A map of the poor or a poor map?

Paul Corral,[*] Kristen Himelein,[†] Kevin McGee[‡] and Isabel Molina[§][¶]

March 14, 2021

Abstract

This paper evaluates the performance of different small area estimation methods using model and design-based simulation experiments. Design-based simulation experiments are carried out using the Mexican Intra Censal survey as a census of roughly 3.9 million households from which 500 samples are drawn using a two-stage selection procedure similar to that of Living Standards Measurement Study (LSMS) surveys. The estimation methods considered are that of Elbers, Lanjouw and Lanjouw (2003), the empirical best predictor of Molina and Rao (2010), the twofold nested error extension presented by Marhuenda et al. (2017), and finally an adaptation, presented by Nguyen (2012), that combines unit and area level information, and which has been proposed as an alternative when the available census data is outdated. The findings show the importance of selecting a proper model and data transformation so that model assumptions hold. A proper data transformation can lead to a considerable improvement in MSE. Results from design-based validation show all small area estimation methods represent an improvement, in terms of MSE, over direct estimates. However, methods that model unit level welfare using only area level information suffer from considerable bias. Because the magnitude and direction of the bias is unknown ex ante, methods relying only on aggregated covariates should be used with caution but may be an alternative to traditional area level models when these are not applicable.

Key words: Small area estimation, ELL, Poverty mapping, Poverty map, Empirical best, Parametric bootstrap, nested error model, twofold nested error model

JEL classification: C55, C87, C15

[*] The World Bank Group - Human Development Group Chief Economist's Office (pcorralrodas@worldbank.org)

[†] The World Bank Group - Poverty and Equity Global Practice

[‡] The World Bank Group - Development Economics Data Group

[§] Universidad Carlos III de Madrid - Department of Statistics (imolina@est-econ.uc3m.es)

[¶] The authors acknowledge financial support from the World Bank. We thank Sandra Carolina Segovia for research assistance. We thank J.N.K. Rao for comments and suggestions. We also thank Roy van der Weide, Alexandru Cojocaru, David Newhouse, and Samuel Freije-Rodriguez for comments. Additionally, we thank Carolina Sanchez and Roberta Gatti for providing support and space to pursue this work. Finally, we thank the Global Solutions Group on Data for Policy, as well as all attendants of the Summer University courses on Small Area Estimation. Any error or omission is the authors' responsibility alone.

1 Introduction

The eradication of poverty was the first Millennium Development Goals (MDGs) established by the United Nations in 2000 and continues as Sustainable Development Goals (SDG) 1.1.1, but governments can only properly target poverty if they know where it is. Traditionally, for a given country, the best source for information on the living standards of its population are household surveys. These surveys are a powerful tool towards defining and addressing the needs of people. These surveys, however, usually offer reliable information only at highly aggregated levels of the population. In other words, *direct* survey estimates tend to be adequate for very large populations but inadequate for smaller populations.

Small area estimation (SAE) is a branch of statistics focused on improving the reliability of estimates and the associated measures of uncertainty for populations where samples cannot produce sufficiently reliable estimates (Rao and Molina, 2015). Small areas can be any population subgroup and are not necessarily tied to geographical areas. According to Ghosh and Rao (1994), the use of small area statistics increased towards the end of the 20th century due to improved computing power and the advent of theoretically sound statistical methods. The main principle behind small area statistical methods is to use modeling to "borrow strength" from auxiliary data sources (e.g. census or administrative data) to produce more efficient estimators than direct survey data alone.

Model-based techniques for small area estimation commonly fall within two groups: i) area based models; for examples see Fay and Herriot (1979) and Torabi and Rao (2014); ii) unit-level models; for examples see Molina and Rao (2010) and Elbers, Lanjouw and Lanjouw (2003). The focus of this paper is to conduct model and design-based evaluation of different unit-level SAE methods. Model-based methods assume a data generating process for the population that follows the model's assumptions, and the target parameters are assumed to be random (Tzavidis et al., 2018). In this paper, model-based evaluation is done considering two-stage sampling strategies and thus assumes a model that includes variation at the domain level as well as at the cluster level, where clusters are the primary sampling units, nested within domains (see Marhuenda et al., 2017). Design-based simulation entails applying the chosen methods under realistic conditions (Tzavidis et al., 2018). Here, a design-based evaluation of several estimators is performed using the 2015 Mexican Intra Censal Survey, which is used here as a census, and from which we draw repeated samples. The estimation methods considered are that of Elbers, Lanjouw and Lanjouw (2003 - ELL henceforth), the empirical best predictor of Molina and Rao (2010 - MR henceforth),[1] the twofold nested error extension of MR (2010) presented by Marhuenda et al. (2017), and finally an adaptation, presented by Nguyen (2012), that combines unit and area level information, and which has been proposed as an alternative when the available census data is too outdated to be considered for use.

The paper's contribution is twofold: i) it presents a thorough evaluation of the considered methods as well as their pros and cons, and ii) it employs the sampling strategies most often encountered

[1] In reality, we apply the Census EB variant introduced by Guadarrama, Molina and Rao (2016)

in real world applications to conduct design-based validations. The results are expected to provide a guideline to others undertaking small area estimation. The results from design-based validation show that all small area estimation methods represent an improvement, in terms of MSE, over direct estimates. However, given the inherent high level of bias of methods relying solely on aggregate information (see Nguyen, 2012; Lange et al., 2018; Masaki et al., 2020) these should be used with caution and only under certain circumstances.

The paper first presents an overview of the small area estimation methods considered, directing readers who are interested in a more in-depth description to the appropriate sources. Then the model-based simulation design is detailed and results of the simulation experiments are presented. This section is followed by a description of the 2015 Mexican Intra Censal Survey and how it is adapted to represent a census and how samples are drawn from the created census. Results from the design-based simulation experiments are then presented. Finally, conclusions and lessons learned are provided.

2 Unit level models

The nested error model used for small area estimation by ELL (2003) and Molina and Rao (2010) was originally proposed by Battese, Harter and Fuller (1988) to produce county-level corn and soybean crop area estimates for the American state of Iowa. For the estimation of poverty and welfare, the ELL and MR methods assume the transformed welfare y_{ch} for each household h within each location c in the population is linearly related to a $1 \times K$ vector of characteristics (or correlates) x_{ch} for that household, according to the nested error model:

$$y_{ch} = x_{ch}\beta + \eta_c + e_{ch},\ h = 1, \ldots, N_c,\ c = 1, \ldots, C, \tag{1}$$

where η_c and e_{ch} are respectively location and household-specific idiosyncratic errors, assumed to be independent from each other, following:

$$\eta_c \overset{iid}{\sim} N\left(0, \sigma_\eta^2\right),\ e_{ch} \overset{iid}{\sim} N\left(0, \sigma_e^2\right)$$

where the variances σ_η^2 and σ_e^2 are unknown. Here, C is the number of locations in which the population is divided and N_c is the number of households in location c, for $c = 1, \ldots, C$. Finally, β is the $K \times 1$ vector of coefficients. Under the original ELL methodology, the locations indexed with c are supposed to be the clusters, or primary sampling units (PSUs) of the sampling design and do not necessarily correspond to the level at which the estimates will be ultimately produced. In fact, clusters are typically nested within the areas of interest (e.g. census enumeration areas within large administrative areas). Presenting estimates at a higher aggregation level than the clusters (for which random effects are included in the model) may not be appropriate in cases of considerable between-area variability, and may underestimate the estimator's standard errors (Das and Chambers, 2017). The recommended approach to mitigate this issue is to include covariates that sufficiently explain the between-area heterogeneity in the model (*ibid*). In this regard, ELL (2002) suggests the inclusion of cluster-level covariates as a way to explain location effects. Nevertheless, this approach is context specific and may not always suffice to ameliorate the issues with between-area heterogeneity. In this regard, Marhuenda et al. (2017) recommend and show that location effects should be at the same aggregation level at which estimation is desired.

If the location effect is specified at the same level where estimation is desired, then the difference between ELL (2003) and MR (2010) reduces to differences in how estimates are obtained and the addition of Empirical Best (EB) prediction by MR (2010). The EB method from MR (2010) conditions on the survey sample data and thus makes more efficient use of the information at hand, whereas ELL does not include this component. In essence, under ELL, for any given area present in the sample the ELL estimator of the census area mean $\bar{y}_c$ is obtained by averaging across M simulated censuses and is given by $\bar{y}_c^{*(m)} \approx \bar{X}_c'\beta + \eta_c^{*(m)} + \bar{e}_c^{*(m)}$, $m = 1, \ldots, M$, where $E[\eta_c^*] = 0$ and $E[e_{ch}^*] = 0$. Thus, the ELL estimator $\frac{1}{M}\sum_{m=1}^{M} \bar{y}_c^{*(m)}$, which approximates $E\left(\bar{y}_c\right)$, reduces to the regression synthetic estimator, $\bar{X}_c'\beta$ (MR, 2010). On the other hand, under MR (2010), conditioning on the survey sample ensures the estimator includes the random location effect, since $E[\bar{y}_c|\eta_c] \approx \bar{X}_c'\beta + \eta_c$. Mechanically, however, conditioning on survey data requires the linking of areas across surveys and census, something that is not always straightforward.[2] Other differences between ELL (2003) and MR (2010) are the computational algorithms used to obtain point and noise estimates; see Corral, Molina and Nguyen (2020 - CMN henceforth) for further discussion.

ELL's (2003) approach to obtain estimates builds upon the multiple imputation (MI) literature in that it uses a single algorithm that produces point and noise estimates by varying model parameters across simulations (see Tarozzi and Deaton, 2009 as well as CMN, 2020). The use of MI methods for obtaining point and noise estimates has shortcomings, however. Under multiple imputation, the method that yields the lowest MSE does not necessarily yield valid statistical inference (Van Buuren, 2018), which is contradictory to the goal of small area estimation in improving precision. Accordingly, CMN (2020) show using simulated populations that the noise estimate (referred to as variance by ELL) is not an appropriate estimate of the MSE. In contrast, MR's (2010) point estimates are obtained through two separate procedures. Point estimates are obtained by a Monte Carlo simulation and noise estimates are obtained by a parametric bootstrap procedure originally proposed by González-Manteiga et al. (2008).[3]

In light of the emerging literature, including MR (2010), the World Bank expanded the original ELL design by adding EB predictors and revising its fitting methodology to incorporate heteroskedasticity and survey weights to account for complex sampling designs (Van der Weide, 2014). The survey weights are incorporated in the estimates of the regression coefficients and in the variance of the components following Huang and Hidiroglou (2003) based on Henderson method III (Henderson,

[2] Under ELL, when including area level covariates, linking the survey and the census areas is also required. Note that it is not necessarily the case that enumeration areas for a census and survey will match.

[3] For more details, see Corral, Molina and Nguyen (2020).

1953) as well as in the predicted area effects as in the pseudo empirical best linear unbiased predictor (EBLUP) proposed by You and Rao (2002). In the absence of survey weights and heteroskedasticity, the fitting method yields parameter estimates that are quite similar to the restricted maximum likelihood (REML) fitting method used by MR (2010). The difference remains, however, in how point and noise estimates of the target indicators are obtained.

As further detailed in CMN (2020), however, even following these revisions, the bootstrap procedure, as implemented by `PovMap` software (Zhao, 2006) for EB and later detailed in Nguyen et al. (2018) differed from the original EB procedure from MR (2010). CMN (2020) note the remaining issues rested in the continued reliance on MI methods as the basis of the methodology. Because under the updated fitting methodology, Van der Weide (2014) does not offer an estimate of var $\left(\sigma_{\eta}^{2}\right)$ bootstrap samples of the data must be taken for each simulation to obtain the fitting parameters $\left(\hat{\beta}^{*}, \hat{\sigma}_{\eta}^{2*}, \hat{\sigma}_{e}^{2*}\right)$.[4] This approach was taken to allow for an algorithm similar to the procedure used in the original implementation of ELL. However, if clusters were equal to areas, then it is unlikely that all areas are included in the sample and thus an area could benefit from EB only in a subset of the simulations, introducing bias into the resulting point and noise estimates. CMN (2020) also present evidence on the fact that, even if the location effect was modeled at the domain level and bootstrap samples of clusters are taken, some bias would still likely remain in the resulting estimates of the original implementation of EB in `PovMap` and the `sae` Stata package.

The Stata package for small area estimation developed by Nguyen et al. (2018) has been updated to integrate the fitting methods from Van der Weide (2014) with the prediction and bootstrap methods from MR (2010). The new method is referred to as the H3-CensusEB (Molina, 2019), and the previous method (i.e. original bootstrap using EB from Van der Weide (2014)) is called H3-CBEB.[5] CMN (2020) performs a model-based validation of the different methods: i) CensusEB, ii) EB, iii) H3-CBEB and iv) ELL. CMN (2020) extend the simulations done by MR (2010) by i) including the area means of the covariates as additional variables in the model; ii) considering a model that has considerably higher explanatory power by adding more covariates; iii) considering larger population sizes and smaller sampling fractions; and iv) generating errors from a Student's t_5 instead of a normal distribution. However, in all these simulations, population data are generated under model (1).

Currently, the software available for small area estimation of non-linear parameters, such as the `sae` Stata package by Nguyen et al. (2018) as well as the R package `sae` by Molina and Marhuenda (2015), only allow for estimation under the nested error model specified in (1). However, since household surveys often use two stage sampling, it seems appropriate to consider a twofold nested error model. Marhuenda et al. (2017) extend the EB method from MR (2010) to a twofold nested error model, given by:[6]

[4] ELL (2002) provides the derivation of the estimate for var $\left(\sigma_{\eta}^{2}\right)$ in their appendix.

[5] The method is called CensusEB because it does not link survey and census households. The old method is called the clustered bootstrap EB (CBEB). For the updated Stata package, see https://github.com/pcorralrodas/SAE-Stata-Package

[6] For simplicity, we omit the heteroskedasticity weights considered by Marhuenda et al. (2017).

$$y_{ach} = x_{ach}\beta + \eta_a + \eta_{ac} + e_{ach},\ h = 1, \dots, N_{ac},\ c = 1, \dots, C_a,\ a = 1, ..., A, \tag{2}$$

where η_a is the random effect for area a and η_{ac} is the random effect of cluster c within area a. These effects along with the individual model errors, e_{ach}, represent the unexplained variation of the transformed welfare, y_{ach}, across areas, clusters, and households (Marhuenda et al., 2017).[7] All three components are assumed to be mutually independent, following:[8]

$$\eta_a \overset{iid}{\sim} N\left(0, \sigma_a^2\right),\ \eta_{ac} \overset{iid}{\sim} N\left(0, \sigma_{ac}^2\right),\ e_{ach} \overset{iid}{\sim} N\left(0, \sigma_e^2\right).$$

Using the assumed twofold nested error model, Marhuenda et al. (2017) study the effect of a misspecified model (i.e. when assuming a cluster effect and presenting results at the area level) on the MSE estimator. The argument posited by Das and Chambers (2017) is that auxiliary variables should explain the between-area variation of the response variable. If this fails, there may be model misspecification, which can lead to an underestimation of the true MSE of the ELL estimator (Marhuenda et al., 2017). Through model-based simulation experiments under the assumed model (2), Marhuenda et al. (2017) reach three important conclusions under the model-based setup:

1. The relative values of σ_{ac}^2 and σ_a^2 are of key importance; the larger the value of σ_a^2 relative to σ_{ac}^2, the more problematic it is to apply models where effects are specified at the cluster level, including the original ELL and EB with locations specified at the cluster level. Additionally, EB with location effects specified at the cluster level, while performing better than ELL,[9] will also perform worse the larger the value of σ_a^2 is relative to σ_{ac}^2.

2. Even if the true model contains random effects only at a single level, the assumption of a twofold model practically does not entail loss in efficiency.

3. EB estimates under model (1) with random effects specified at the area level will have similar performance to the twofold EB estimates. The recommendation is that if the estimation is done using model (1) because of software availability or simplicity, then the model's random effects should be specified at the level at which results are desired.

As noted by Marhuenda et al. (2017), small area estimators based on unit level models often achieve very large reductions in MSE compared to direct estimators. EB estimators based on unit level models are also likely to achieve considerable gains in terms of MSE over Fay-Herriot (FH) area level models.[10] For example, Molina and Morales (2009) obtained mild gains over direct poverty

[7] Marhuenda et al. (2017) presents these effects as domains and sub-domains.

[8] For full derivation of EB predictors under the twofold model, see Marhuenda et al. (2017).

[9] The performance is better than ELL because it includes the average of the cluster effects as opposed to ELL which only includes the linear fit.

[10] Fay-Herriot models were introduced by Fay and Herriot (1979) and are a popular area-level small area estimation model.

and poverty gap estimates when using a FH model. Molina and Rao (2010) though, using the same data sources as Molina and Morales (2009) but applying unit level models and EB prediction, obtain considerably larger gains. However, when the available census and survey are not from the same year, small area estimates based on unit level models may result in biased estimates. In such scenarios, FH models offer an alternative because these do not require census microdata. Alternatively, twofold models such as that of Torabi and Rao (2014) could also be considered to achieve larger gains as noted by Molina (2019).

Another potential solution is to use only aggregated covariates in the model for the household level welfare. This alternative is presented by Nguyen (2012) in an application for Vietnam. The author proposes a model where the dependent variable is household level logarithm of per capita expenditure from a recent survey, in this case the Vietnam Household Living Standard Survey from 2006, whereas all covariates are commune level means. These means are obtained from a dated (1999) census, although the author notes geographic information system data (GIS) could also be included into the set of covariates. Nguyen (2012) obtains ELL estimates for small areas under that model and compares the performance with that of typical ELL estimates obtained using unit level covariates from the Vietnam Household Living Standard Survey from 2006 and the 2006 Rural Agriculture and Fishery Census. The author finds provinces and districts hovering around the middle of the distribution suffer from considerable re-rankings across methods, but those at the top and the bottom are relatively stable.

Lange et al. (2018) present an approach similar to Nguyen's (2012) which the authors suggest as an alternative in cases when census and survey data are not from similar periods, though the same issues noted above for the ELL method would likely persist in a model using only area-level covariates. Masaki et al. (2020) use a similar modeling approach to Nguyen's (2012), but take measures to address some of the shortcomings of a standard ELL approach and obtain EB estimators of Molina and Rao (2010) which appear to be more precise. The authors conduct a design-based validation study using a wealth index constructed with principal component analysis using census data for Sri Lanka and Tanzania. Their results show the approach may hold promise.

In the sections that follow, the different procedures discussed here are tested under model-based and design-based simulation experiments.

3 Model-based simulation experiments

The simulation experiment described here is based on those conducted by Marhuenda et al. (2017) in which the true data generating process is a twofold nested error model. Such a model will better accommodate the usual applications of poverty mapping, where household surveys use two-stage sampling. In fact, the traditional ELL method includes random effects only at the cluster level but estimates are given for a higher level.

In this simulation experiment a census data set of $N = 20,000$ observations is created, where observations are allocated among 40 areas $(a = 1, \ldots, A)$. Within each area, observations are uniformly spread over 10 clusters $(c = 1, \ldots, C_a)$. Each cluster, c, consists of $N_{ac} = 50$ observations, and each cluster is labeled from 1 to 10. The assumed model contains both cluster and area effects. Cluster effects are simulated as $\eta_{ac} \overset{iid}{\sim} N\left(0, 0.1^2\right)$, area effects as $\eta_a \overset{iid}{\sim} N\left(0, 0.05^2\right)$ and household specific residuals as $e_{ach} \overset{iid}{\sim} N\left(0, 0.5^2\right)$, where $h = 1, \ldots, N_{ac}$; $c = 1, \ldots, C_a$; $a = 1, \ldots, A$. Covariates are simulated as follows:[11]

1. x_1 is a binary variable, taking value 1 when a random uniform number between 0 and 1, at the household level, is less than or equal to $0.3 + 0.5\frac{a}{40} + 0.2\frac{c}{10}$.
2. x_2 is a binary variable, taking value 1 when a random uniform number between 0 and 1, at the household level, is less than or equal to 0.2.
3. x_3 is a binary variable, taking value 1 when a random uniform number between 0 and 1, at the household level, is less than or equal to $0.1 + 0.2\frac{a}{40}$.
4. x_4 is a binary variable, taking value 1 when a random uniform number between 0 and 1, at the household level, is less than or equal to $0.5 + 0.3\frac{a}{40} + 0.1\frac{c}{10}$
5. x_5 is a discrete variable, simulated as the rounded integer value of the maximum between 1 and a random Poisson variable with mean $\lambda = 3\left(1 - 0.1\frac{a}{80}\right)$.
6. x_6 is a binary variable, taking value 1 when a random uniform value between 0 and 1 is less than or equal to 0.4. Note that the values of x_6 are not related to the area's label.
7. x_7 is a binary variable, taking value 1 when a random uniform number between 0 and 1 is greater than or equal to $0.2 + 0.4\frac{a}{40} + 0.1\frac{c}{10}$

The welfare vector for each cluster within an area is created from the model with these covariates, as follows:

$$y_{ach} = 3 + 0.09x_{1ach} - 0.04x_{2ach} - 0.09x_{3ach} + 0.4x_{4ach} - 0.25x_{5ach} + 0.1x_{6ach} + 0.33x_{7ach} + \eta_a + \eta_{ac} + e_{ach}, \tag{3}$$

The dependent variable, y_{ach}, is in the log scale. The poverty line in this scenario is fixed at $z = 12$.

