiased predictor (BLUP) is equal to:

$$\hat{P}_i^{blup} = x_{ij}' \beta + b_i \quad (20)$$

where  $b_i = \gamma_i (\hat{P}_i^{dir} - x_i' \beta)$ , and  $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{\epsilon,i}^2}$ . The assumption that  $\sigma_{\epsilon,i}^2 = \hat{\sigma}_{\epsilon,i}^2$  enables  $\beta$  and  $\sigma_u^2$  to be estimated using restricted information maximum likelihood (REML). We use the "ampl" method for estimation developed by [Li and Lahiri \(2010\)](#), which ensures strictly positive variance estimates, as implemented in the Stata *fayherriot* command ([Halbmeier et al. \(2019\)](#))

To obtain Empirical Best Predictors, we obtain smoothed direct estimates of  $\sigma_{\epsilon,i}^2$  from the survey and treat them as known. We first use the the Horvitz-Thompson direct variance estimator to obtain raw estimates of the variance of the area-level direct poverty estimates, denoted as  $\hat{\sigma}_{\hat{P},i}^2$  :

$$\hat{\sigma}_{\hat{P},i}^2 \approx \frac{1}{\left(\sum_{j=1}^{n_i} w_{i,j}\right)^2} \left[ \sum_{j=1}^{n_i} w_{i,j} (w_{i,j} - 1) I(y_{i,j} < Z) \right] \quad (21)$$

where  $I(y_{i,j} < Z)$  indicates if sample household  $j$  in area  $i$  has per capital income  $y_{i,j}$  below the poverty line  $Z$ ,  $n_i$  represents the number of sample households in area  $i$ , and  $w_{i,j}$  indicates the sample weight, equal to the inverse probability of selection, for household  $j$ .

We then obtained smoothed estimates of the variance by regressing the log of the raw variance estimates on the log number of sample households, following [You \(2022\)](#).

$$\ln(\hat{\sigma}_{\hat{P},i}^2) = \eta_0 + \ln(n_i) \eta_1 + \epsilon_i \quad (22)$$

where  $\epsilon_i \sim N(0, \psi)$ . We finally obtain the smoothed variance estimate  $\tilde{\sigma}_{\hat{P},i}^2$  as:

$$\tilde{\sigma}_{\hat{P},i}^2 = \exp\left(\hat{\eta}_0 + \ln(n_i) \hat{\eta}_1 + \frac{\hat{\psi}}{2}\right) \quad (23)$$

The Empirical Best Linear Unbiased Estimator (EBLUP) of  $\hat{P}_i$  under the Fay-Herriot

model can be expressed as:  $\hat{P}_i^{FH} = \hat{\gamma}_i \hat{P}_i^{dir} + (1 - \hat{\gamma}_i) x'_i \hat{\beta}$ , where  $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{\epsilon,i}^2}$ .  $\hat{P}_i^{FH}$  is therefore a weighted average of the direct estimates and the predictions, with the weights determined by the relative variance of the estimated sampling and model error. At one extreme, the Fay-Herriot estimates will equal the synthetic prediction estimate  $x'_i \hat{\beta}$  when there is zero shrinkage towards the direct estimate, which occurs when it is an out of sample area, the sample variance for that area ( $\hat{\sigma}_{\epsilon,i}^2$ ) approaches infinity or when the predictors  $x'_i$  perfectly predict poverty rates across areas ( $\hat{\sigma}_u^2 = 0$ ). At the other extreme, when using the raw variance estimates, the Fay-Herriot estimate would equal the direct estimate  $\hat{P}_i^{dir}$  in areas in where all sample households are either poor or non-poor and the estimated sample variance ( $\hat{\sigma}_{\epsilon,i}^2$ ) equals zero. In that case, there would be full shrinkage to the sample estimate. The variance smoothing procedure described above, however, avoids full shrinkage by ensuring that the smoothed variance is strictly positive, even in cases where all sample households in an area are poor or non-poor.

