Mapping poverty

Small Area Estimation at the World Bank

Poverty and Equity Global Practice

July 1st, 2021 – ECLAC webinar on SAE techniques

Paul Corral

pcorralrodas@worldbank.org

WORLD BANK GROUP

Based on work done with:
Isabel Molina, Kristen Himelein, Kevin McGee, and Minh Nguyen

Disclaimer: The findings, interpretations, and conclusions expressed in this presentation do not necessarily reflect the views of the World Bank, the Executive Directors of the World Bank or the governments they represent.

The presentation is based on work from the following papers:

Corral, P., Molina, I., & Nguyen, M. (2021). Pull your small area estimates up by the bootstraps. Journal of Statistical Computation and Simulation, 1-54.

- Revised the bootstrap methods
- Shows how this is an improvement over the previous methods using model-based validation

Corral, P., Himelein, K., Mcgee, K., & Molina, I. (2021). A map of the poor or a poor map? World Bank Policy Research Working Paper, (9620).

- Validation done using Mexican Intra censal survey from 2015, and extracting 500 samples mimicking LSMS surveys
- Shows performance of onefold, twofold nested error models, ELL, and Unit-context models with real data

Outline

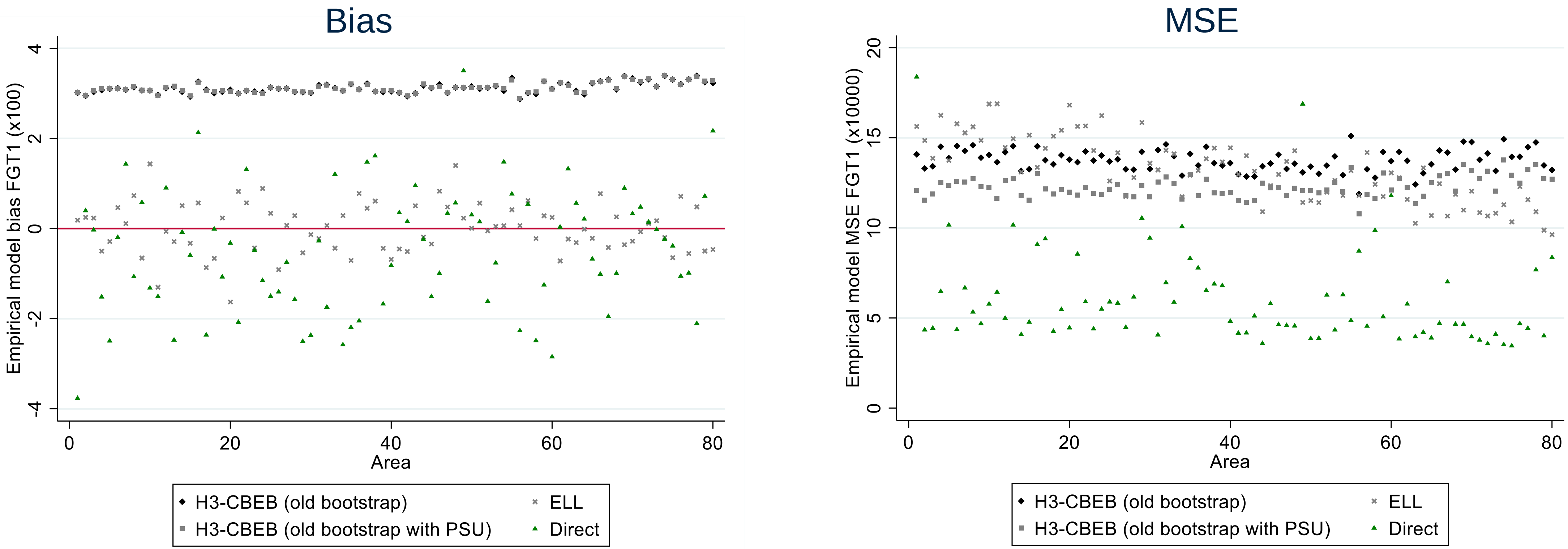
1. Updating the bootstrap and EB methods used by the World Bank
   - Original toolkit and methods
   - The need for an update
   - The update
   - Results
2. Conduct a rigorous validation exercise of broader range of methods using Mexican Intra-censal Survey of 2015
   - Importance of data transformation
   - Methods tested – Census EB (one fold and two-fold), H3-CBEB, ELL

The latest Stata sae package can be obtained from: https://github.com/pcorralrodas/SAE-Stata-Package
We'll be soon submitting to SSC for an update.