From the created "census," 20% of the observations are sampled in each of the clusters using simple random sampling without replacement;[12] this yields our "survey" data. The generation and sampling processes are repeated $L = 10,000$ times. In each simulation replicate, the following quantities are computed for the poverty rates and gaps in each area:

[11] Covariates are simulated following the approach from Molina and Rao (2010) and Marhuenda et al. (2017), with slight modifications.

[12] The same units are sampled in every simulation and the values of x_1 to x_7 for all the census units are also kept fixed; this implies that the values of these covariates for the sample units are always the same across simulations.

1. True poverty indicators τ_a, using the "census".

2. Direct estimators $\hat{\tau}_a^{DIR}$ using the "survey", defined as the sample versions of τ_c.

3. Census EB estimators $\hat{\tau}_a^{CEB_{ac}}$ presented in Marhuenda et al. (2017) based on a twofold nested error regression model, and obtained using a Monte Carlo approximation with $M = 50$ replicates. Note that, for this estimator, the fitted model agrees with the true data generating process (3).

4. Census EB estimators $\hat{\tau}_a^{CEB_a}$ presented in CMN (2020) based on a nested error model with only **area** random effects obtained using a Monte Carlo approximation with $M = 50$ replicates.

5. Census EB estimators $\hat{\tau}_a^{CEB_c}$ presented in CMN (2020) based on a nested error model with only **cluster** random effects and including the aggregate cluster and area means of the considered auxiliary variables, where $M = 50$.

6. Traditional ELL estimators $\hat{\tau}_a^{ELL_c}$, based on a nested error model with only **cluster** random effects and including the aggregate cluster and area means of the considered auxiliary variables, where $M = 50$.

7. Unit-context Census EB estimators $\hat{\tau}_a^{UC-CEB_a}$ based on a nested error model with random effects at the **area level**. This estimator follows the approach from Masaaki et al. (2020) that is a modified version of that of Nguyen (2012), which uses only area means for some of the right hand side variables.[13] Nguyen (2012) proposes this solution for the case when only a dated census and a recent survey are available.

8. Unit-context **two-fold nested error** Census EB estimators $\hat{\tau}_a^{UC-CEB_{ac}}$ based on a two-fold nested error model with random location effects at the **area and cluster level**. This estimator follows the approach from Masaaki et al. (2020) and Nguyen (2012), where only area means for some of the right hand side variables are used.[14]

Model bias and MSE are approximated empirically as in MR (2010), as the averages across the $L = 10,000$ simulations of the prediction errors in each simulation (l), $\hat{\tau}_a^{j(l)} - \tau_a^{(l)}$ and of the squared prediction errors respectively, where j stands for one of the methods: $DIR, CEB_{ac}, CEB_a, CEB_c, ELL_c, UC-CEB_a$, $UC - CEB_{ac}$.[15] Model bias and root MSE for a given area's estimate are computed at the area level as follows:

$$Bias\left(\hat{\tau}_a^j\right) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\tau}_a^{j(l)} - \tau_a^{(l)})$$

[13] The covariates used in this model are: $\bar{x}_{1ac}$, $\bar{x}_{3a}$, $\bar{x}_{4ac}$, $\bar{x}_{5a}$, and $\bar{x}_{7ac}$.

[14] The covariates used in this model are: $\bar{x}_{1ac}$, $\bar{x}_{3a}$, $\bar{x}_{4ac}$, $\bar{x}_{5a}$, and $\bar{x}_{7ac}$.

[15] $E\left(\hat{\tau}_c^j - \tau_c\right)$ for the bias and $E\left(\hat{\tau}_c^j - \tau_c\right)^2$ for the MSE, where $E\left(.\right)$denotes expectation under model 2.

$$RMSE\left(\hat{\tau}_a^j\right) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(\hat{\tau}_a^{j(l)} - \tau_a^{(l)})^2}$$

3.1 Results

The section presents the results from the model-based simulation experiments where the goal is to compare the performance of the different methods. Marhuenda et al. (2017) consider multiple scenarios where they simulate different values for σ_a^2 and σ_{ac}^2. The authors note what matters are the relative values. In this instance the interest is to assess how results differ when the random cluster effect is considerably smaller than the random area effect and when the random cluster effect is considerably larger than the random area effect. ELL would commonly specify its random location effect at the cluster level and then aggregate results to the area level. Consequently, we expect ELL to perform better when the random cluster effect is larger than the area random effect.

We consider two scenarios:

1. $\eta_{ac} \overset{iid}{\sim} N\left(0, 0.1^2\right)$ and $\eta_a \overset{iid}{\sim} N\left(0, 0.05^2\right)$
2. $\eta_{ac} \overset{iid}{\sim} N\left(0, 0.05^2\right)$ and $\eta_a \overset{iid}{\sim} N\left(0, 0.1^2\right)$

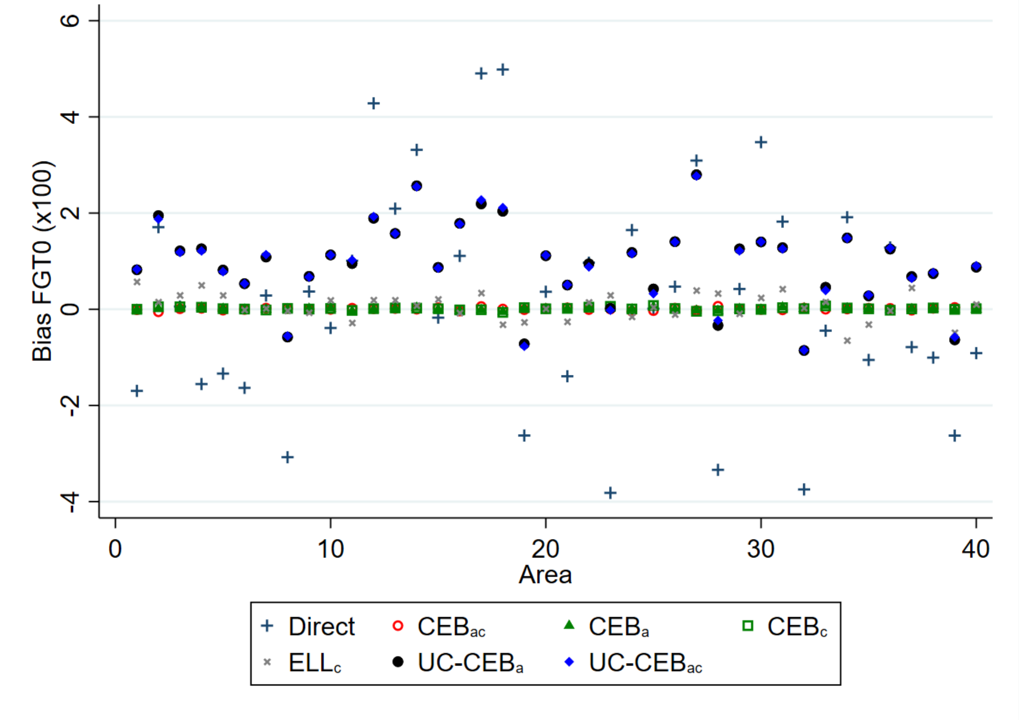
Simulation results under the two considered scenarios are presented respectively in Figures 1 and 3 for bias, and Figures 2 and 4 for MSE.

The results for bias and MSE are presented at the area level. With the exception of a few areas in which the ELL and unit-context methods demonstrate a slightly higher bias, all examined methods perform better than the direct estimators, in most cases by a substantial margin. For MSE, the result varies between the two simulation scenarios. Under scenario 1 with results shown in Figure 2, where the variance of cluster effects is double that of the area effects, the methods considered, including ELL and the unit-context methods, perform relatively well and in nearly all cases now do as well as or better than direct estimates, though again ELL and the unit-context methods perform worse than the other options.[16] However, in Figure 4, where the variance of the area effects is now much larger, ELL in particular performs poorly in terms of MSE likely due to the error misspecification and the contextual variables not explaining sufficiently the variability of the area effects. To a lesser extent a similar effect can be observed for the CensusEB estimator, based on a model with only cluster effects and contextual variables (CEB_c with context), which performs well in terms of MSE under scenario 1 but under-performs in scenario 2.

The twofold model results are aligned to the results presented by Marhuenda et al. (2017); the bias and MSE of estimates obtained under twofold fitting and onefold CensusEB fitting at the area level

[16] Despite the result, other issues from the implementation of ELL (2003) noted by CMN (2020) like the underestimation of the MSE still remain, unless the method to estimate MSE is adjusted.

Figure 1: Empirical model bias of FGT0 estimators for $\sigma_{ac} = 0.1$ and $\sigma_{a} = 0.05$



are largely indistinguishable. This result is interesting in that it resonates with the findings from Marhuenda et al. (2017); in the absence of a software solution for a twofold nested error regression, it is preferable to specify the random effects at the level at which results are desired. This will ensure that MSEs are minimized despite mistaken model assumptions.

Surprisingly, the two unit-context models used to obtain CensusEB estimators, one with random effects only at the area level and another with random effects at the cluster and area levels, show more bias than ELL within any given area.[17] The results shown here are not evident under the model based simulation conducted in Masaki et al. (2020)[18] because under the simulation presented here, true welfare is generated from household level covarites as is likely the case in real world scenarios. In Masaki et al. (2020) the authors chose to model the dependent variable using only 2 subdomain level covariates which are constant for all households in the subdomain.

The bias observed in the simulations conducted here for unit-context models is in part due to omitted variable bias.[19] If the true model includes x_{5ach} then when we only consider $\bar{x}_{5ac}$ we are omitting $z_{5ach} = x_{5ach} - \bar{x}_{5ac}$ from the model. Note that $\bar{x}_{5ac}$ is obtained in the "census" and thus $\bar{x}_{5ac}$ will be correlated to z_{5ach} in the survey due to it being a sample and not the entire population, we also know that z_{ach} is correlated to the dependent variable thus leading to the omitted variable

[17] The covariates used in this model are: $\bar{x}_{1ac}$, $\bar{x}_{3a}$, $\bar{x}_{4ac}$, $\bar{x}_{5a}$, and $\bar{x}_{7ac}$. In other simulations run, not shown, all the covariates' aggregates at the cluster level are used and similar results are obtained.

[18] See page 36 of Masaki et al. (2020)

[19] The unit-context models also display an upward bias in simulations where the population (20,000) is used to fit the model and to obtain the FGT0 predictors.

Figure 2: Empirical model MSE of FGT0 estimators; $\sigma_{ac} = 0.1$ and $\sigma_a = 0.05$

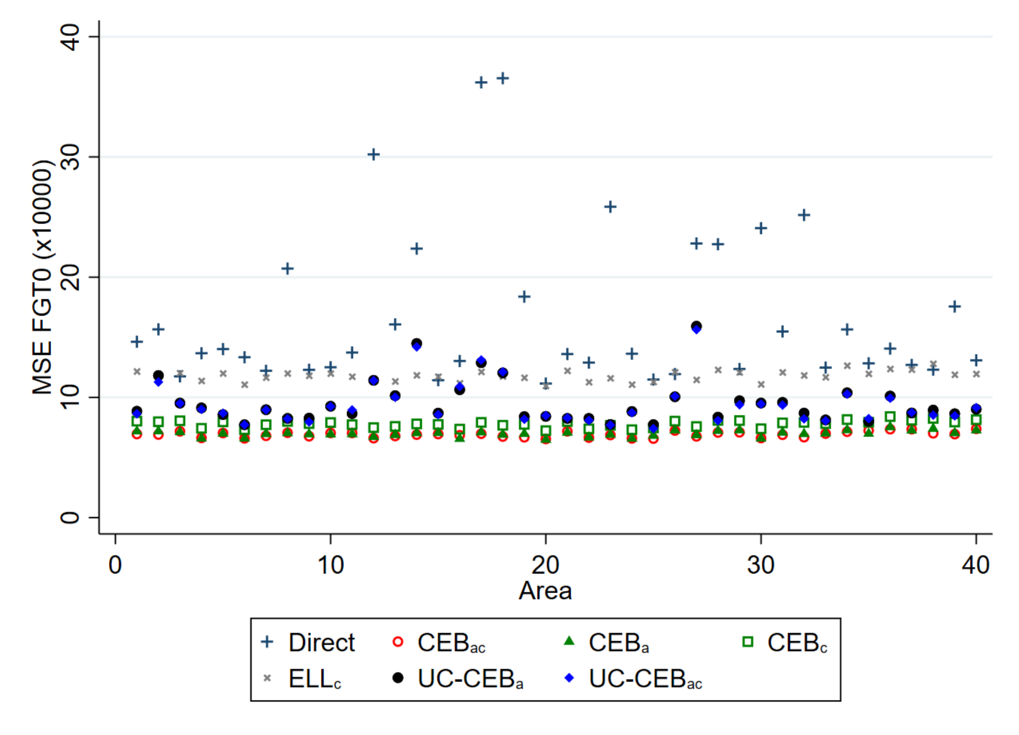


Figure 3: Empirical model bias of FGT0 estimators for $\sigma_{ac} = 0.05$ and $\sigma_a = 0.1$

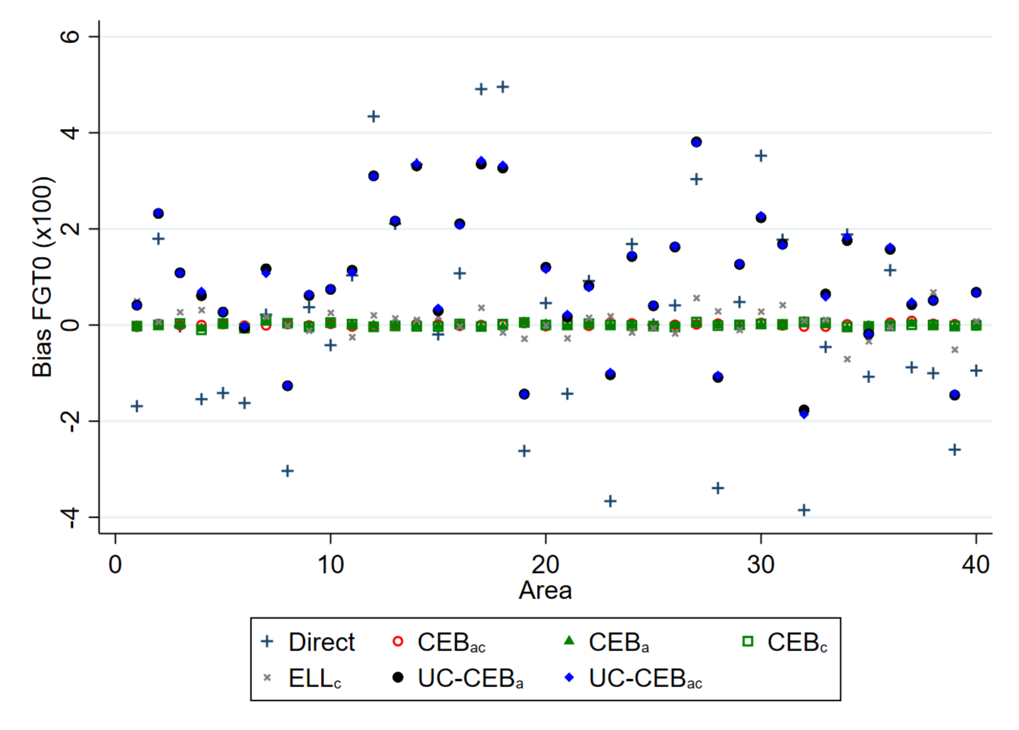


Figure 4: Empirical model MSE of FGT0 estimators; $\sigma_{ac} = 0.05$ and $\sigma_a = 0.1$

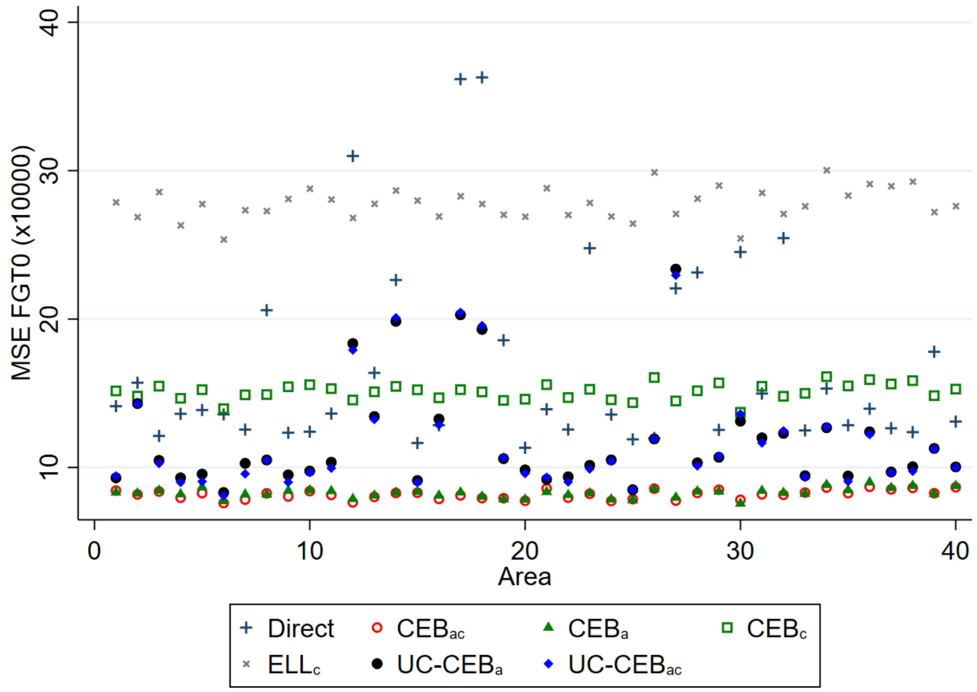


Figure 5: Omitted variable bias under unit-context models

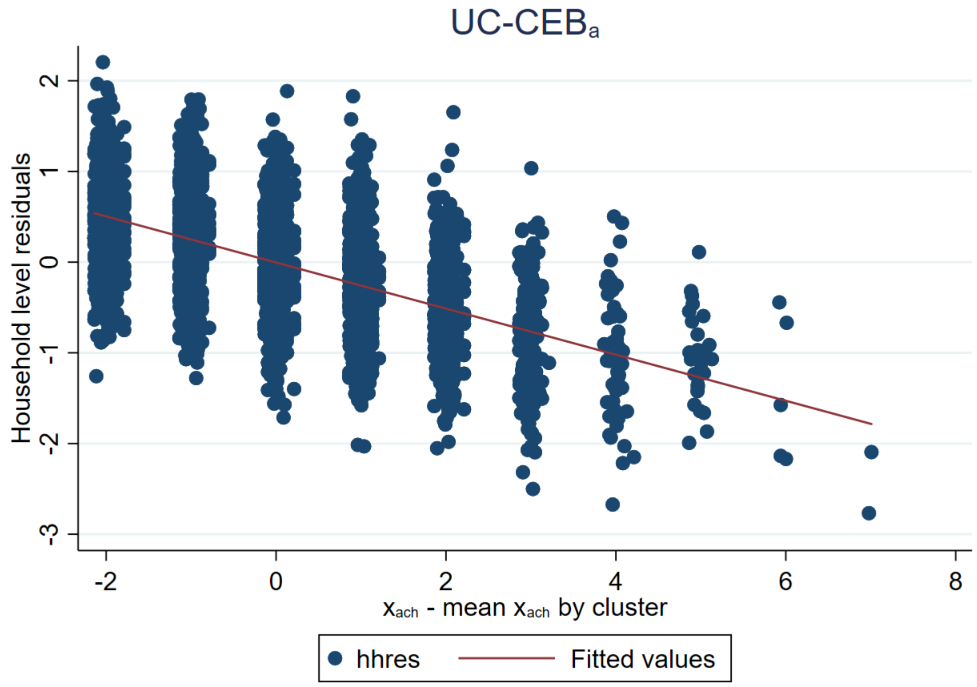


Table 1: Aggregate results across areas (FGT0)

	1	2	3	4
Model R^2	~0.42	~0.42	~0.42	~0.42
σ_a	0.05	0.1	0.05	0.1
σ_{ac}	0.1	0.05	0.1	0.05
σ_e	0.5	0.5	0.5	0.5
Sample per cluster	10 per all clusters	10 per all clusters	10 per 20% clusters	10 per 20% clusters
Population	20,000	20,000	100,000	100,000
Direct				
AAB (×100)	1.780	1.783	1.802	1.795
ARMSE (×100)	4.039	4.041	4.396	4.563
$\boldsymbol{CEB_{ac}}$				
AAB (×100)	0.020	0.023	0.018	0.023
ARMSE (×100)	2.630	2.859	2.493	2.191
$\boldsymbol{CEB_{a}}$				
AAB (×100)	0.020	0.024	0.021	0.017
ARMSE (×100)	2.642	2.876	2.515	2.237
$\boldsymbol{CEB_{c}}$				
AAB (×100)	0.024	0.029	0.045	0.020
ARMSE (×100)	2.793	3.885	4.731	2.614
$\boldsymbol{ELL_{c}}$				
AAB (×100)	0.238	0.236	0.144	0.139
ARMSE (×100)	3.433	5.269	5.048	2.768
$\boldsymbol{UC-CEB_{a}}$				
AAB (×100)	1.115	1.362	1.296	0.986
ARMSE (×100)	3.074	3.404	3.068	2.584
$\boldsymbol{UC-CEB_{ac}}$				
AAB (×100)	1.106	1.370	1.272	0.929
ARMSE (×100)	3.059	3.383	3.046	2.591

AAB: Average absolute bias; ARMSE: Average root MSE bias.[20] This misspecification is manifested as failure of the linearity assumption and the magnitude of this bias depends on the form of the covariates (see Figure 5 and appendix). Additionally, it will vary for each type of indicator. For example, when estimating mean income in a model without log transformation, linearity is preserved when aggregating and thus the approach is expected to perform well.[21]

However, the empirical MSEs for the unit-context models in most areas outperform those from ELL (see Table 1 for average results). In fact, under scenario 1 (Figure 2) in some areas the MSE for the unit-context models are only slightly worse than those of the CEB_a and CEB_{ac}. Under scenario 2, where the area effects have a higher variance relative to that of the cluster effect, which is likely the case in real world scenarios, the empirical MSE for the unit-context models show considerable variability for both fitting models. The results suggest the method could outperform direct estimates as well as misspecified models with random effects for the area level only, in terms of MSE, and may be an alternative under scenarios where census and survey data are not aligned. However, bias in these unit-context models may be a considerable concern. As one can see in the results from Table 1 the unit context method yields FGT0 estimators with an average absolute bias that is 56 times larger than the bias of the CensusEB estimators, and almost 5 times larger than ELL's average absolute bias.[22] In the sections that follow this is further explored.