### A.3.2 Unit-context EBP model with area-level predictors

The final option we consider is a unit-context model with area-level predictors. This differs in only one way from the Empirical Best Predictor (EBP) model with household variables described in section A.1.1: The predictor variables now only consist only of area-level predictors. Therefore the model estimates:  $\ln y_{ij} = x'_{ij} \beta + u_i + \epsilon_{ij}$ , where all parameters are defined as in the household model described in section A.1.1, except that  $x'_{ij}$  now only contains municipal level variables and a constant. Compared with the Fay-Herriot model, this method is less flexible because it assumes that both the variance of both error terms  $u_i$  and  $\epsilon_{ij}$  is constant. In contrast, the Fay-herriot model allows the variance of  $\epsilon_{ij}$  to vary by municipality. However, because the unit-context model assumes an equal variance for all municipalities, it is estimated more precisely than in the Fay-Herriot case.<sup>24</sup> In addition, the model predicts household per capita income rather than municipal poverty rates, which contains more information and therefore may, all else equal, improve predictive performance. Therefore, the relative performance of the Fay-Herriot and unit-context model when only area-level predictors are available is an empirical question.

## B Annex B: Use of Survey Weights with EBP models

In general, small area estimation involves estimating a model using standard household survey data. These are often collected using a two-stage sample with PSUs selected with probability proportional to population size, known as PPS sampling. Because PSU population size is systematically correlated with household income and nearly every other outcome of interest, failing to properly adjust for weights will lead to what statisticians call informative sampling bias, and what economists refer to as endogenous sampling or sample selection bias. As a result, it is standard practice to adjust for sample weights when

---

<sup>24</sup>A recently developed approach allows the parameters in the EBP model to vary across areas (Lahiri and Salvati, 2023)

estimating descriptive statistics or model parameters (Solon et al. (2015)). The household sample weights are set equal to the product of the inverse probability of selection, which eliminates this source of bias.

Of the seven approaches to estimation outlined in the previous section, four are Empirical Best Predictor (EBP) models. Unlike most econometric models, properly adjusting for sample weights when estimating EBP models is complex, and there is not yet a consensus in the literature on the best method to use. This is an important issue because the treatment of weights in EBP models can have large impacts on the accuracy of small area estimates. Ideally, weighting for EBP models would utilize the first stage selection probabilities, or the probability that each sampled PSU was selected (Rabe-Hesketh and Skrondal, 2006). Unfortunately, however, this information is not typically included in household survey data files.

This section reviews four methods that incorporate sample weights when estimating EBP models and the first order selection probabilities are unknown. As noted above, the basic log-linear EBP model can be written as:  $\log y_{ij} = x_{ij}^T \beta + b_i + u_i + \epsilon_{ij}$  using the notation defined in section 2.

## B.1 Conditional weights

The conditional weighting scheme is described in detail in Bates et al. (2015). Essentially it assumes that  $(\log y_{ij}|b_i) \sim N(X_{ij}^T \beta + b_i, \sigma_\epsilon^2 W_{ij}^{-1})$ , where  $W_{ij}$  is an exogenous weight reflecting the inverse selection probability of household  $j$  into the sample.  $W_{ij}$  thus acts as a "heteroscedasticity weight", because it is modeled directly into the variance, and is incorporated into the estimation algorithm in order to weight observations according to their specified inverse variance (see Bates et al. (2015) for details).