4 Design-based simulation

In this section we present a design-based simulation experiment based on the Mexican Intra Censal Survey of 2015 (Encuesta Intracensal). The purpose of the simulation is to observe performance of the different methods under more realistic scenarios. The survey is carried out by the Mexican National Institute of Statistics and Geography (Instituto Nacional de Estadistica y Geografia - INEGI). The survey has a sample of 5.9 million households and is representative at the national, state (32 states) and municipal or delegation level (2,457 municipalities), as well as for localities with a population of 50,000 or more inhabitants.

The 2015 Intra Censal Survey questionnaire includes the following topics related to the housing unit: dwelling features, size and use of the dwelling, conditions for cooking, ownership and access conditions, access to water, sanitation facilities and sanitation, electric power, solid waste, equipment, appliances and automobile; and information and communication technologies (ICT). It also includes the following demographic information: total population and structure, birth registration, marital status, health services, ethnicity, education, economic characteristics, nonpaid work, migration, daily mobility, fertility and mortality, household composition, non-labor household income,

[20] The correlation is also present between $\bar{x}_{5ac}$ and other omitted deviations, z_{ach}.

[21] The empirical average relative root MSE of the mean for the UC approach is slightly higher than that of the CensusEB.

[22] The unit context method also yields mean welfare estimators that are 38 times more biased than Census EB variants (not shown).

food security, agricultural land use, relationship to the household head, indigenous language, occupation, economic activity, and accumulated education. The survey also includes indicators for states, municipalities, and counties.

One of the key features of the survey is its size and the fact that it includes a measure of income at the household level.[23] The inclusion of an income measure allows for a design-based validation of the different methods presented above. The next section describes how the survey is modified to create a census dataset and how samples are then drawn from the created census data with the goal of obtaining small area estimates of poverty at the municipal level.

4.1 Creating a census and survey

Because the goal of this exercise is to test how well the different methods perform under a real world scenario, the Intra Censal Survey is modified to mimic a Census in order to allow for a design-based simulation. The first step consists of randomly removing 90 percent of households that reported an income of 0. This is done to ensure that some households with an income of 0 are included in the population, but not as many as in the original data to make it more realistic.[24] In the second step, all municipalities with less than 500 households are removed, thus observations by municipality range from 501 to 23,513.[25] The final “Census” consists of 3.9 million households and 1,865 municipalities.

To draw survey samples, primary sampling units (PSU) need to be created.[26] Within each municipality, the original data’s PSUs are sorted[27] and joined until each created PSU has close to 300 households. Under the proposed approach, original PSUs are never split, just joined to others. Additionally, all original PSUs that were larger than 300 households are designated as a created PSU. The entire process yields 16,297 PSUs.

The resulting Census data is used to draw 500 survey samples to conduct a design-based simulation to establish how a method will perform over repeated sampling from a finite population (Tzavidis et al., 2018). The sampling approach reflects standard designs used in most face-to-face household surveys conducted in the developing world, such as those conducted by the Living Standards Measurement Study (LSMS) program of the World Bank (Grosh and Muñoz, 1996), with some simplification for convenience. First, the thirty-two states that comprise Mexico are treated as the intended domains of the sample. The main indicator of interest for the survey is the welfare measure: household per capita income. The desired level of precision to be achieved in the sample is assumed to be a relative standard error (RSE) of 7.5 percent for mean per capita income

[23] Income is defined as money received from work performed during the course of the reference week by individuals of age 12 or older within the household.

[24] Welfare values of 0 and or missing are a common feature of household surveys.

[25] The median municipality has 1,613 households

[26] The Intra Censal Survey has its own PSUs, however many of these have too few observations to properly work as PSUs in the created “Census” data.

[27] This assumes that the clusters’ numbering is tied to how proximate each cluster is to one and other.

in each state.[28] A two-stage clustered design is assumed with clusters (defined above) serving as primary sampling units (PSUs) selected in the first stage within each domain and then a sample of 10 households within each cluster is selected in the second stage. With these design features established, the trimmed census data was analyzed to identify the parameter estimates for per capita income (mean and standard deviation) for each state and the target standard error implied by the parameter estimates that corresponds to an RSE of 7.5 percent. The minimum sample size required for state s, given these parameters, under simple random sampling (SRS) design is then obtained as:

$$n_s^{SRS} = \left(\frac{\sigma_s}{\bar{y}_s \times RSE^{tgt}}\right)^2 = \left(\frac{\sigma_s}{\bar{y}_s \times 0.075}\right)^2$$

where σ_s is the standard deviation of per capita income in state s, RSE^{tgt} is the target standard error of 7.5 percent in state s, and $\bar{y}_c$ is the mean per capita income in state s. The minimum sample size under SRS must then be adjusted to account for the clustered design. This design effect due to clustering is accounted for by estimating the intra-cluster correlation (ICC) for per capita income within each state using the trimmed census data. The ICC estimates can then be applied to the SRS size obtained above to arrive at the minimum sample size for state s under the clustered design employed here, given by:

$$n_s = n_s^{SRS} \times DEFF_s = n_s^{SRS} \times (1 + (n_{psu} - 1)\,\rho_s)$$

where $DEFF_s$ is the design effect in state s, n_{psu} is the number of households selected within each cluster (10 in this case), and ρ_s is the ICC for per capita income in state s. The minimum number of clusters to achieve n_s (assuming 10 households per cluster) was calculated and then the final (household) sample size established by multiplying the number of clusters by 10.[29]

Taking the sample size requirements from above as fixed, the sample in each simulation is selected in accordance with the two-stage design. PSUs within each state, referred to here as clusters, are selected with probability proportional to size (PPS) without replacement, where the measure of size is the number of households within the cluster. Then 10 households were selected within each cluster via simple random sampling. According to this design, the inclusion probability for household h in cluster c and in state s is approximated[30] as:

$$p_{sch} = \frac{N_{sc}C_s}{N_s} \times \frac{n_{sc}}{N_{sc}}$$

[28] The desired precision of an RSE of 7.5 percent is somewhat arbitrary but corresponds to precision targets in similar surveys and yields a sample of reasonable size.

[29] Full estimates and results for the sample size determination are available upon request.

[30] The equation used to calculate inclusion probabilities assumes sampling with replacement but is used here as an approximation of inclusion probabilities under PPS selection without replacement. This should provide a reasonable approximation in this case since there are a relatively large number of PSUs present in the frame.

where N_s is the total number of households in the census for state s, N_{sc} is the number of households in cluster c from state s, C_s is the number of clusters selected in state s, and n_{sc} is the number of households selected in cluster c within state s, which is fixed at 10.[31]

The sample size across the 500 samples is roughly 23,540. Under the proposed sampling scenario, not all municipalities are included, and the number of municipalities included varies from sample to sample, ranging between 951 to 1,020 municipalities. The median municipality included in a given sample, is represented by a sole PSU and thus its sample size is of 10 households.

4.2 Model selection

Model selection is conducted using the first sample drawn from the scenario detailed in the previous section. The target variable is household per capita income. However, this variable is highly skewed and to achieve an approximately normal distribution we test three transformations: i) natural logarithm,[32] ii) log-shift transformation, and iii) Box-Cox transformation of the natural logarithm.[33] As one can see in Figures 6, 7 and 8, the Box-Cox transformation as well as the log shift fix the skewness in the distribution of model residuals that appears after taking the natural logarithm of per capita income.[34]

The goal of the model selection process is to arrive at a model that only includes stable covariates. Under each transformation, model selection is done using a least absolute shrinkage and selection operator, commonly known as lasso,[35] where the candidates for covariates include household characteristics and characteristics at the PSU, municipal and state level. Two models are selected: i) a model that includes household characteristics and characteristics at the PSU, municipal, and state levels and ii) another model that only includes characteristics at the PSU, municipal and state levels. The second model is used for the unit-context approach. All household level characteristics that are included also as aggregates at the PSU, municipality, and state levels have been previously standardized to ensure that these have mean 0 and standard deviation 1 for each PSU, municipality and state respectively. Note that aggregated covariates have been obtained from the "census," thus, the aggregated covariates will be the same within the sample and the census. The lasso model selection process applied here ignores the error structure presented in (1) and (2) and does not ensure that selected covariates will be significant under the assumed models. Consequently, after the initial lasso model selection, model (1) is fit using Henderson's Method III (with sampling weights) and random effects specified at the municipality level. Because the resulting model may

[31] The design weight for each household is simply the inverse of the inclusion probability. In a typical survey, the design weights would be further adjusted for nonresponse and calibrated to known population characteristics. However, since the sampling is only a simulation exercise, there is no nonresponse and thus no nonresponse adjustment is required. Calibration or post-stratification could be performed but was not implemented to simplify the process.

[32] In any given sample, roughly 11 observations have an income of 0, these are assigned an income of 1 prior to transformation.

[33] For more details on transformations, see Tzavidis et al. (2018).

[34] Figures make use of sample from a two-stage clustered design.

[35] Model is selected using 20 fold cross validation and shrinkage parameter (λ) that is within 1 standard error of the one that minimizes the cross validated MSE.

Figure 6: Histogram of residuals from unit level onefold nested error model fitted to Nat. log. of per capita income (municipal random effects)

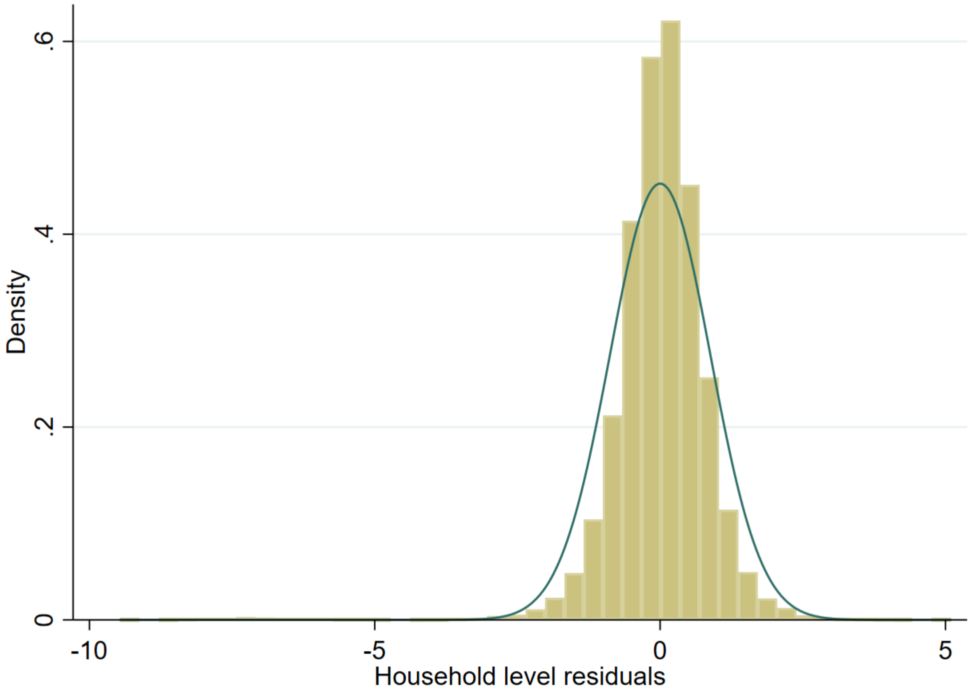


Figure 7: Histogram of residuals from unit level onefold nested error model fitted to log-shift transformation of per capita income (municipal random effects)

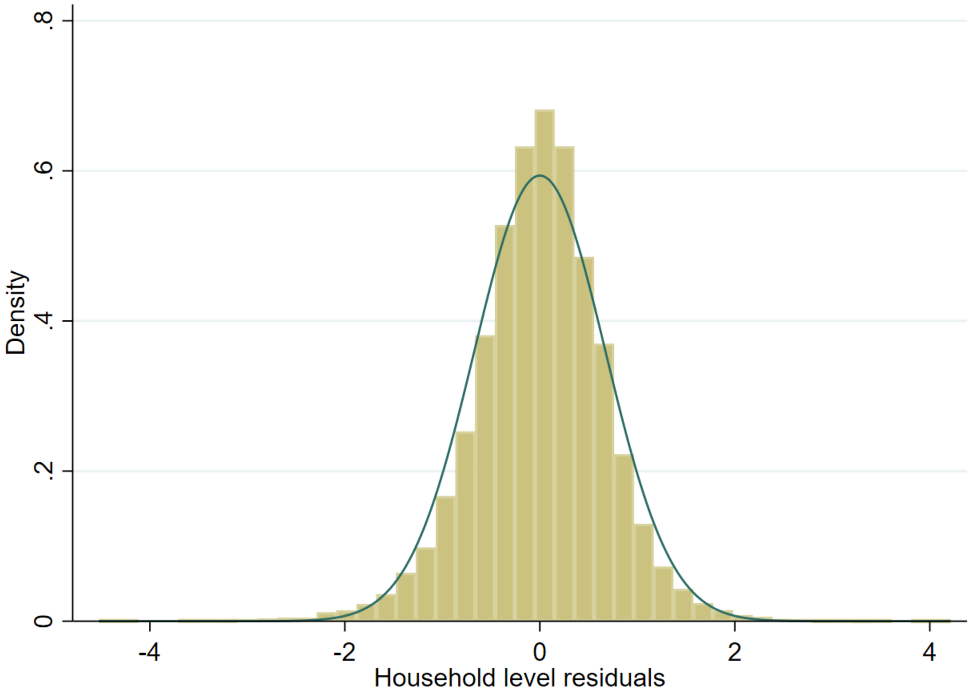


Figure 8: Histogram of residuals from unit level onefold nested error model fitted to Box-Cox of Nat. log. of per capita income (municipal random effects)

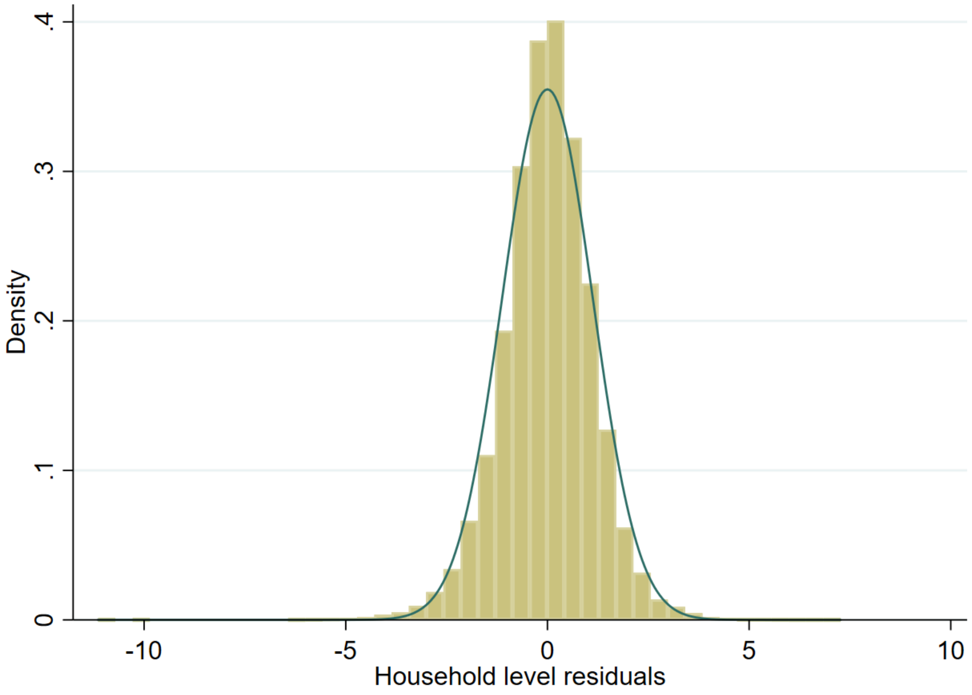


Figure 9: Normal Q-Q-plot of unit level onefold nested error model - household level and predicted subdomain effects (Nat. log transformation)

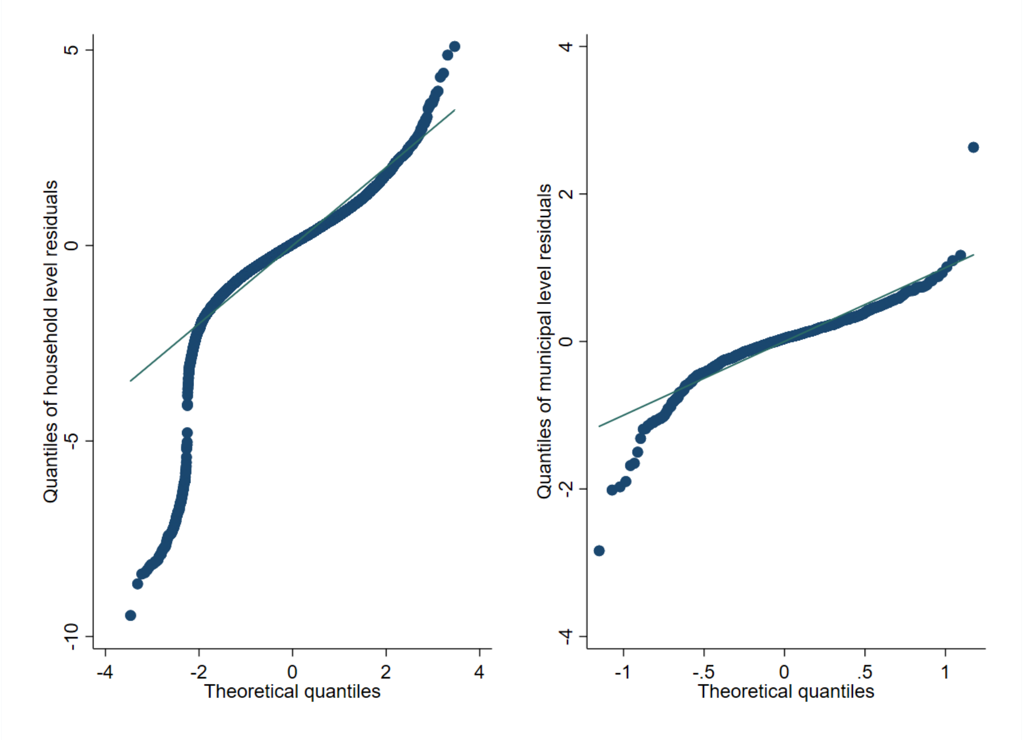


Figure 10: Normal Q-Q-plot of unit level onefold nested error model - household level and predicted subdomain effects (log-shift transformation)

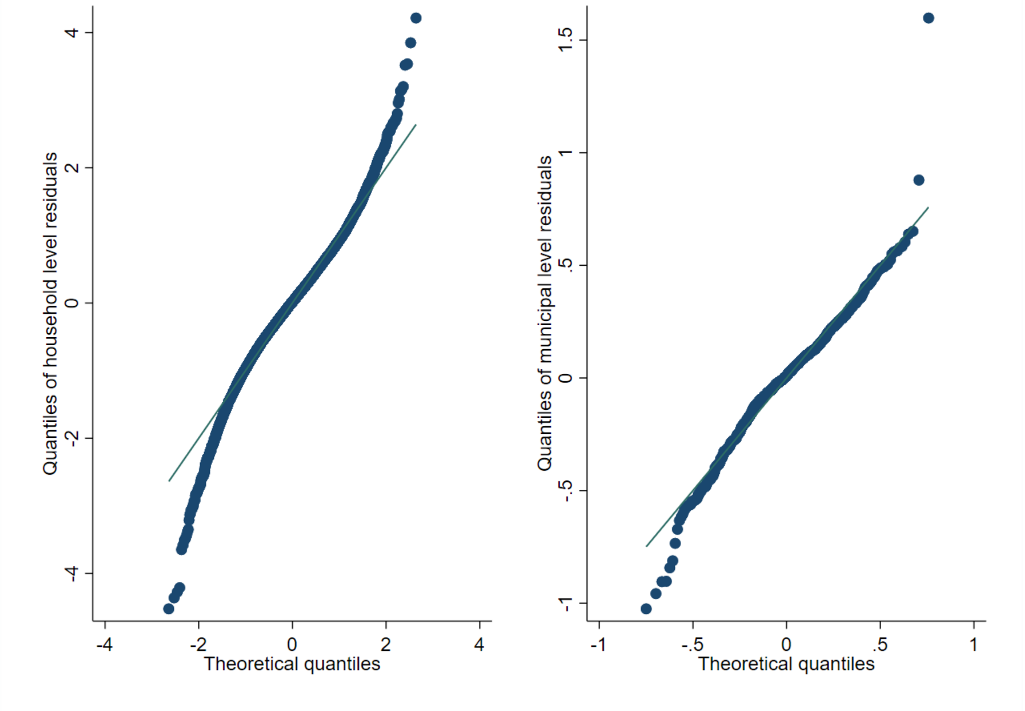
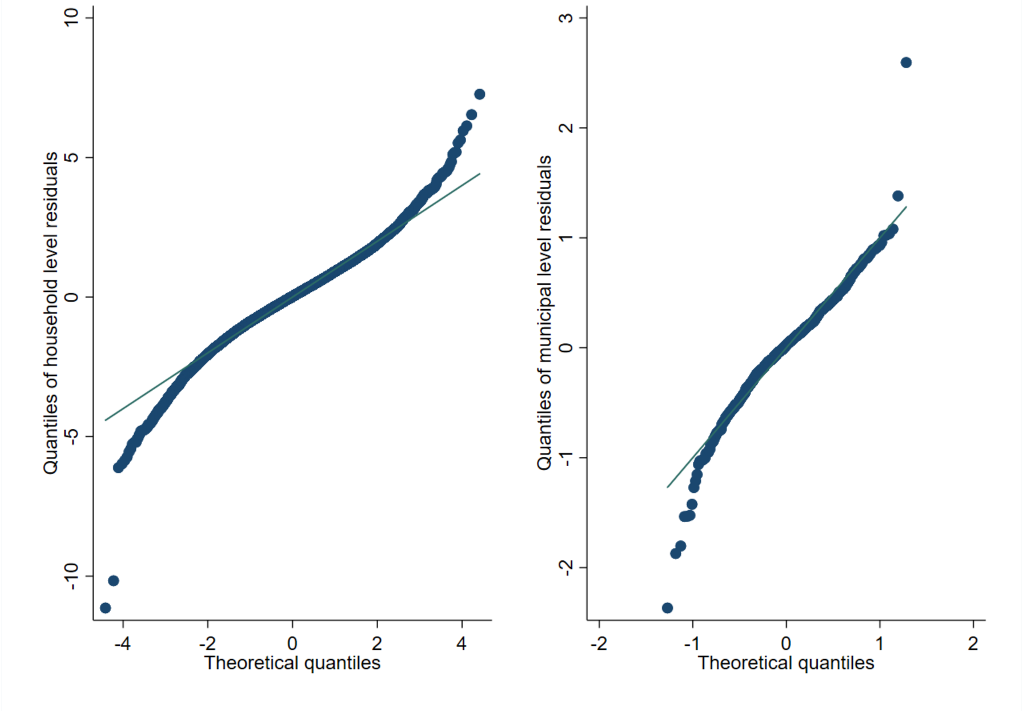


Figure 11: Normal Q-Q-plot of unit level onefold nested error model - household level and predicted subdomain effects (Box-Cox of nat. log transformation)



still include covariates that are not significant under the fitted model, all non-significant (at the 1% level) covariates are removed sequentially.[36] Finally, we check for multicollinearity and remove covariates with a variance inflation factor (VIF) above 3.

Fitted models using the first sample and the described model selection are presented in Tables 3 and 4 in the appendix. Since in small area estimation models are used for prediction, assessing the predictive power of the model is important. This may be measured using the coefficient of determination, R^2 (Tzavidis et al., 2018). For the household level model used in this exercise the R^2 is close to 0.45 while that of the unit-context model centers around 0.25.

Figures 9, 10, and 11 represent checks on the normality assumptions using normal Q-Q-plots of household level residuals and estimated area effects for the onefold unit level model with random municipality location effects. The resulting plots for the unit-context models can be seen in the appendix figures 26, 27, and 28. The natural logarithm transformation (Fig. 9) presents evidence of deviation from normality. Marhuenda et al. (2017) notes in applications that use real data the exact fit to a distribution is barely met; however the Box-Cox and log-shift transformations provide considerably better normal approximations (see Figures 10 and 11).[37]

4.3 Results

Once the models to be used have been selected and $L = 500$ samples under two-stage sampling described in section 4.1 have been taken from the "census" created using the Mexican Intra Censal Survey of 2015 we obtain estimates using the different considered models. The target parameters for this simulation are mean welfare and the Foster et al. (1984) - FGT class of decomposable poverty measures for municipalities present in the census. The true values of the target parameters at the municipal level are based on the census data.

As a first step, we compare results using the three transformations discussed in the previous section. Figures 12 and 13 show box plots of empirical absolute relative bias and MSE of CensusEB estimates of poverty rates based on model (1) under each transformation. The Box-Cox and log-shift transformation yield estimates that are not only less biased, but also results with a lower empirical design MSE. As can be seen in the results from the first 3 columns in Table 2 the log-shift transformation yields aggregate results somewhat preferable to the Box-Cox. Additionally, in Table 2 it is quite clear that the natural logarithm may actually result in aggregate results that present considerably smaller gains over direct ones and thus model and residual checks should always be done to ensure adequate transformations are used to avoid such an outcome. The rest of the discussion will focus on estimates obtained from the log-shift transformation applied to the two-stage samples.[38]

[36] The removal is done sequentially starting with the least significant covariate, then the model is fit again without the covariate. The process is repeated until all covariates in the model are significant at the 1% level.

[37] Box-Cox and log-shift transformations are available under R's `sae` package as well as under Stata's `sae` package. In the World Bank's `PovMap` software, the only available transformation was natural logarithm. However, `PovMap` does allow for drawing residuals from the empirical distribution as well as from a Student's t distribution.

[38] Figures under other transformations are available upon request.

Figure 12: Box plots of Census EB FGT0 design bias for unit level onefold nested error model estimates for fitted model 1 and two-stage samples (Municipal level)

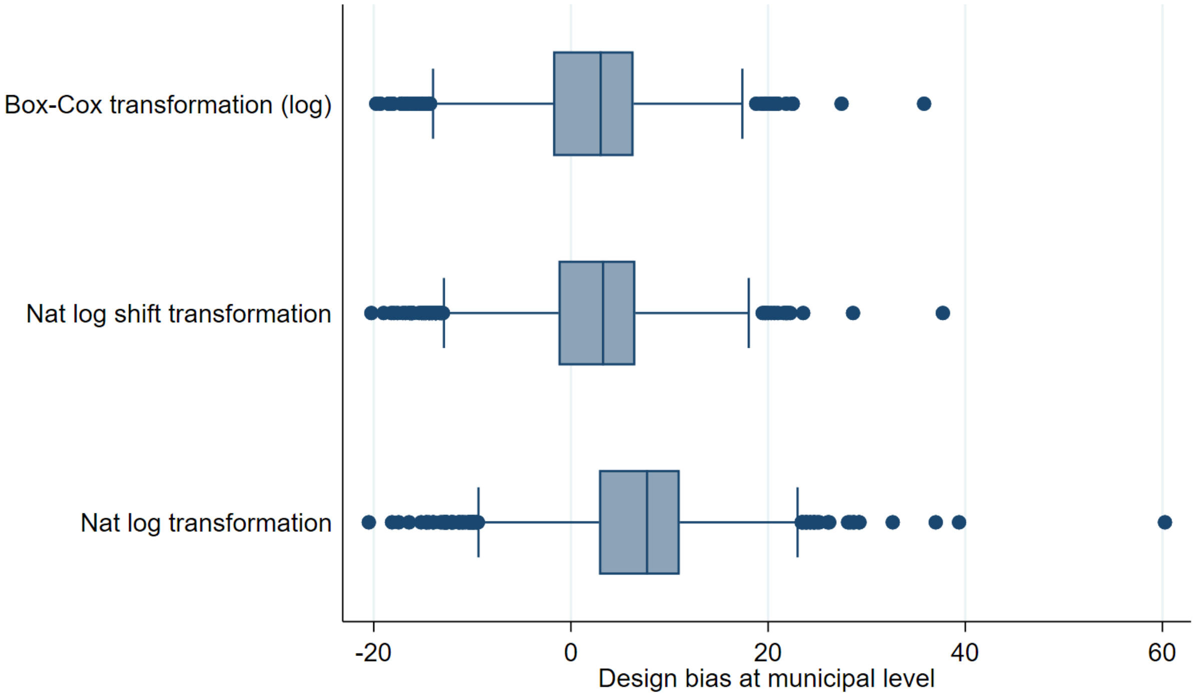
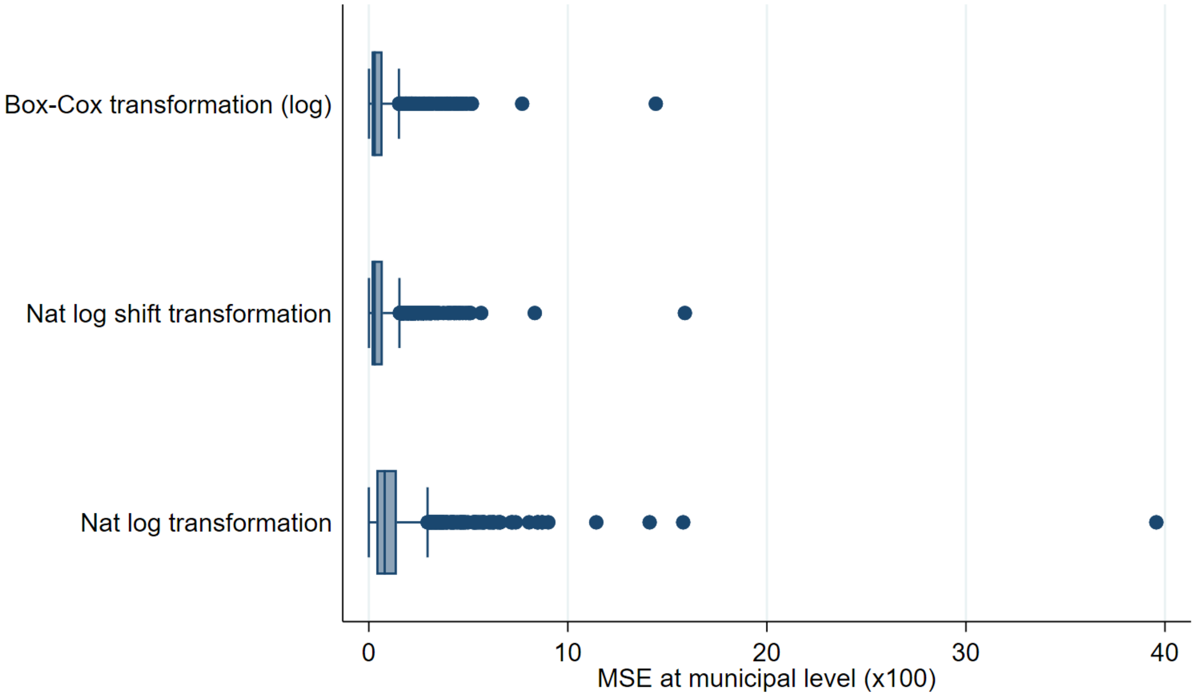


Figure 13: Box plots of empirical MSE of unit level onefold nested error model CensusEB FGT0 estimates for fitted model 1 and two-stage samples (Municipal level)



We consider estimators based on onefold and twofold nested error models, including unit-context (UC) models with only aggregated covariates, and direct estimators. Concretely, we consider:[39]

Direct Estimates

1. Direct estimates from the survey samples for each municipality. These are calculated with weights from the considered design. Specifically, the calculation of FGT indicators under the two-stage design detailed in subsection 4.1 using the inclusion probabilities is obtained as:

   $\hat{\tau}_a^{direct} = \sum_h \frac{p_{ach}}{\sum p_{ach}} F_{ach}$;

   where, p_{ach} is the inclusion probability for household h in cluster c and in municipality a, and F_{ach} is the FGT or welfare measure of interest for household h in cluster c and in municipality a.[40]

Unit-level models:

1. CEB_a: Model fit is done using Henderson's Method III (with sampling weights) and estimates are obtained using CensusEB as noted in CMN (2020). The fitted model reads:

   $y_{ach} = x_{ach}\beta + z_{ac}\alpha + t_a\omega + g_s\lambda + \eta_a + \varepsilon_{ach}$;

   where x_{ach} is a vector of household specific characteristics, z_{ac} contains cluster level characteristics, t_a includes municipality level characteristics and g_s is composed of state level characteristics. The random effects, η_a, are specified at the municipal level.

2. CEB_c: Model fit is done using Henderson's Method III (with sampling weights) and estimates are obtained using Census EB as noted in CMN (2020). The difference with model (1) is that random effects are specified at the PSU level. That is, the fitted model is:

   $y_{ach} = x_{ach}\beta + z_{ac}\alpha + t_a\omega + g_s\lambda + \eta_{ac} + \varepsilon_{ach}$;

   where, η_{ac}, is a random effect for cluster c within municipality a.

3. CEB_{ac}: Model fit is done using REML and estimates are obtained using CensusEB as noted in Marhuenda et al. (2017). Fitted model follows:[41]

   (a) $y_{ach} = x_{ach}\beta + z_{ac}\alpha + t_a\omega + g_s\lambda + \eta_a + \eta_{ac} + \varepsilon_{ach}$; random effects are specified at the municipality and PSU level.

4. CEB_{sa}: Similar to 4, however random effects are specified at the municipal and state level. The goal here is to borrow strength from the state for municipalities that are not in the sample which takes a cue from Marhuenda et al. (2017).

[39] Every model includes a constant term.

[40] Note that depending on the sampling strategy, it is likely that we do not have direct estimates for all areas.

[41] All twofold nested error methods are fit without the use of probability sampling weights and are thus not comparable to those that use survey weights.

5. ELL_c: under the same model as in 2. In cases where we use a transformation different from the natural logarithm, random location effects and household residuals are drawn from their empirical distribution.

Unit-context models:

1. $UC-CEB_a$: Unit-context model originally proposed by Nguyen (2012) but with EB, similar to Masaki et al. (2020). Model fit is done using Henderson's Method III (with sampling weights) and estimates are obtained using CensusEB as noted in CMN (2020). The fitted model follows:

   $y_{ach} = z_{ac}\alpha + t_a\omega + g_s\lambda + \eta_a + \varepsilon_{ach}.$

2. $UC-CEB_{ac}$: Unit-context model fit is done using REML and estimates are obtained using CensusEB as noted in Marhuenda et al. (2017). Fitted model is:

   $y_{ach} = z_{ac}\alpha + t_a\omega + g_s\lambda + \eta_a + \eta_{ac} + \varepsilon_{ach};$

3. $UC-CEB_{sa}$: Similar to model 8, however random effects are specified at the municipal and state levels. Just like 5, the goal here is to borrow strength from the state for municipalities that are not in the sample.

4. $UC-ELL_c$: ELL estimates under model in $UC-CEB_a$ but random effects specified at the PSU level, as was originally proposed by Nguyen (2012) and then by Lange et al. (2018). In cases where we use a transformation different from the natural logarithm, random location effects and household residuals are drawn from their empirical distribution.

   $y_{ach} = z_{ac}\alpha + t_a\omega + g_s\lambda + \eta_{ac} + \varepsilon_{ach}.$

The chosen measures to evaluate performance of the considered predictors are bias, MSE, and root MSE obtained as:

$$Bias\left(\hat{\tau}_a^j\right) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\tau}_a^{j(l)} - \tau_a)$$

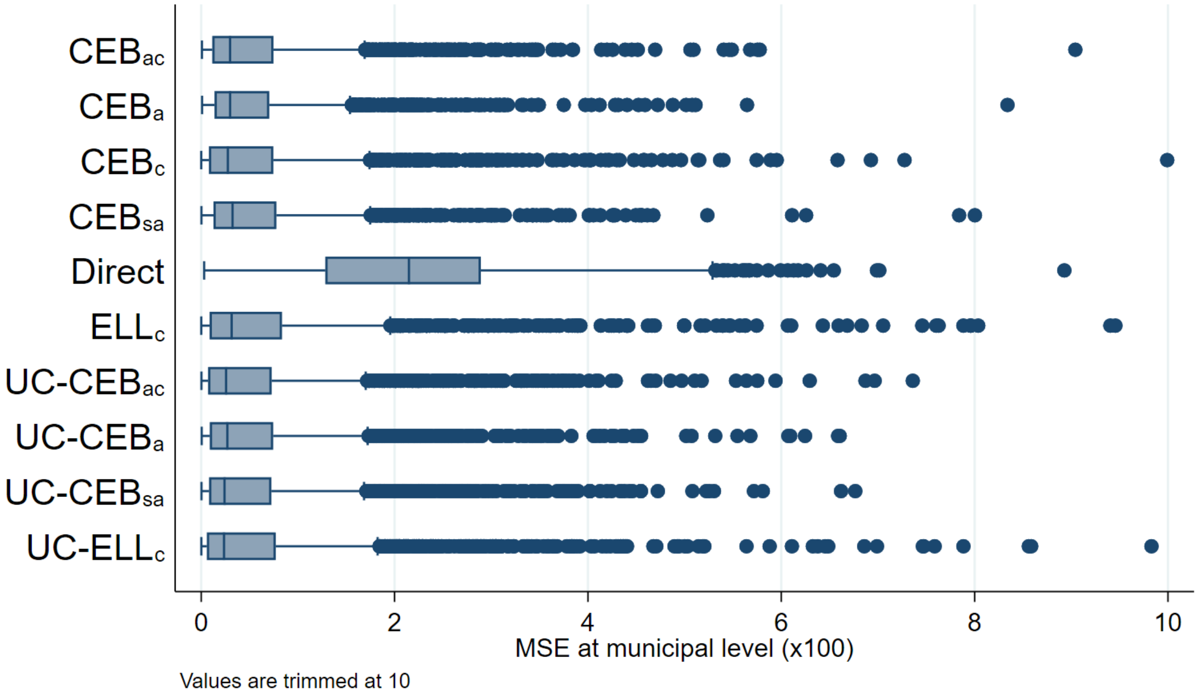
$$MSE\left(\hat{\tau}_a^j\right) = \frac{1}{L}\sum_{l=1}^{L}(\hat{\tau}_a^{j(l)} - \tau_a)^2$$

$$RMSE\left(\hat{\tau}_a^j\right) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(\hat{\tau}_a^{j(l)} - \tau_a)^2}$$

where j stands for one of the methods: CEB_a, CEB_c, *direct*, CEB_{ac}, CEB_{sa}, ELL_c, $UC-CEB_a$, $UC-CEB_{ac}$, $UC-CEB_{sa}$, and $UC-ELL_c$; τ_a denotes the true population parameter for municipality a.[42] Formal evaluation of the MSE estimators is not undertaken here since it is beyond the scope of the exercise and computationally intensive as it would require obtaining a bootstrap MSE for each of the 500 samples.