The main shortcoming of the conditional weighting method is that, during the estimation process, the weights are conditioned on a particular value of the mean random effects  $b_i$ . This implies that weights are not fully taken into account when estimating  $b_i$  itself. In particular, this method does not fully incorporate weights when estimating the area-specific shrinkage factors  $\gamma_a$ , which partly determine  $b_i$ . The conditional weighting method uses the following formula to calculate  $\gamma_a$ :  $\hat{\gamma}_a = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \frac{\hat{\sigma}_\epsilon^2}{n_a}}$ , where  $n_a$  is the number of sample units in area  $i$ , which for a household-level model is the number of sample households in area  $i$ . This is the same formula for  $\hat{\gamma}_a$  as in Battese et al. (1988) and Molina and Rao (2010), which did not adjust for sample weights when estimating the model. However, as noted by You and Rao (2002), Van der Weide (2014) and Guadarrama et al. (2018), the presence of heterogeneous sample weights reduces the effective sample size in each area, which should be incorporated into the formula for  $\gamma_a$  as follows:  $\hat{\gamma}_a = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \delta_a \hat{\sigma}_\epsilon^2}$ , where  $\delta_a = \frac{\sum_{h=1}^{m_a} w_{ij}^2}{(\sum_{h=1}^{m_a} w_{ij})^2}$ , the inverse of the effective sample size, and  $w_{ij}$  represents the sample weight for household  $j$  in area  $i$ . When  $w_{ij} = w_a \forall h$ , the weights are identical for all units in the area, and  $\delta_a = \frac{1}{m_a}$  as in the conditional weighting method. Because the conditional weighting method does not account for  $\delta_a$  when estimating  $\gamma_a$ , it gives too much weight to the direct survey esti-

mates relative to the synthetic model predictions in the estimates, leading to sub-optimal estimates in the typical case where sampling weights are heterogeneous within an area.

## B.2 Partial adjustment weighting

You and Rao (2002) and Guadarrama et al. (2018) recognize this issue and proposed a different method for adjusting for sample weights in EBP estimation. We refer to this method as the "*partial adjustment method*" because it adjusts for weights when estimating  $\beta$  and  $b$  but not the variance components  $\sigma_v^2$  and  $\sigma_\epsilon^2$  (Huang and Hidirolou, 2003). The partial adjustment method consists of three steps:

1. Estimate an unweighted linear mixed model, which in matrix notation can be written as:

$$\log y = x\beta + zb + zu + \epsilon$$

where  $z$  is an  $n$  by  $m$  design matrix mapping observations to areas,  $n$  is the number of sample observations and  $m$  is the number of areas in the sample.

and obtain  $\hat{\beta}$ ,  $\hat{b}$ ,  $\hat{\sigma}_v^2$  and  $\hat{\sigma}_\epsilon^2$ .

2. Calculate  $\hat{\beta}^w$ , a version of  $\hat{\beta}$  that adjusts for sample weights, as follows:

$$\hat{\beta}^w = (x'W\tilde{x})^{-1} (x'W\tilde{y})$$

where  $\tilde{x} = x' - \hat{\gamma}z'\hat{b}$ ,  $\tilde{y} = y - \hat{\gamma}z'\hat{b}$ , and  $W$  is an  $n$  by  $n$  diagonal matrix of weights, with  $n$  representing the total number of units in the sample.  $\hat{\gamma}_a$  and  $\delta_a$  are defined as in the previous section, with  $\hat{\gamma}_a = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \delta_a \hat{\sigma}_\epsilon^2}$ , and  $\delta_a = \frac{\sum_{h=1}^{m_a} w_{ij}^2}{(\sum_{h=1}^{m_a} w_{ij})^2}$ . This estimates

$\hat{\beta}^w$  by "partialling out" the estimated mean random effects  $\hat{b}$  (which are estimated without weights) and applying the standard Generalized Least Squares (GLS) formula to obtain  $\hat{\beta}^w$ , an estimate of the coefficients that adjusts for weights.

3. Calculate  $\hat{b}^w$ , a version of the random effect means that adjust for weights, by calculating for each area  $i$ :  $\hat{b}_a^w = \hat{\gamma}_a \bar{e}_{wa}$  where  $\bar{e}_{wa} = \frac{\sum_{h=1}^{m_a} W_{ij} (y_{ij} - X_{ij} \hat{\beta}^w)}{\sum_{h=1}^{m_a} W_{ij}}$  is the weighted average of the level 0 residuals in area  $i$  (calculated using weighted coefficient estimates), and  $\hat{\gamma}_a$  is defined as above.

As noted above, this procedure adjusts for sampling weights when estimating the coefficients of the predictor variables  $\beta$  and the mean of the random effects  $b$ , but uses the estimated variance components  $\hat{\sigma}_v^2$  and  $\hat{\sigma}_\epsilon^2$  obtained from the unweighted mixed model estimated in step 1. This procedure was originally proposed by You and Rao (2002) for estimating mean outcomes, which in expectation are unaffected by biased estimates of variance components. Guadarrama et al. (2018) recognizes that the best predictors are given by the random effect means and the variance components replaced by their weighted versions.<sup>25</sup> However, the simulation exercises implemented in that article use the partial

---

<sup>25</sup>Guadarrama et al. (2018) p.8

adjustment method, which does not replace the estimated variance components with their weighted versions.<sup>26</sup> This, as shown below, leads to bias when using this weighting method that is particularly large when using unit-context models for poverty estimation in the Mexican context.

### B.3 Hybrid weighting

To address the shortcomings in both the conditional and partial adjustment methods, we propose an extension to the partial adjustment method, which we refer to as "hybrid weights". The extension replaces the estimates of the variance components with their weighted version. This adds a fourth and fifth step to the "partial adjustment weights" developed in [You and Rao \(2002\)](#) and [Guadarrama et al. \(2018\)](#), which utilize the conditional weighting method to estimate a weighted version of the variance components. Essentially, this entails estimating a new weighted mixed model that restricts the coefficients on the predictors to equal  $\hat{b}^w$ . The two additional steps are as follows:

4. Estimate the following weighted mixed model:

$$e_{ij}^w = \beta_0 + b'_i + \eta_a + v_{ij}, \text{ where } e_{ij}^w = \log Y_{ij} - X_{ij}\hat{\beta}^w.$$

This equation is similar to what would be obtained by subtracting  $X_{ij}\beta$  from both sides of the original model:  $\log Y_{ij} = X_{ij}\beta + v_i + \epsilon_{ij}$ . This effectively imposes the restriction in the original model that  $\beta = \hat{\beta}^w$ , except for the intercept which is allowed to vary.

5. Adjust the estimated variance components from step 4 to account for the fact that they are obtained from a regression on residuals, using the following degrees of freedom correction:

$$\hat{\sigma}_v^{2w} = \hat{\sigma}_\eta^2 * \frac{m-1}{m-p}$$

where  $m$  is the number of target areas contained in the sample and  $p$  is the number of predictor variables in the original model including the intercept (the number of columns in  $x$ ).  $m$  is used in this adjustment because  $\hat{\sigma}_\eta^2$  is identified using the variation across  $m$  sample areas. Similarly, we use

$$\hat{\sigma}_\epsilon^{2w} = \hat{\sigma}_v^2 * \frac{n-1}{n-k},$$

where  $n$  is the number of units contained in the sample, to generate an estimate of the variance of the unit-level error term that accounts for weights. These degrees of freedom adjustments are necessary because  $\hat{\sigma}_\eta^2$  and  $\hat{\sigma}_v^2$  are estimated using a model with one predictor ( $\beta_0$ ) while  $\hat{\sigma}_v^{2w}$  and  $\hat{\sigma}_\epsilon^{2w}$  are parameters from models specified with  $k$  predictors, namely:

$y_{ij} = X_{ij}\hat{\beta}^w + \hat{b}_i^w + v_i^w + \epsilon_{ij}^w$ . Therefore, the degrees of freedom adjustments convert  $\hat{\sigma}_\eta^2$  and  $\hat{\sigma}_v^2$  into accurate estimates of  $\hat{\sigma}_v^{2w}$  and  $\hat{\sigma}_\epsilon^{2w}$ , respectively.

---

<sup>26</sup>[Corral et al. \(2021\)](#) and [Corral et al. \(2022\)](#) use a different method for model estimation adjusting for weights based on [Huang and Hidiroglou \(2003\)](#), as described in [Van der Weide \(2014\)](#). We do not consider this method in this version but will include it in a revised version of this paper.