In section 3.1, like Marhuenda et al. (2017), we noticed how the relative size of the random effects affected the precision of the different methods. The first sample from the simulation experiment under two-stage sampling is used to assess the magnitude of the random effects under the unit level model. The values of $\hat{\sigma}^2_{ac}$ and $\hat{\sigma}^2_a$ under the twofold nested error model with a log shift transformation for this sample are equal to 0.021 and 0.013, respectively. The value of $\hat{\sigma}^2_{ac}$ when specifying random effects only at the cluster level is equal to 0.072, and $\hat{\sigma}^2_{ac}/\hat{\sigma}^2_e$ is equal to 0.054. When specifying random effects only at the municipality level, $\hat{\sigma}^2_a$ is equal to 0.022, and the $\hat{\sigma}^2_{ac}/\hat{\sigma}^2_e$ ratio is equal to 0.045.

Figure 14: Box plots of empirical design MSE for FGT0 under two-stage sampling (Nat log shift transformation)



In Figure 14 we can clearly see in general all SAE methods outperform the direct estimates in terms of design MSE.[43] In terms of design bias (Fig. 15), direct estimators are very close to being unbiased under the design and would likely converge to the true estimate as the number of simulations increase. On the other hand, model-based estimators are all biased under the design. The result is similar to the one presented by Marhuenda et al. (2017), where gains in MSE for

[42] Note that here there is only one true population parameter because our census is fixed. In the model-based simulation we assume the target parameters are random.

[43] See the appendix figure 31 for the untrimmed version of the figure.

Table 2: Aggregate results for 1,865 municipalities in “Census” (FGT0) - Results from 500 samples

	1	2	3	4
Transformation	Box-Cox	Nat. log.	Log. Shift	Log. Shift
Direct				
AAB (×100)	11.314	11.314	11.314	9.722
ARMSE (×100)	14.051	14.051	14.051	11.997
$\boldsymbol{CEB_{ac}}$				
AAB (×100)	6.277	8.642	6.273	5.393
ARMSE (×100)	6.580	9.169	6.574	5.964
$\boldsymbol{CEB_{sa}}$				
AAB (×100)	6.382	8.953	6.380	5.740
ARMSE (×100)	6.695	9.414	6.687	5.997
$\boldsymbol{CEB_{a}}$				
AAB (×100)	6.080	8.800	6.092	5.395
ARMSE (×100)	6.584	9.525	6.589	6.034
$\boldsymbol{CEB_{c}}$				
AAB (×100)	6.253	8.690	6.277	6.054
ARMSE (×100)	6.363	8.847	6.384	6.181
$\boldsymbol{ELL_{c}}$				
AAB (×100)	6.685	8.781	6.685	6.961
ARMSE (×100)	6.820	8.854	6.820	7.087
$\boldsymbol{UC-CEB_{a}}$				
AAB (×100)	6.171	7.531	6.016	5.982
ARMSE (×100)	6.607	8.274	6.461	6.672
$\boldsymbol{UC-CEB_{ac}}$				
AAB (×100)	6.121	7.596	5.998	5.926
ARMSE (×100)	6.446	8.186	6.332	6.652
$\boldsymbol{UC-ELL_{c}}$				
AAB (×100)	6.250	8.019	6.250	7.581
ARMSE (×100)	6.421	8.100	6.421	7.758
$\boldsymbol{UC-CEB_{sa}}$				
AAB (×100)	6.002	7.539	5.875	5.937
ARMSE (×100)	6.414	8.284	6.294	6.570

AAB: Average absolute bias; ARMSE: Average root MSE

ELL under Box-Cox and log shift, model is still for Nat. log. but errors are drawn from the empirical distribution.

Columns 1, 2, and 3 under two-stage samples (Section 4.1)

Column 4 results are under 1% SRS by PSU, all PSUs are included

model-based estimators are achieved at the expense of design bias. As an additional check, we add simulations where $L = 500$ samples are taken, each consisting of a 1% SRS without replacement in every PSU within the fixed census population. Aggregate results for this scenario are presented in column 4 of Table 2 and box plots for MSE and bias are presented in appendix Figures 29 and 30, respectively. The first thing to note in these figures is there are fewer extreme outliers when compared to results from the two-stage sample scenario in Figure 31 in the appendix. Despite fewer outliers, the results under the additional sampling scenario mimic those from two-stage samples. However, under the two-stage sampling used here, direct estimates for most municipalities are not available across all 500 samples (see subsection 4.1). Consequently, direct estimates are not included under the remaining figures that discuss results from the two-stage samples.

Under the simulation experiment with two-stage sampling, the empirical MSE of ELL_c estimates, where the location effect has been specified at the PSU level, appear to have a tighter spread than direct estimates (Fig. 14). Consequently, though ELL appears to perform relatively well, relative to direct estimates, the number of ELL outliers with a high MSE is considerable and ELL_c performs worse than all other small area approaches. This result was not expected, as the FGT0 results for $UC - ELL_c$ appear to perform better than the traditional ELL_c which has a considerably better model fit.[44] Nevertheless, $UC - ELL_c$ does very poorly on mean welfare where it has an MSE that is as large as that of direct estimates and is also considerably biased (see appendix Table 7). These results also hold under the simulation experiment with the 1% SRS by PSU samples.

As expected, CEB_a estimates show a considerably tighter MSE spread than direct estimates, but still with outliers. Another interesting finding is that under CensusEB but with random effects specified at the PSU level (CEB_c), the empirical MSE is tighter than that of the direct estimates and also displays better properties than ELL_c as seen in Table 2. However, given the discussion in the model-based validation and the results shown here, the results with the random effects at the municipality level are preferred over the ones where the effect is specified at the PSU level.

Perhaps the most surprising result is the low MSE of the $UC - CEB_a$ method with only contextual variables and the other $UC - CEB$ variants, as proposed by Masaki et al. (2020). The results display a tight spread and the results from Table 2 corroborate the finding. Nevertheless, beyond UC models being an alternative if contemporaneous census data is not available, the method presents a couple of advantages over traditional Fay Herriot (FH) area level models under sampling scenarios that follow a two-stage design like the one considered here. First, as noted toward the end of section 4.1, the majority of municipalities are represented by 1 PSU and thus have quite small sample sizes which makes the likelihood of direct FGT0 estimates being equal to 0 or 1 much higher. Under these cases, the method from Nguyen (2012) is a valid alternative to FH because FH is not applicable in these municipalities with a sampling variance of 0. An additional advantage is that the model can be used for multiple indicators whereas the FH requires a model for each indicator considered.

[44] Note that under the MI inspired bootstrap of ELL, a very poor model fit will be heavily penalized and will likely yield noise estimates for $UC - ELL_c$ that fare worse than direct estimates in most applications.

A possible explanation for the good performance of unit-context models under the considered experiments could be due to estimated random effects, where $\hat{\sigma}^2_{ac} > \hat{\sigma}^2_a$, which coincides with the better performing scenario in subsection 3.1. Although this observation is likely a coincidence since the performance of unit-context models depends on the particular covariates included in the model and on the shape of the target indicator. Under the Mexican data with the covariates used, the unit-context model performs rather well. However, the method considerably lags all others for estimation of the welfare mean by area. A similar result is observed in the results presented in subsection 3.1.[45]

Under the simulations of column 3 of Table 2, the gains in MSE for the unit-context methods appear to come in municipalities with larger populations.[46] This is expected because in our sampling scenarios, municipalities with a larger population are more likely to be included in all the samples. For example, in the census there are 16,293 PSUs spread over 1,865 municipalities. The median municipality in the created census has 7 PSUs, and the median municipality in a given two-stage sample is only represented by one PSU. Consequently, it is not surprising the unit-context variants under-perform relative to unit models in terms of bias and MSE in municipalities with smaller populations, as can be observed in Figure 16. For municipalities with larger populations, and hence those likely to consist of more PSUs and more of these PSUs in the sample, direct estimates for FGT0 are more precise and unit-context models begin to catch up to unit models (see Fig. 32 in the appendix).

Under the simulations from a 1% SRS by PSU, shown in Table 2, column 4 and in the appendix Figures 29 and 30, we notice the unit models (CEB_a) present a slight upward bias that seems to become more pronounced in more populous municipalities (Fig. 17). Notice that some bias is acceptable since gains in MSE are achieved at the expense of bias, however in this case the presence of outliers seems to lead to increased upward bias that affects more municipalities as we move to more populous deciles (box-plots for CEB_a in Fig. 17). Unit-context ($UC-CEB_a$) models on the other hand, appear to have a downward bias. In the box-plots in Figure 17 for $UC - CEB_a$, the bias in lower deciles is downward and as we move to upper deciles the downward bias is considerably reduced.

As noted in section 3.1, the problem faced by unit-context models is omitted variable bias manifested as a lack of linearity (see Fig. 18).[47] Figure 19 further presents the issue. Under CEB_a residuals appear to follow a random pattern, however under $UC-CEB_a$ households in municipalities represented by only one PSU in the sample will all have the same linear fit. This manifests itself in the figures for $UC-CEB_a$ as a column of vertical dots. The problem can also be observed

[45] See Table 7 in the appendix.

[46] Similar findings are obtained under the simulations for column 1.

[47] The true model includes household size at the household level. Consequently, it is a determinant of the dependent variable. Household size at the household level (x_{ach}) can be broken down into $z_{ach} - \bar{x}_{ac}$, thus the omitted component, z_{ach}, is also a determinant of the dependent variable. The unit-context model only includes the PSU average household size ($\bar{x}_{ac}$) obtained from the census as a covariate. When a given survey sample is taken, the non included covariate (z_{ach}) is correlated with the PSU average household size and the dependent variable.

Figure 15: Boxplot of empirical design bias for FGT0 under two-stage sampling (Nat log shift transformation)

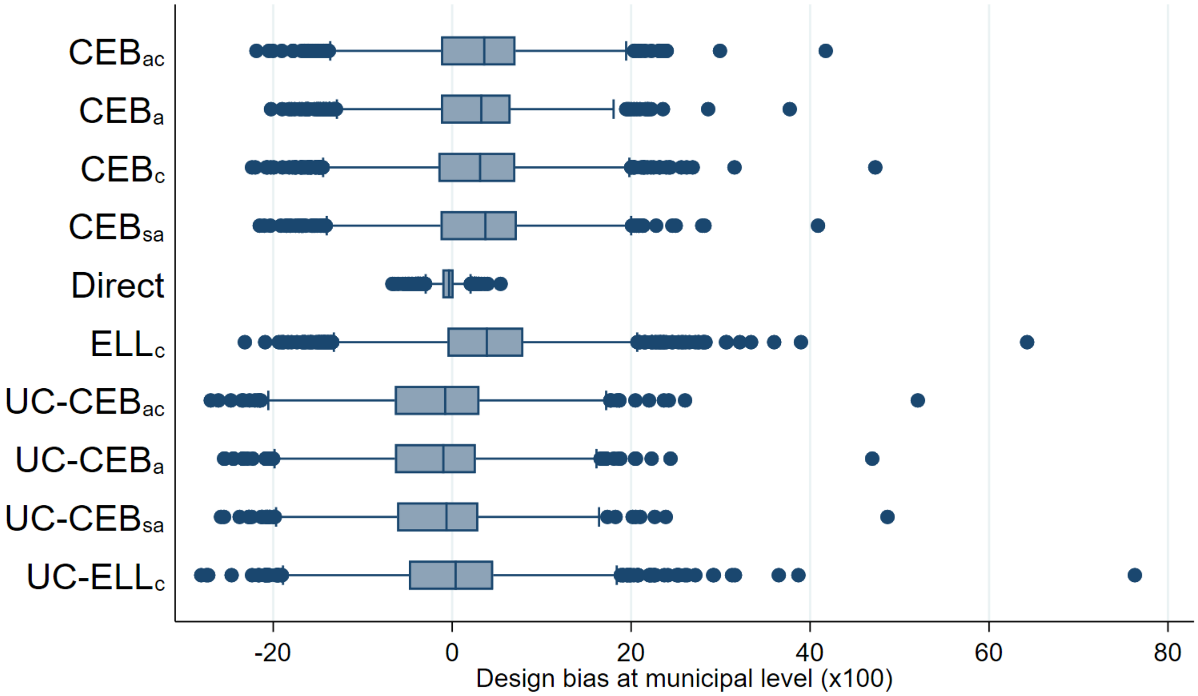


Figure 16: Average empirical MSE for FGT0 under two-stage sampling by municipality population deciles (Nat log shift transformation)

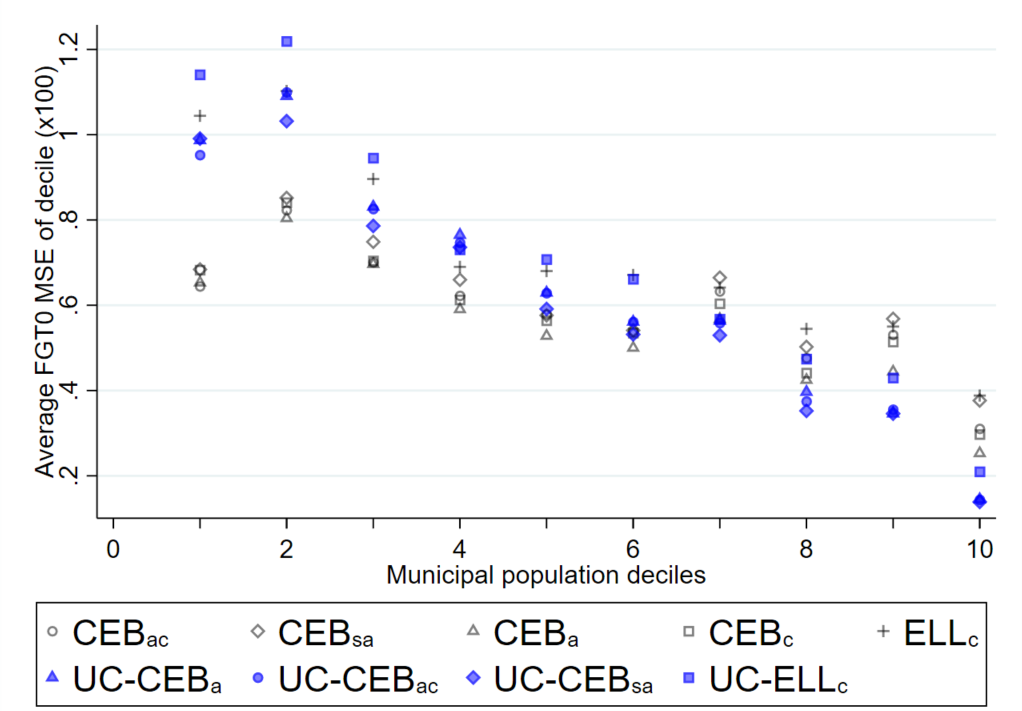


Figure 17: Box plots of design bias under 1% SRS by PSU sampling, by municipality population deciles (Nat log shift transformation)

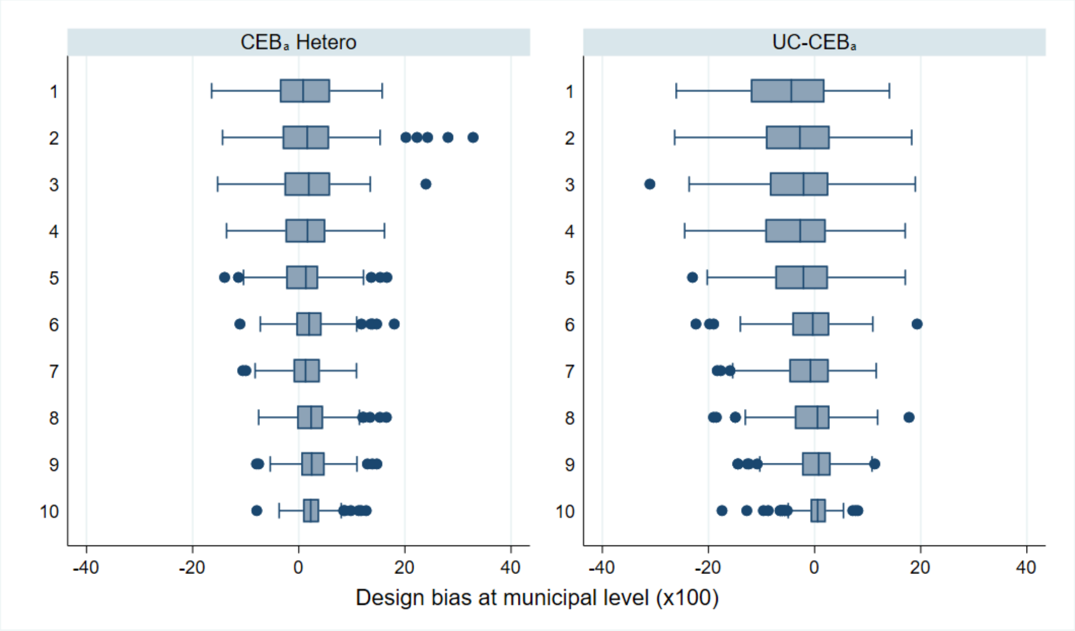


Figure 18: Omitted variable bias under unit-context models

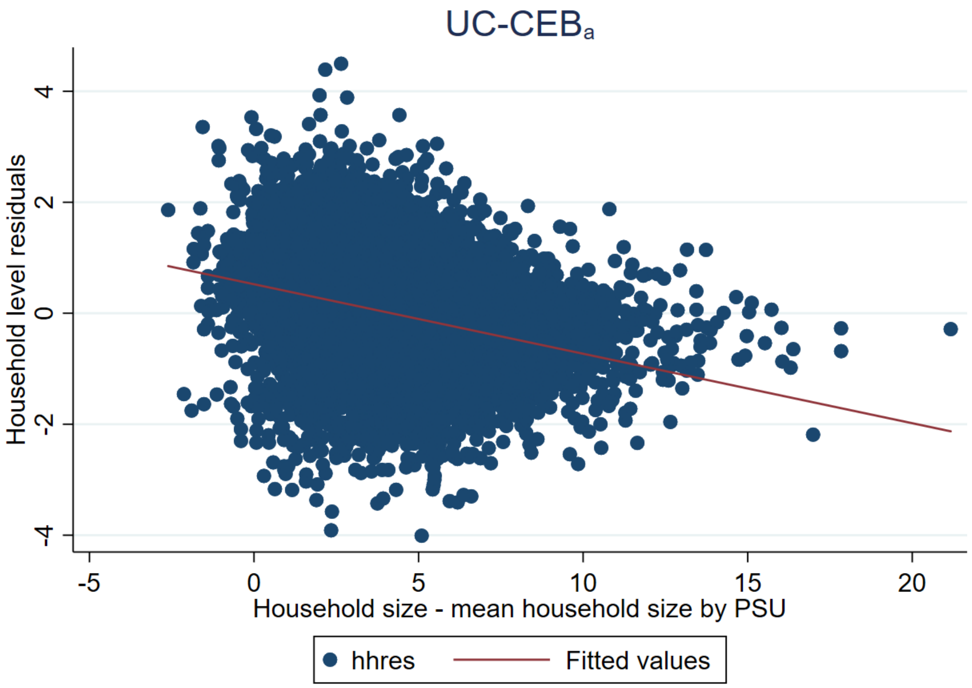
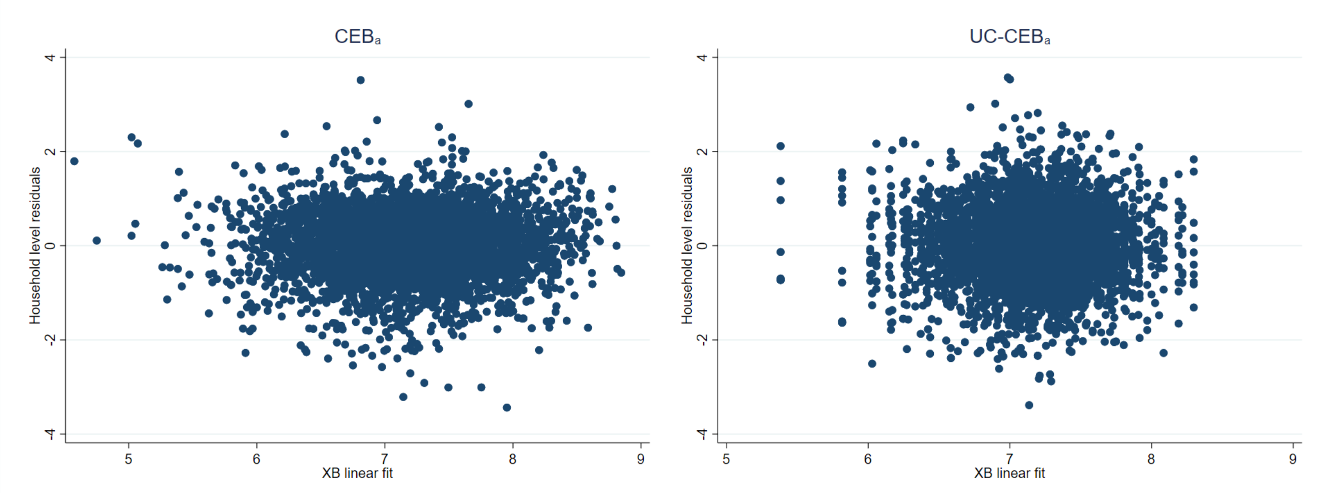