The estimates of  $\hat{\sigma}_\eta^2$  and  $\hat{\sigma}_v^2$  obtained from the mixed effects model in step four incorporate sample weights using conditional weighting (Pinheiro et al. (2015), Bates et al. (2015)). We therefore consider this method to be a "hybrid" approach, because it extends the partial adjustment weighting method using the conditional weighting method. This procedure yields the same estimates of  $\beta$  and  $b$  that are obtained using the partial adjustment method, and improves on the partial adjustment method by accounting for weights when estimating the variance components  $\sigma_v^2$  and  $\sigma_\varepsilon^2$ . This reduces bias when sample weights are "informative" or correlated with the outcome. The reduction can be significant and the variance of the area effect,  $\sigma_v^2$ , is significant and when the effective sample size is small. Effective sample sizes tend to be small, in turn, either when the actual sample size  $n$  is small, or when estimating unit-context models in which predictors only vary across PSUs and/or target areas.

## B.4 Weighted Generalized Least Squares (GLS)

This method was originally derived by Huang and Hidioglou (2003), based on You and Rao (2002), and then adopted by Van der Weide (2014) and Nguyen et al. (2018) for the original Povmap and Stata SAE packages developed by the World Bank. The estimation process consists of four main steps, described using the notation in Van der Weide (2014):

1. Estimate a weighted OLS regression to obtain estimates of  $\sigma_{\eta,ols}^2$ ,  $\sigma_{\varepsilon,ols}^2$  using "Henderson's method 3" decomposition (Henderson, 1953), where

$X$  is the  $n$  by  $K$  matrix of predictors,  $n$  is the number of sample observations, and  $y$  is an  $n$  by 1 vector containing the dependent variable.

2. Estimate  $\hat{\beta}_{gls} = (X^T V_w^{-1} X)^{-1} (X^T V_w^{-1} y)$  where

$V_w$  is an  $n$  by  $n$  matrix equal to  $I_N \hat{\sigma}_{\varepsilon,ols}^2 W^{-1} + \Omega \hat{Q}$

$I_n$  is the  $n$  by  $n$  identity matrix

$W$  is the  $n$  by  $n$  diagonal matrix where each diagonal element contains the value of the sample weight  $w_{ij}$  for the corresponding household.

$\Omega$  is an  $n_i$  by  $n_i$  diagonal matrix where each element contains  $\frac{\sum_1^{n_i} w_{ij}}{\sum_1^{n_i} w_{ij}^2}$

$\hat{Q}$  is a block-diagonal  $n$  by  $n$  matrix in which each block representing area  $i$  is set equal to  $\hat{\sigma}_{u,ols}^2 1_{n_i} 1_{n_i}^T$

$1_{n_i}$  is a column vector of ones of dimension  $n_i$ , the number of sample households in area  $i$

3. Estimate GLS variance components  $\hat{\sigma}_{u,gls}^2$  and  $\hat{\sigma}_{\varepsilon,gls}^2$  using the formula provided in the annex of Van der Weide (2014) and Huang and Hidioglou (2003).

4. Estimate the mean random effects for each area  $i$  as  $b_i = \left( \frac{\sum_1^{n_i} w_{ij}}{\sum_1^{n_i} w_{ij}^2} \right) \hat{\sigma}_{u,gls}^2 1_{n_i}^T V_{i,w}^{-1} e_i$

where  $V_{i,w}$  is the  $n_i$  by  $n_i$  block diagonal of the  $V_w$  matrix corresponding to area  $i$  and  $e_i$  is the  $n_i$  by 1 vector of GLS residuals, where each element  $e_{i,j} = y_{ij} - X^T \hat{\beta}_{gl_s}, \forall j$

Like the hybrid and conditional weighting methods, the GLS method adjusts for weights when estimating all model parameters. Unlike conditional weighting, it appropriately accounts for heterogeneity in the weights when estimating variance components.