Figure 19: Household residuals plotted against linear fit under two-stage sampling (Nat. log shift transformation)



Only includes municipalities represented by one PSU

Figure 20: Linear fit plotted against municipalities under two-stage sampling (Nat. log shift transformation)

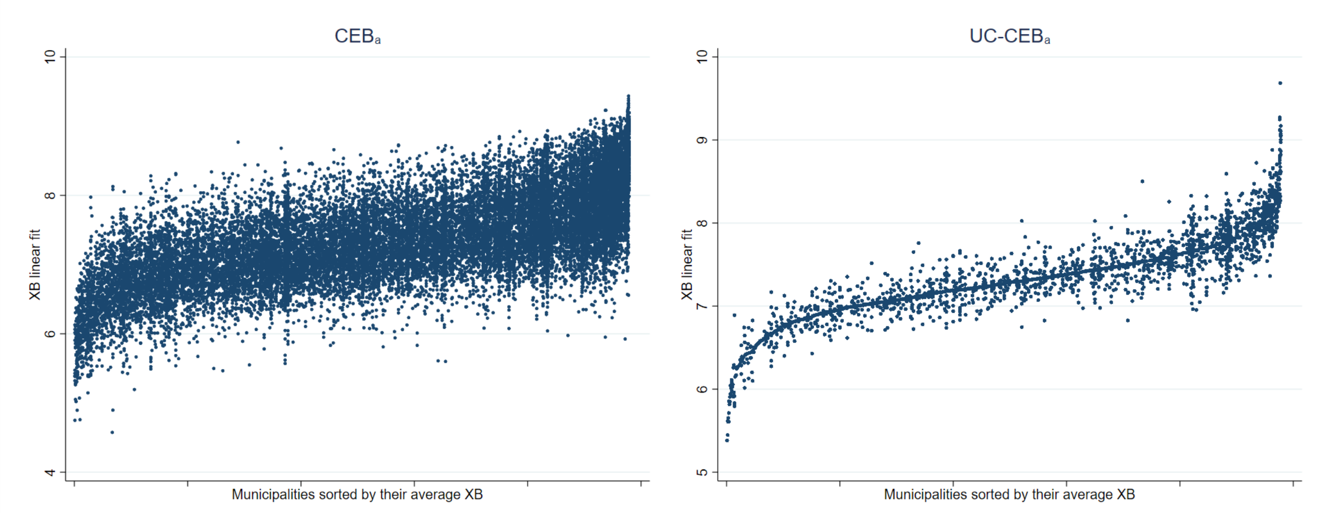


Figure 21: Box plots of design bias under ordered quantile normalization by municipality population deciles (Two-stage sampling)

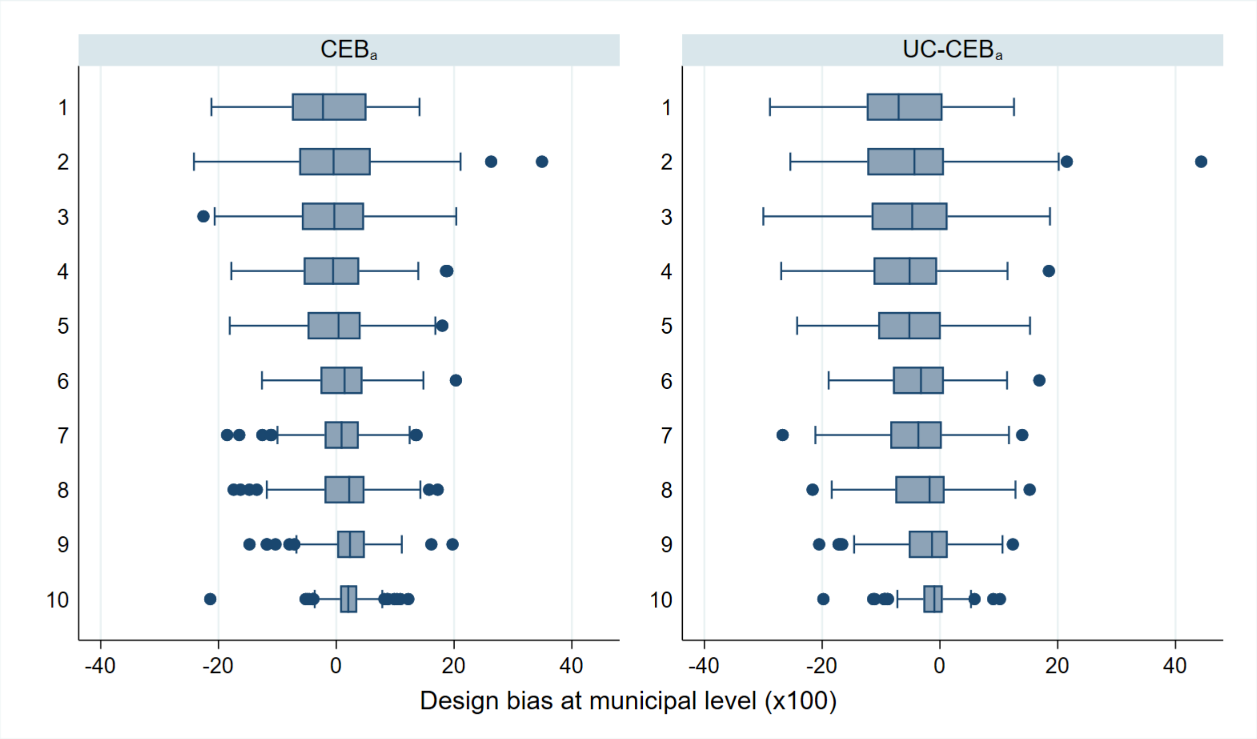
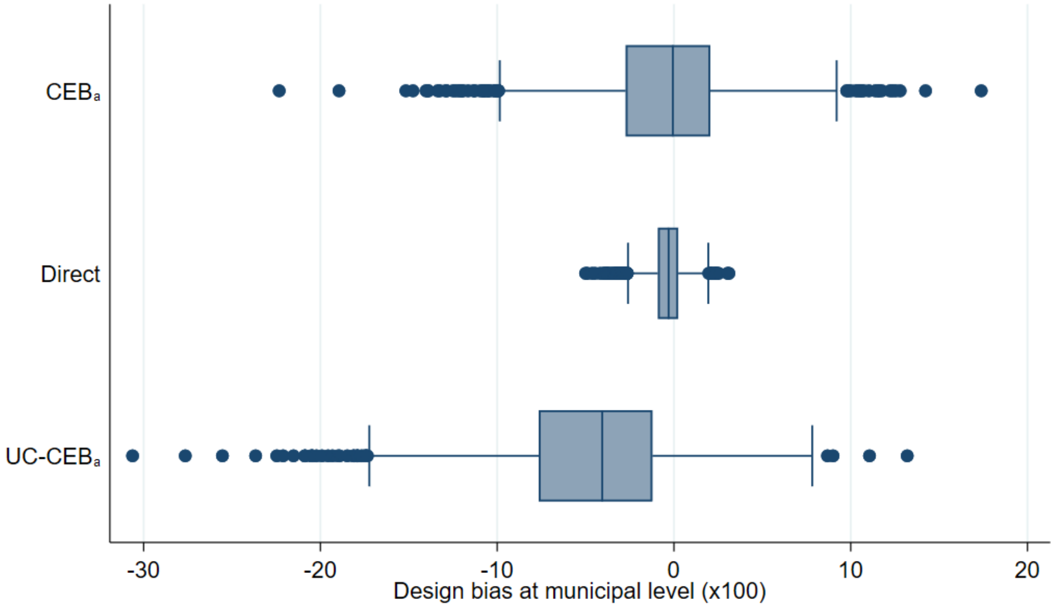
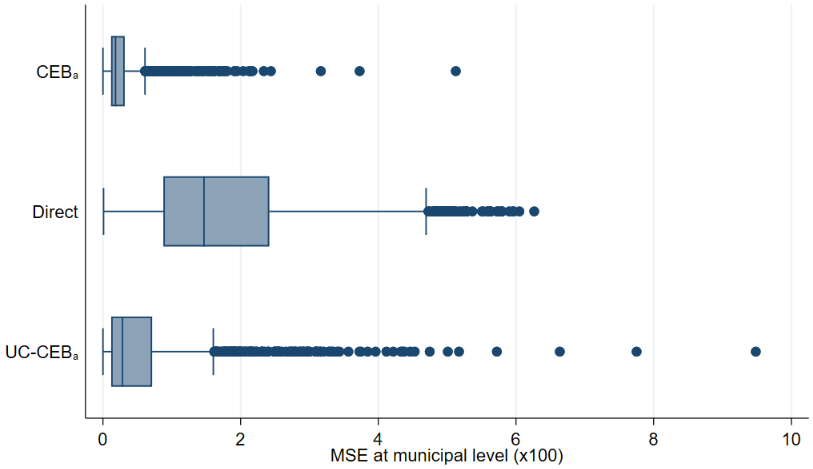


Figure 22: Box plots of design bias under 1% SRS by PSU sampling (Hybrid simulation)



in Figure 20 where the households within municipalities with just one PSU in the sample will have the same linear fit. More specifically, all the points corresponding to different households from the same PSU become superposed in Figure 20 right, and for those municipalities with just one PSU, there is a single point representing the same predicted value for all the households in that municipality.

Figure 23: Box plots of empirical MSE under 1% SRS by PSU sampling by municipality population deciles (Hybrid simulation)



To check whether the upward bias of estimators based on unit-level models is due to deviations from model assumptions, specifically deviation from normality, we apply a normalization transformation called ordered quantile normalization (Peterson and Cavanaugh, 2019).[48] In the absence of tied values the transformation is guaranteed to produce a normally distributed transformed data. The transformation is of use only for FGT0 because it cannot be fully reversed 1 to 1 without the original data and thus has to extrapolate values not observed in the original data (*ibid*).[49] The result for this transformation can be seen in Figure 21, and it shows that the deviation from normality may be an issue in our models. The previously observed upward bias in the CEB_a models is less evident in these results. However, now that the deviation from normality is less of an issue, the $UC-CEB_a$ models show a clear downward bias. The figure adds further evidence of a possible bias inherent in the $UC - CEB_a$ model offsetting the bias due to the deviation from the model's assumptions - in this case normality. Offsetting of biases is not guaranteed to always occur.

As an additional check, we also performed a hybrid experiment that consists in using the census data created in section 4.1 and a 5% SRS from each PSU to construct a synthetic census based on a twofold model. The 5% SRS sample is used to select a model with all eligible covariates (household and aggregate) following the same process described in section 4.2. Using the model's resulting parameter estimates from a twofold model as in (2) we create a new welfare vector in the

[48] The transformation is the same one utilized by Masaki et al. (2020) in their design based simulations.

[49] The original functional transformation is only defined when a given value is in the observed original data (Peterson and Cavanaugh, 2019).

Figure 24: Box plots of design bias under 1% SRS by PSU sampling, by municipality population deciles (Hybrid simulation)

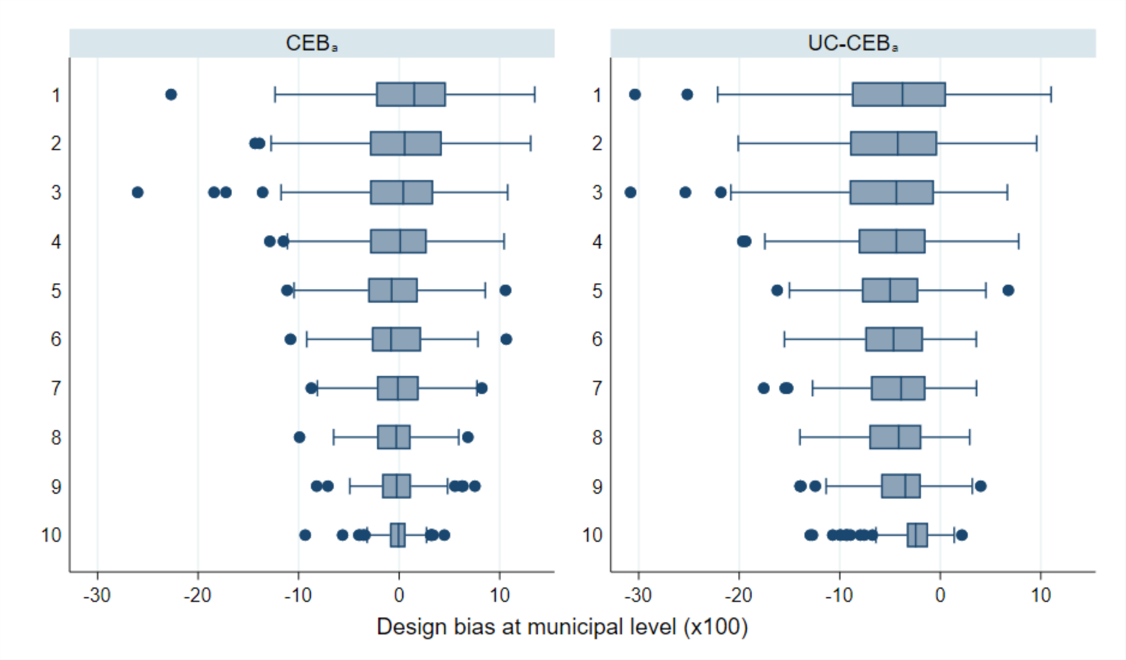
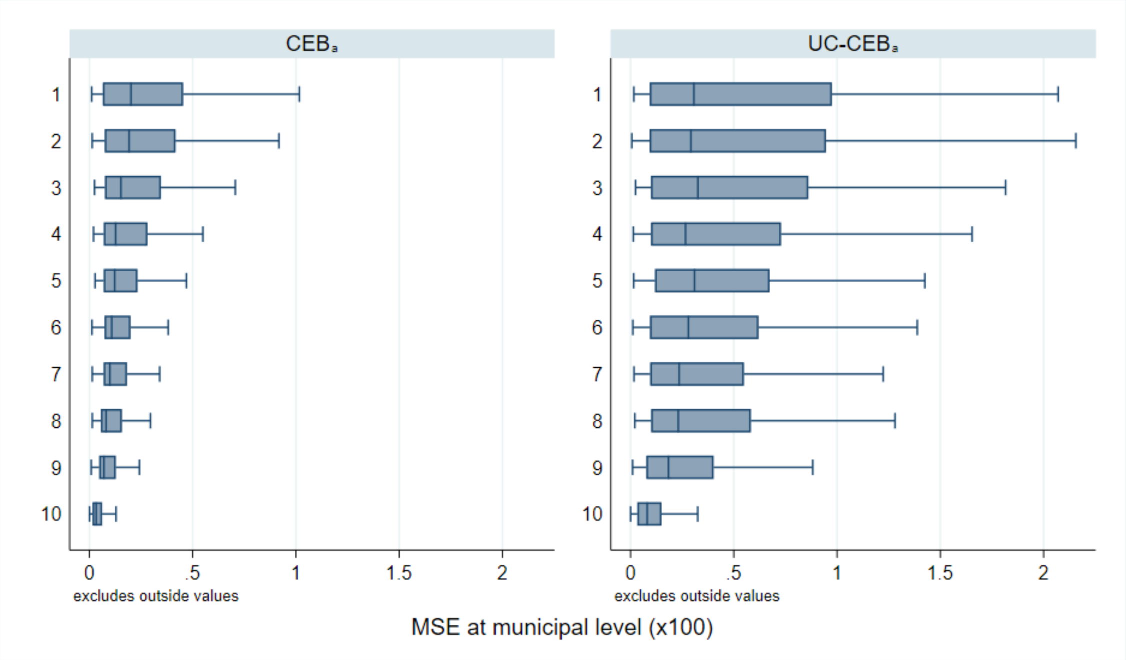


Figure 25: Box-plot of empirical MSE for FGT0 under 1% SRS by PSU sampling by municipality population deciles (Hybrid simulation)



census for all households. Then a unit-context model and a new unit model are selected, once again following the process described in section 4.2 using the first out of 500 samples where 1% SRS by PSU is selected. This is done to remove the issue of outliers from the data and to ensure that the data generating process follows the one assumed in Eq. (2). The simulation removes the potential misspecification due to deviations from normality in the data and allows us to isolate the problem present in the unit-context model ($UC - CEB_a$).

Results for the new hybrid simulations are presented in Figures 22 and 23. Note that in this simulation, where we have removed the normality issue, the upward bias that was present in the unit level model (CEB_a) is no longer evident. On the other hand, the previously suspected downward bias of the unit context models ($UC - CEB_a$) is salient, as can be seen in Figure 22 and by municipality deciles sorted by population in Figure 24. Note that under the $UC - CEB_a$ model more than 75% of the municipalities present a downward bias (Fig. 22). This finding is aligned to what we observed in Figure 17. However, because there is no deviation from normality in the hybrid simulation, the downward bias of the unit-context models ($UC - CEB_a$) is never offset, and is quite considerable and leading to substantially larger empirical MSEs for the unit context models (Fig. 25). Simulations were repeated, where, instead of performing model selection, the selected model for CEB estimators contains exactly the same covariates as those used to generate the welfare, and considering only the area aggregates for U-C models. This was done just to check whether the observed biases could be due to model mispecification, in the sense that the selected covariates are different from those in the true model. Results were very similar to those observed in the previous hybrid simulation with a model selection step. Hence, the results suggest deviations from the assumed model are an issue and the countering of biases is what is driving the seemingly good results for unit-context models in the two-stage sampling scenario, highlighting the importance of proper data transformations and model selection to ensure that model assumptions hold when using Census EB methods.

Given the direction of the bias of unit-context models is not known *a priori*[50] - and that these might present high bias - unit-context models are unlikely to be preferred over traditional FH methods when the census auxiliary data are not aligned to survey microdata, unless the calculation of variances is not possible for various locations as noted before.[51] In this case benchmarking is not a recommended procedure and may not help in reducing the bias. EB estimators are approximately model unbiased and optimal in terms of minimizing the MSE for a given area, thus when adjusted afterwards for benchmarking, that is, so that these match usual estimates at higher aggregation levels, the optimal properties are lost and estimators usually become worse in terms of bias and MSE under the model. When benchmarking adjustments are large, as those likely required by UC variants, it is an indication that the model does not really hold for the data. In the case of UC models we have shown that the model will not hold due to omitted variable bias.

Also, note bias can lead to considerable re-ranking of locations and thus a limit on the accept-

[50] Notice how under the simulations presented in Figures 1 and 3 the method appears to be upward biased.

[51] This holds for other measures of welfare, and particularly for ELL variants of the unit-context models.

able bias should usually be determined according to need. This is of particular importance when determining priorities across areas based on small area estimates. If an area's true poverty rate is of 50% and the method yields an estimator of 10% due to a biased model, there is a real risk that this area may not be given assistance when needed. Molina (2019) suggests 5 or 10 percent of absolute relative bias as acceptable thresholds. An additional problem for unit-context models in many applications is it is not possible to match census and survey PSUs; in some cases it is due to confidentiality reasons and in others it is due to different sampling frames used for the survey. The latter is something that is likely to affect applications where census and surveys correspond to different years. Under these scenarios, unit-context models are unlikely to be superior to FH and alternative area models.

5 Conclusions

In this paper we have illustrated that one of the most important aspects of SAE applications with Census EB methods under unit models is selecting a proper model; specially, the issue of data transformation. Such a finding is not new, and has been quite often echoed by others in the area (see Marhuenda et al. 2017; Molina and Marhuenda 2015; Tzavidis et al. 2018). Here we show how data transformations can lead to improved results. For example, under onefold nested error models, the aggregate gains from moving to a log-shift transformation as opposed to just taking the natural logarithm are close to 30 percent in terms of MSE. Nevertheless, as Marhuenda et al. (2017) note, finding an appropriate transformation is not always straightforward and the resulting data could stray from normality which would lead to biased estimates as we also find here. Consequently, model checks and residual analysis are something that every SAE application should include in order to test if the model's assumptions are not completely invalidated. In case of data deviating from model assumptions, the model should be changed accordingly. For instance, in case of area outliers, fixed effects could be included for those outlying areas in the model to reduce the design bias. If their sample sizes are not too small, the efficiency of the resulting model-based estimates might be acceptable in this case, even if specific model parameters are specific to these areas.

Second, we have validated SAE applications under model-based simulation and design-based simulation methods. Under model-based validation, where the data generating process follows a twofold nested error model, we note the ELL method still performs poorly in terms of MSE even with contextual level variables. The result is most evident in scenarios where area random effects are larger than cluster random effects, since contextual variables do not explain a sufficient amount of the area's variability. Issues regarding the underestimation of this noise under ELL are not evaluated here, but should be considered by future practitioners, particularly when the noise is estimated under the MI inspired bootstrap method (see Corral et al. 2020; Das and Chambers 2017; Marhuenda et al. 2017). On the other hand, model-based simulations conducted here provide further evidence to the finding from Marhuenda et al. (2017) that misspecification of the model under the onefold CensusEB, i.e. modeling random effects at the area level only, when the true model has cluster and area level random effects, with clusters nested within areas, entails virtually no loss of efficiency when estimating area level welfare-based indicators.

Under design-based validation, where the sampling strategy mimics real world scenarios such as those implemented under LSMS surveys, SAE methods present improvements over direct estimators. We have also investigated estimators based on unit-context models, originally proposed by Nguyen (2012), which can be applied when census auxiliary data at the household level is not valid. Given that under the two-stage samples used, many municipalities, which are the target areas, are represented by a small amount of observations (or even zero), these data are not always suitable for FH area level models. Despite model-based simulations yielding poor results in terms of bias for unit-context models the CensusEB variant does considerably better than the ELL variant under the design-based simulations. The reason for the positive results is shown to be due to the bias that is inherent in the unit-context model, which was being offset by bias due to deviations from normality. This offset is something that is not guaranteed to occur in all scenarios because the direction of the bias of unit-context models is not known *a priori*. In simulations where deviations from normality are not an issue, the bias of the unit-context method becomes quite clear. Additionally, the method's performance is contingent upon the number of subdomains present in the domain. Moreover, PSU or subdomains must be matched across census and survey for unit-context models, something that is not always feasible. Finally, since the direction and size of the bias of the unit-context model are not known beforehand, the models may be considered an alternative to FH and other area level models only when area level models are not applicable, like in our design-based experiment with the more realistic two-stage sampling.

References

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28–36. http://www.jstor.org/stable/2288915

Corral, P., Molina, I., & Nguyen, M. C. (2020). Pull your small area estimates up by the bootstraps. *World Bank Policy Research Working Paper*, (9256).

Das, S., & 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.

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

Elbers, C., Lanjouw, J. O. [Jean Olson], & Lanjouw, P. (2002). Micro-level estimation of welfare. *World Bank Policy Research Working Paper*, (2911).

Fay III, R. E., & 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*(366a), 269–277.

Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society*, 761–766.

Ghosh, M., & Rao, J. (1994). Small area estimation: An appraisal. *Statistical science*, *9*(1), 55–76.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, *78*(5), 443–462.

Grosh, M. E., & Muñoz, J. (1996). *A manual for planning and implementing the living standards measurement study survey*. The World Bank.

Guadarrama, M., Molina, I., & Rao, J. (2016). A comparison of small area estimation methods for poverty mapping. *Statistics in Transition new series*, *1*(17), 41–66.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, *9*(2), 226–252.

Huang, R., & Hidiroglou, M. (2003). Design consistent estimators for a mixed linear model on survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association (2003)*, 1897–1904.

Lange, S., Pape, U. J., & Pütz, P. (2018). Small area estimation of poverty under structural change. *World Bank Policy Research Working Paper*, (9383).

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 1111–1136.

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2020). Small area estimation of non-monetary poverty with geospatial data. *World Bank Policy Research Working Paper*, (9383).

Molina, I. (2019). Desagregación de datos en encuestas de hogares: Metodologías de estimación en áreas pequeñas.

Molina, I., & Marhuenda, Y. (2015). Sae: An R package for small area estimation. *The R Journal*, *7*(1), 81–98.

Molina, I., & Morales, D. (2009). Small area estimation of poverty indicators. *Estadistica e Investigacion Operativa*, *25*(3).

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

Nguyen, M. C., Corral, P., Azevedo, J. P., & Zhao, Q. (2018). Sae: A stata package for unit level small area estimation. *World Bank Policy Research Working Paper*, (8630).

Nguyen, V. C. (2012). A method to update poverty maps. *The Journal of Development Studies*, *48*(12), 1844–1863. https://doi.org/10.1080/00220388.2012.682983

Peterson, R. A., & Cavanaugh, J. E. (2019). Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*.

Rao, J., & Molina, I. (2015). *Small area estimation* (2nd). John Wiley & Sons.

Tarozzi, A., & 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.

Torabi, M., & 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, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(4), 927–979.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman; Hall/CRC.

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).

You, Y., & Rao, J. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, *30*(3), 431–439.

Zhao, Q. (2006). User manual for povmap. *World Bank. http://siteresources. worldbank. org/INTPGI/Resources/34 1092157888460/Zhao_ ManualPovMap. pdf.*

Appendix

Figure 26: Normal Q-Q-plot of onefold unit-context model and predicted subdomain effects (Nat. log transformation)

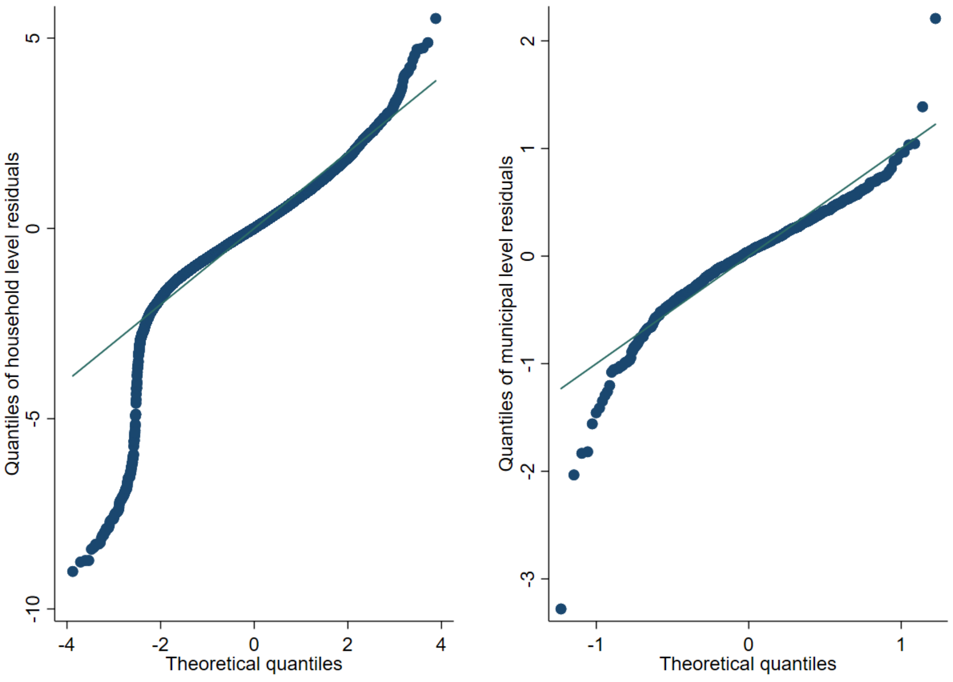


Figure 27: Normal Q-Q-plot of onefold unit-context model and predicted subdomain effects (log-shift transformation)

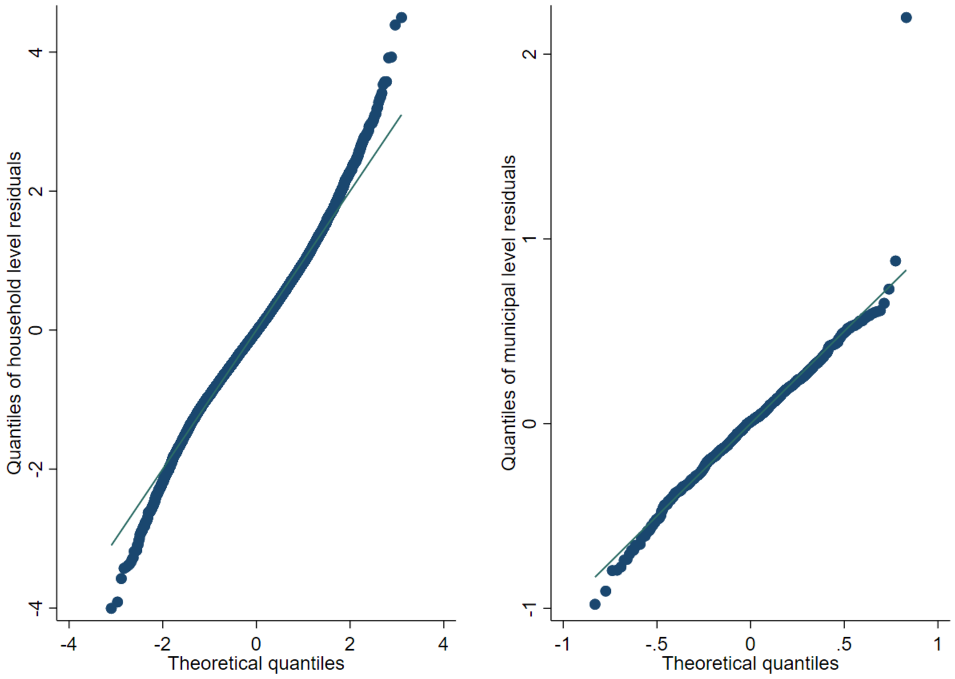


Figure 28: Normal Q-Q-plot of onefold unit-context model and predicted subdomain effects (Box-Cox of nat. log transformation)

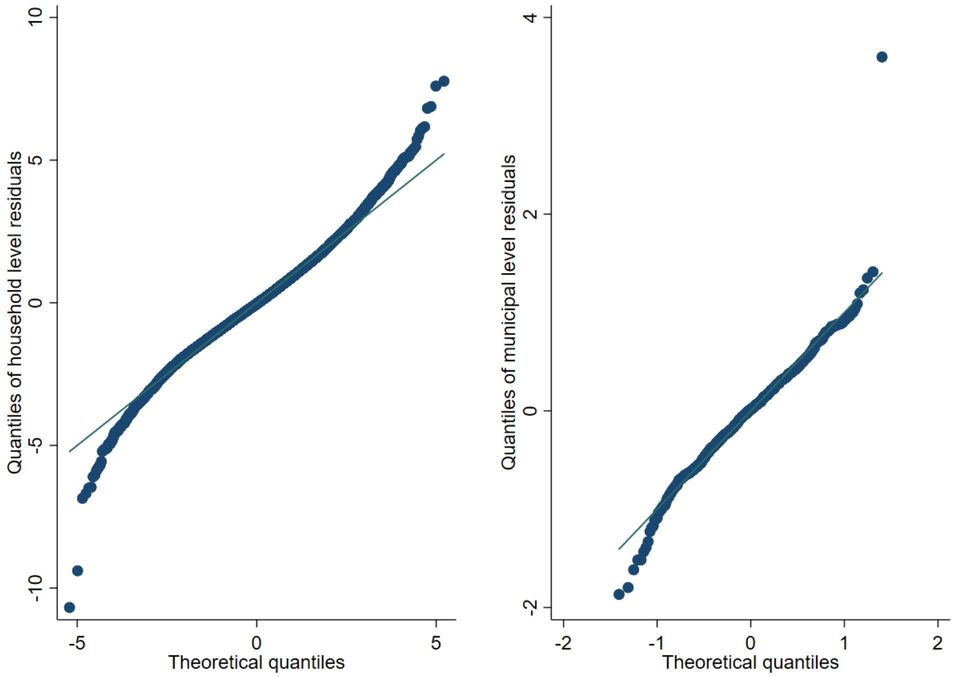


Figure 29: Box plots of empirical design MSE for FGT0 under 1% SRS by PSU sampling (Nat log shift transformation)

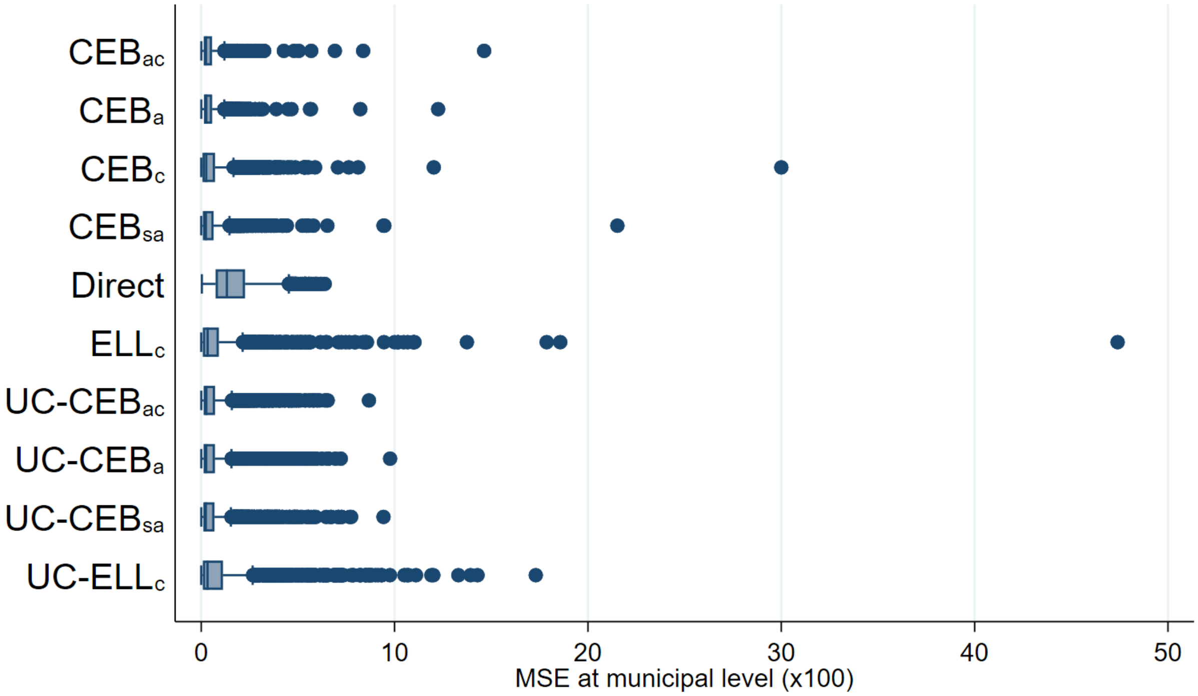


Figure 30: Box plots of empirical design bias for FGT0 under 1% SRS by PSU sampling (Nat log shift transformation)

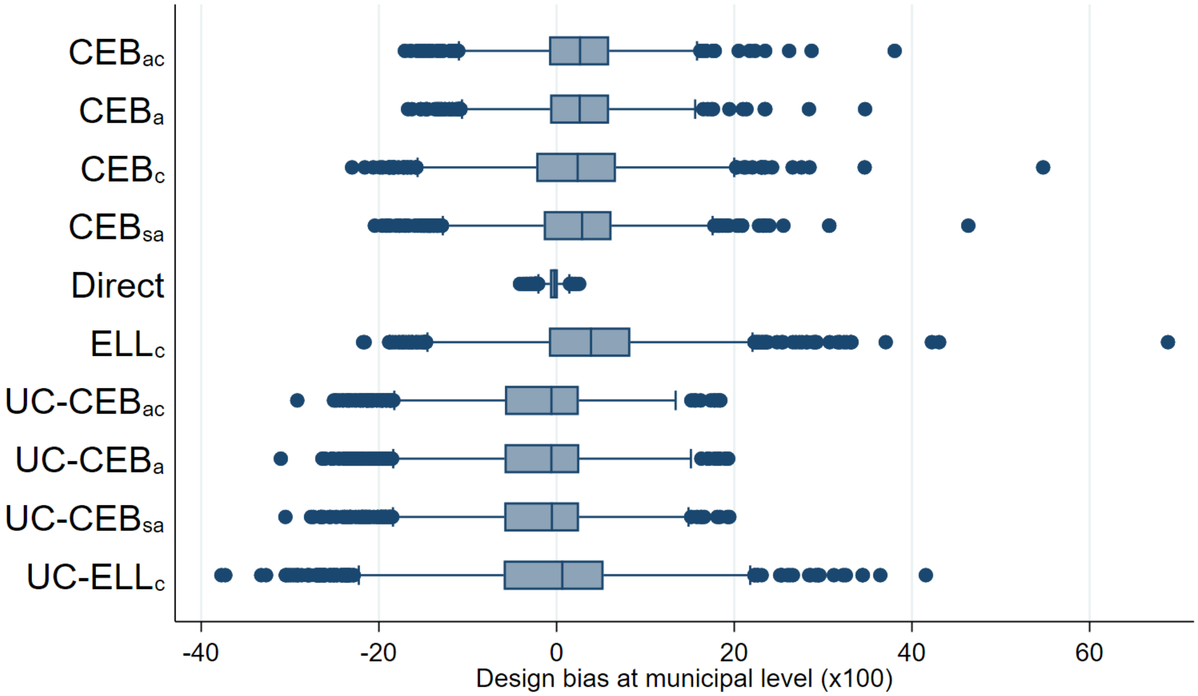


Figure 31: Box plots of empirical design MSE for FGT0 under two-stage sampling (Nat log shift transformation)

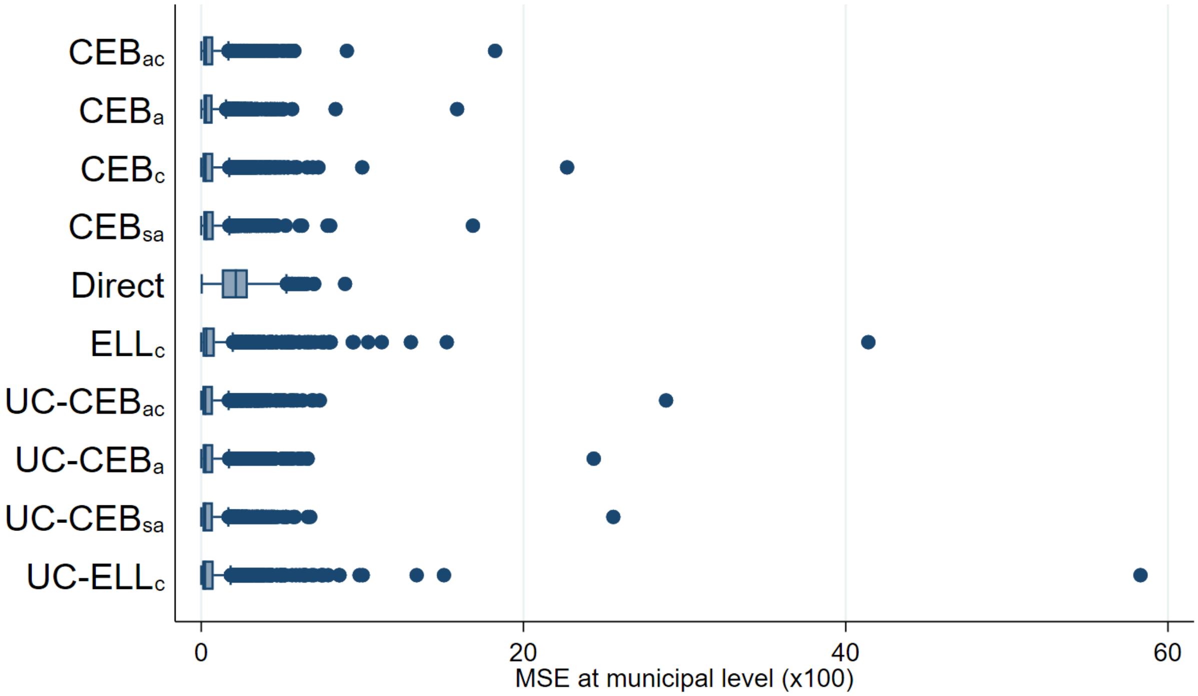


Figure 32: Average empirical MSE for FGT0 under two-stage sampling by municipality population deciles (Nat log shift transformation)

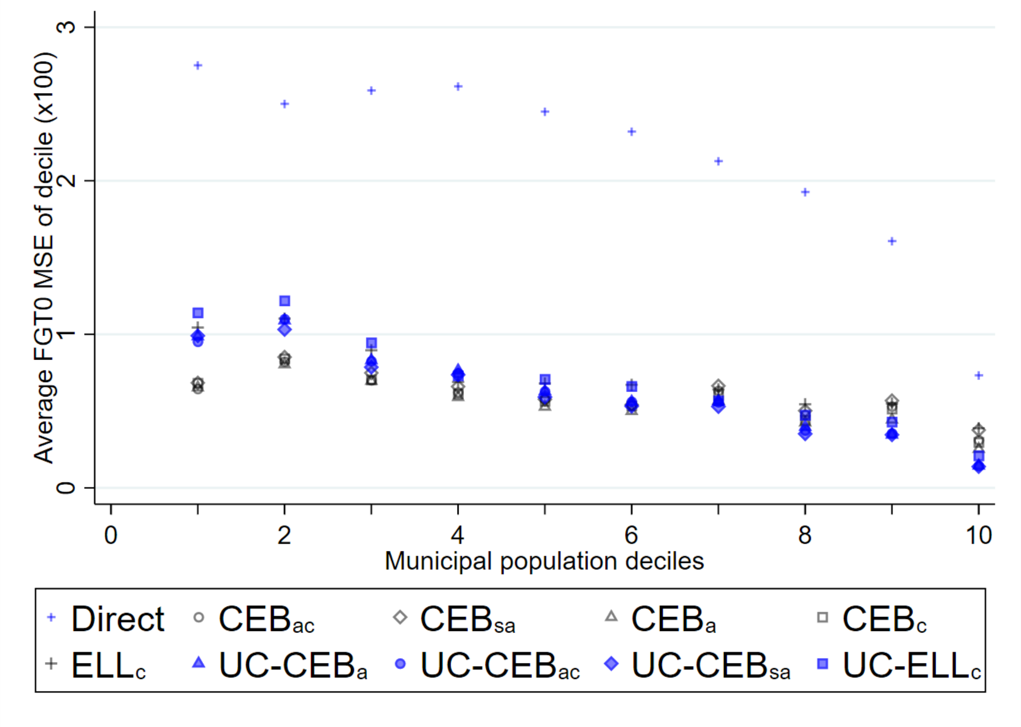


Table 3: Unit level model - under log shift transformation

Variable	Coefficient	Std. Error
PSU HH head age	-0.0726***	(0.00700)
PSU % hh with internet	0.133***	(0.00937)
PSU % of hh with television	0.141***	(0.00864)
Head's age	-0.00127***	(0.000406)
HH owns a cellphone	0.214***	(0.0155)
HH owns a computer	0.190***	(0.0178)
Number of hh members	-0.118***	(0.00315)
HH has access to internet	0.105***	(0.0181)
Male HH head	0.0640***	(0.0136)
Max. education is tertiary	0.393***	(0.0156)
Share of adult hh members	0.688***	(0.0247)
State % of male head hh	-0.0683***	(0.00870)
State share of elderly population	-0.0514***	(0.0120)
HH owns a washing machine	0.119***	(0.0136)
Constant	6.984***	(0.0351)
Observations	23,516	
Adj. R^2	0.445	
σ_a^2	0.0220	
σ_a^2/σ_e^2	0.0452	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: Unit-context model - under log shift transformation

Variable	Coefficient	Std. Error
PSU avg. hh number of members	-0.127***	(0.0105)
PSU share of male headed households	-0.0595***	(0.00976)
PSU share of households with max. tertiary education	0.219***	(0.0105)
PSU avg. share of elderly	-0.0760***	(0.00783)
PSU avg. share of female members	-0.0283***	(0.00913)
Municipal share of hh owning a television	0.168***	(0.0116)
State share of hh owning a computer	0.123***	(0.0110)
Constant	7.353***	(0.0104)
Observations	23,516	
Adj. R^2	0.253	
σ_a^2	0.0234	
σ_a^2/σ_e^2	0.0357	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5: Aggregate results for 1,865 municipalities of Mexico (FGT1) - Results from 500 samples

	1	2	3	4
Transformation	Box-Cox	Nat. log.	Log. Shift	Log. Shift
Direct				
AAB (×100)	5.549	5.549	5.549	4.773
ARMSE (×100)	6.975	6.975	6.975	5.943
$\boldsymbol{CEB_{ac}}$				
AAB (×100)	3.583	5.843	3.699	3.210
ARMSE (×100)	3.747	6.217	3.867	3.523
$\boldsymbol{CEB_{sa}}$				
AAB (×100)	3.665	6.068	3.773	3.417
ARMSE (×100)	3.830	6.386	3.941	3.553
$\boldsymbol{CEB_{a}}$				
AAB (×100)	3.427	5.889	3.533	3.205
ARMSE (×100)	3.693	6.406	3.807	3.553
$\boldsymbol{CEB_{c}}$				
AAB (×100)	3.548	5.779	3.680	3.556
ARMSE (×100)	3.614	5.904	3.746	3.631
$\boldsymbol{ELL_{c}}$				
AAB (×100)	4.209	5.783	4.209	4.516
ARMSE (×100)	4.300	5.844	4.300	4.600
$\boldsymbol{UC-CEB_{a}}$				
AAB (×100)	3.329	4.589	3.166	3.319
ARMSE (×100)	3.543	5.080	3.400	3.668
$\boldsymbol{UC-CEB_{ac}}$				
AAB (×100)	3.287	4.606	3.165	3.284
ARMSE (×100)	3.451	4.998	3.344	3.653
$\boldsymbol{UC-ELL_{c}}$				
AAB (×100)	3.463	4.730	3.463	4.522
ARMSE (×100)	3.574	4.797	3.574	4.637
$\boldsymbol{UC-CEB_{sa}}$				
AAB (×100)	3.265	4.628	3.136	3.319
ARMSE (×100)	3.470	5.122	3.357	3.638

AAB: Average absolute bias; ARMSE: Average root MSE

ELL under Box-Cox and log shift, model is still for Nat. log. but errors are drawn from the empirical distribution.

Columns 1, 2, and 3 under two-stage samples (Section 4.1)

Column 4 results are under 1% SRS by PSU, all PSUs are included

Table 6: Aggregate results for 1,865 municipalities of Mexico (FGT2) - Results from 500 samples

	1	2	3	4
Transformation	Box-Cox	Nat. log.	Log. Shift	Log. Shift
Direct				
AAB (×100)	3.901	3.901	3.901	3.441
ARMSE (×100)	5.017	5.017	5.017	4.361
$\boldsymbol{CEB_{ac}}$				
AAB (×100)	2.432	4.135	2.566	2.267
ARMSE (×100)	2.539	4.421	2.682	2.476
$\boldsymbol{CEB_{sa}}$				
AAB (×100)	2.498	4.289	2.617	2.411
ARMSE (×100)	2.605	4.535	2.731	2.501
$\boldsymbol{CEB_{a}}$				
AAB (×100)	2.327	4.144	2.434	2.259
ARMSE (×100)	2.498	4.540	2.619	2.492
$\boldsymbol{CEB_{c}}$				
AAB (×100)	2.415	4.073	2.555	2.514
ARMSE (×100)	2.462	4.173	2.602	2.564
$\boldsymbol{ELL_{c}}$				
AAB (×100)	3.053	4.085	3.053	3.364
ARMSE (×100)	3.124	4.137	3.124	3.435
$\boldsymbol{UC-CEB_{a}}$				
AAB (×100)	2.354	3.200	2.185	2.329
ARMSE (×100)	2.482	3.558	2.335	2.547
$\boldsymbol{UC-CEB_{ac}}$				
AAB (×100)	2.306	3.204	2.182	2.303
ARMSE (×100)	2.406	3.492	2.299	2.535
$\boldsymbol{UC-ELL_{c}}$				
AAB (×100)	2.478	3.297	2.478	3.312
ARMSE (×100)	2.560	3.352	2.560	3.405
$\boldsymbol{UC-CEB_{sa}}$				
AAB (×100)	2.307	3.228	2.177	2.334
ARMSE (×100)	2.431	3.591	2.319	2.532

AAB: Average absolute bias; ARMSE: Average root MSE

ELL under Box-Cox and log shift, model is still for Nat. log. but errors are drawn from the empirical distribution.

Columns 1, 2, and 3 under two-stage samples (Section 4.1)

Column 4 results are under 1% SRS by PSU, all PSUs are included

Table 7: Aggregate results for 1,865 municipalities of Mexico (Mean income) - Results from 500 samples

	1	2	3	4
Transformation	Box-Cox	Nat. log.	Log. Shift	Log. Shift
Direct				
AAB (×100)	38,838.44	38,838.44	38,838.44	30,028.16
ARMSE (×100)	54,351.71	54,351.71	54,351.71	41,841.21
$\boldsymbol{CEB_{ac}}$				
AAB (×100)	18,374.61	45,904.53	18,276.26	16,908.87
ARMSE (×100)	19,998.56	48,531.76	19,899.96	19,569.77
$\boldsymbol{CEB_{sa}}$				
AAB (×100)	19,574.38	45,440.07	19,452.75	18,426.15
ARMSE (×100)	21,207.41	47,861.59	21,069.74	19,745.09
$\boldsymbol{CEB_{a}}$				
AAB (×100)	18,146.66	43,445.38	18,158.09	17,016.17
ARMSE (×100)	20,596.54	47,108.70	20,591.85	19,967.33
$\boldsymbol{CEB_{c}}$				
AAB (×100)	19,048.90	46,845.91	19,093.24	21,112.72
ARMSE (×100)	19,534.06	47,576.32	19,570.16	21,653.29
$\boldsymbol{ELL_{c}}$				
AAB (×100)	32,576.30	48,594.36	32,576.30	44,808.60
ARMSE (×100)	33,650.59	49,018.62	33,650.59	45,504.33
$\boldsymbol{UC - CEB_{a}}$				
AAB (×100)	31,330.70	67,999.59	30,715.13	29,647.72
ARMSE (×100)	33,900.14	72,555.97	33,309.27	33,513.86
$\boldsymbol{UC - CEB_{ac}}$				
AAB (×100)	32,718.31	69,364.83	31,893.15	29,700.46
ARMSE (×100)	34,599.54	72,841.40	33,823.68	33,765.53
$\boldsymbol{UC - ELL_{c}}$				
AAB (×100)	53,184.14	73,673.70	53,184.14	71,189.32
ARMSE (×100)	54,672.16	74,068.42	54,672.16	72,282.49
$\boldsymbol{UC - CEB_{sa}}$				
AAB (×100)	33,148.54	69,355.82	32,296.15	29,642.77
ARMSE (×100)	35,502.18	73,879.73	34,677.96	33,217.28

AAB: Average absolute bias; ARMSE: Average root MSE

ELL under Box-Cox and log shift, model is still for Nat. log. but errors are drawn from the empirical distribution.

Columns 1, 2, and 3 under two-stage samples (Section 4.1)

Column 4 results are under 1% SRS by PSU, all PSUs are included

Omitted variable bias for unit-context models

Consider that data are generated as:

$$y_{ah} = \beta_0 + \beta_1 x_{ah}^1 + \cdots + \beta_p x_{ah}^p + \eta_a + e_{ah};\ h = 1, \ldots, N_a;\ a = 1, \ldots, A \tag{4}$$

Let us decompose x_{ah}^k as follows:

$x_{ah}^k = \left(x_{ah}^k - \bar{X}_a^k\right) + \bar{X}_a^k$, where $\bar{X}_a^k = N_a^{-1} \sum_{h=1}^{N_a} x_{ah}^k$ is the population mean of x_{ah}^k in area a. However under unit-context model we fit:

$$y_{ah} = \alpha_0 + \alpha_1 \bar{X}_a^1 + \cdots + \alpha_p \bar{X}_a^p + \eta_a + e_{ah};\ h = 1, \ldots, N_a;\ a = 1, \ldots, A \tag{5}$$

Note that here we are omitting variables:

$$\tilde{x}_{ah}^k = x_{ah}^k - \bar{X}_a^k;\ k = 1, \ldots, p$$

Let us write model (4) in matrix notation, for the sample data. For this we define the vectors:

$$\mathbf{y}_a = \begin{pmatrix} y_{a1} \\ \vdots \\ y_{an_a} \end{pmatrix},\ \mathbf{X}_a = \begin{pmatrix} 1 & x_{a1}^1 & \cdots & x_{a1}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{an_a}^1 & \ldots & x_{an_a}^p \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\ \mathbf{e}_a = \begin{pmatrix} e_{a1} \\ \vdots \\ e_{an_a} \end{pmatrix}$$

Then, the model is given by:

$$\mathbf{y}_a = \mathbf{X}_a \beta + \mathbf{1}_{n_a} \eta_a + \mathbf{e}_a,\ a = 1, \ldots, A$$

Finally, define:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_A \end{pmatrix},\ \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_A \end{pmatrix},\ Z = \begin{pmatrix} \mathbf{1}_{n_1} & & \\ & \ddots & \\ & & \mathbf{1}_{n_A} \end{pmatrix},\ \eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_A \end{pmatrix}\ \mathbf{e}_a = \begin{pmatrix} e_{a1} \\ \vdots \\ e_{an_a} \end{pmatrix}$$

where $\mathbf{1}_{n_A}$ is an $n_A \times 1$ column of ones.

Then the model in (4) may be written as:

$$\mathbf{y} = \mathbf{X}\beta + Z\eta + \mathbf{e}$$

On the other hand, for model (5) we define:

$$\bar{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \bar{X}_1^1 \mathbf{1}_{n_1} & \cdots & \bar{X}_1^p \mathbf{1}_{n_1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_A} & \bar{X}_A^1 \mathbf{1}_{n_A} & \cdots & \bar{X}_1^p \mathbf{1}_{n_A} \end{pmatrix}$$

Then model (5) can be written as, for $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)'$:

$$\mathbf{y} = \bar{X}\alpha + Z\eta + \mathbf{e}$$

the WLS estimator of α in model (5):

$$\hat{\alpha}^{UC} = \left(\bar{X}'V^{-1}\bar{X}\right)^{-1} \bar{X}'V^{-1}\mathbf{y} \tag{6}$$

However, $\mathbf{y}$ actually follows model (4), that is:

$$\mathbf{y} = \tilde{X}\beta_{(0)} + \bar{X}\beta + Z\eta + \mathbf{e} \tag{7}$$

where

$$\beta_{(0)} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \ \tilde{X} = \begin{pmatrix} \tilde{X}_1 \\ \vdots \\ \tilde{X}_A \end{pmatrix}, \ \tilde{X}_a = \begin{pmatrix} x_{a1}^1 - \bar{X}_a^1 & \cdots & x_{a1}^p - \bar{X}_a^p \\ \vdots & \ddots & \vdots \\ x_{an_a}^1 - \bar{X}_a^1 & \cdots & x_{an_a}^p - \bar{X}_a^p \end{pmatrix}$$

replacing (7) in to (6), we get:

$$\hat{\alpha}^{UC} = \left(\bar{X}'V^{-1}\bar{X}\right)^{-1} \bar{X}'V^{-1}\bar{X}\beta + \left(\bar{X}'V^{-1}\bar{X}\right)^{-1} \bar{X}'V^{-1}\tilde{X}\beta_{(0)} + \left(\bar{X}'V^{-1}\bar{X}\right)^{-1} \bar{X}'V^{-1}(Z\eta + \mathbf{e})$$

Taking expectations, and since $E\left[\eta\right] = 0$ and $E\left[\mathbf{e}\right] = 0$, we get:

$$E\left[\hat{\alpha}^{UC}\right] = \beta + B\left[\hat{\alpha}^{UC}\right]$$

where the bias is equal to:

$$B\left[\hat{\alpha}^{UC}\right] = \left(\bar{X}'V^{-1}\bar{X}\right)^{-1} \bar{X}'V^{-1}\tilde{X}\beta_{(0)}$$

where

$$\bar{X}'V^{-1}\bar{X} = \frac{1}{\sigma_\eta^2}\begin{pmatrix} \sum_a \gamma_a & \sum_a \gamma_a \bar{X}_a^1 & \cdots & \sum_a \gamma_a \bar{X}_a^p \\ \sum_a \gamma_a \bar{X}_a^1 & \sum_a \gamma_a \left(\bar{X}_a^1\right)^2 & \cdots & \sum_a \gamma_a \bar{X}_a^1 \bar{X}_a^p \\ \vdots & \vdots & \ddots & \vdots \\ \sum_a \gamma_a \bar{X}_a^p & \sum_a \gamma_a \bar{X}_a^1 \bar{X}_a^p & \cdots & \sum_a \gamma_a \left(\bar{X}_a^p\right)^2 \end{pmatrix}$$

and where

$$\gamma_a = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \frac{\sigma_e^2}{n_a}}$$

Additionally, we obtain

$$\bar{X}'V^{-1}\tilde{X}\beta_{(0)} = \begin{pmatrix} \sum_a \mathbf{1}_{n_a}^{'} V_a^{-1}\tilde{X}_a\beta_{(0)} \\ \sum_a \bar{X}_a^1 \mathbf{1}_{n_a}^{'} V_a^{-1}\tilde{X}_a\beta_{(0)} \\ \vdots \\ \sum_a \bar{X}_a^p \mathbf{1}_{n_a}^{'} V_a^{-1}\tilde{X}_a\beta_{(0)} \end{pmatrix}$$

noting that:

$$\mathbf{1}_{n_a}^{'} V_a^{-1}\tilde{X}_a = \frac{1}{\sigma_\eta^2}\frac{\gamma_a}{n_a}\mathbf{1}_{n_a}^{'}\tilde{x}_a$$

where $\tilde{x}_a = \left(\tilde{x}_a^1, \ldots, \tilde{x}_a^p\right)'$, with

$$\tilde{x}_a^k = \frac{1}{n_a}\sum_{h \in S_a}\left(x_{ah}^k - \bar{X}_a^k\right) = \bar{x}_a^k - \bar{X}_a^k$$

and where S_a is the survey sample households in area a. Therefore,

$$\bar{X}'V^{-1}\tilde{X}\beta_{(0)} = \frac{1}{\sigma_\eta^2}\begin{pmatrix} \sum_a \gamma_a \sum_k \tilde{x}_a^k \beta_k \\ \sum_a \gamma_a \bar{X}_a^1 \sum_k \tilde{x}_a^k \beta_k \\ \vdots \\ \sum_a \gamma_a \bar{X}_a^p \sum_k \tilde{x}_a^k \beta_k \end{pmatrix}$$

Finally, the bias is given by:

$$B\left[\hat{\alpha}^{UC}\right] = \begin{pmatrix} \sum_a \gamma_a & \sum_a \gamma_a \bar{X}_a^1 & \cdots & \sum_a \gamma_a \bar{X}_a^p \\ \sum_a \gamma_a \bar{X}_a^1 & \sum_a \gamma_a \left(\bar{X}_a^1\right)^2 & \cdots & \sum_a \gamma_a \bar{X}_a^1 \bar{X}_a^p \\ \vdots & \vdots & \ddots & \vdots \\ \sum_a \gamma_a \bar{X}_a^p & \sum_a \gamma_a \bar{X}_a^1 \bar{X}_a^p & \cdots & \sum_a \gamma_a \left(\bar{X}_a^p\right)^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_a \gamma_a \sum_k \tilde{x}_a^k \beta_k \\ \sum_a \gamma_a \bar{X}_a^1 \sum_k \tilde{x}_a^k \beta_k \\ \vdots \\ \sum_a \gamma_a \bar{X}_a^p \sum_k \tilde{x}_a^k \beta_k \end{pmatrix}$$

Consequently, the bias of $\hat{\alpha}^{UC}$ is due to the discrepancy between the sample mean of a given covariate and the population mean of that covariate, $\tilde{x}_a^k = \bar{x}_a^k - \bar{X}_a^k$.

Bias of individual predictors under unit-context models

Under model (5),

$$y_{ah} = \bar{X}_a'\alpha + \eta_a + e_{ah}$$

the conditional expectation under model (5) for $h \in S_a$ is:

$$\mu_{ah|s}^{UC} = E\left[y_{ah}|y_s\right] = \bar{X}_a'\alpha + \tilde{\eta}_a, \text{ for } \tilde{\eta}_a = \gamma_a\left(\bar{y}_a - \bar{X}_a'\alpha\right)$$

Then, and replacing α with its estimate, $\hat{\alpha}^{UC}$

$$\mu_{ah|s}^{UC} = \bar{X}_a'\hat{\alpha}^{UC} + \gamma_a\left(\bar{y}_a - \bar{X}_a\hat{\alpha}^{UC}\right) = \gamma_a\bar{y}_a + (1-\gamma_a)\,\bar{X}_a'\hat{\alpha}^{UC}$$

and taking its expectation we get:

$$E\left[\mu_{ah|s}^{UC}\right] = \gamma_a E\left[\bar{y}_a\right] + (1-\gamma_a)\,\bar{X}_a' E\left[\hat{\alpha}^{UC}\right],$$

where $E\left[\bar{y}_a\right] = \bar{x}_a'\beta$ and $E\left[\hat{\alpha}^{UC}\right] = \beta + B\left[\hat{\alpha}^{UC}\right]$. Consequently,

$$E\left[\mu_{ah|s}^{UC}\right] = \bar{X}_a'\beta + \gamma_a\left(\bar{x}_a - \bar{X}_a\right)\beta + (1-\gamma_a)\bar{X}_a'B\left[\hat{\alpha}^{UC}\right]$$

Note that once again, the discrepancy between the sample and population means play a role in the bias.