SAE at the World Bank until recently was based on the methods from ELL (2003)

- Until 2018 the World Bank relied on PovMap stand alone software
  - Coded in C
  - Very efficient and fast
  - Point and click interface made it hard to work with
  - Difficult for other people to contribute
- In 2016 we began work on a Stata, statistical software widely used at the World Bank, version of PovMap
  - Replication of PovMap version can be read about in:
    - Nguyen, M., 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).
  - This allowed us to conduct more rigorous research with the methods we were using

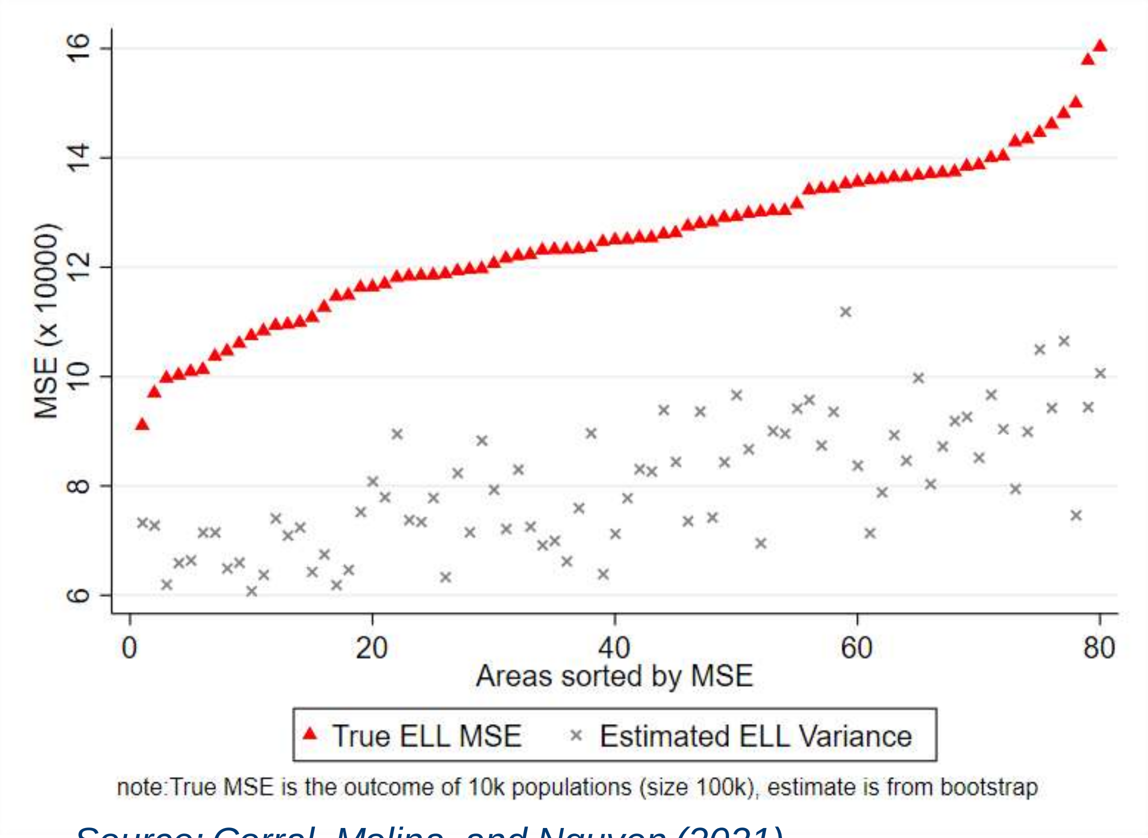
We replicated simulations from Molina and Rao (2010), and it became apparent that the methods we were using were not "Best"



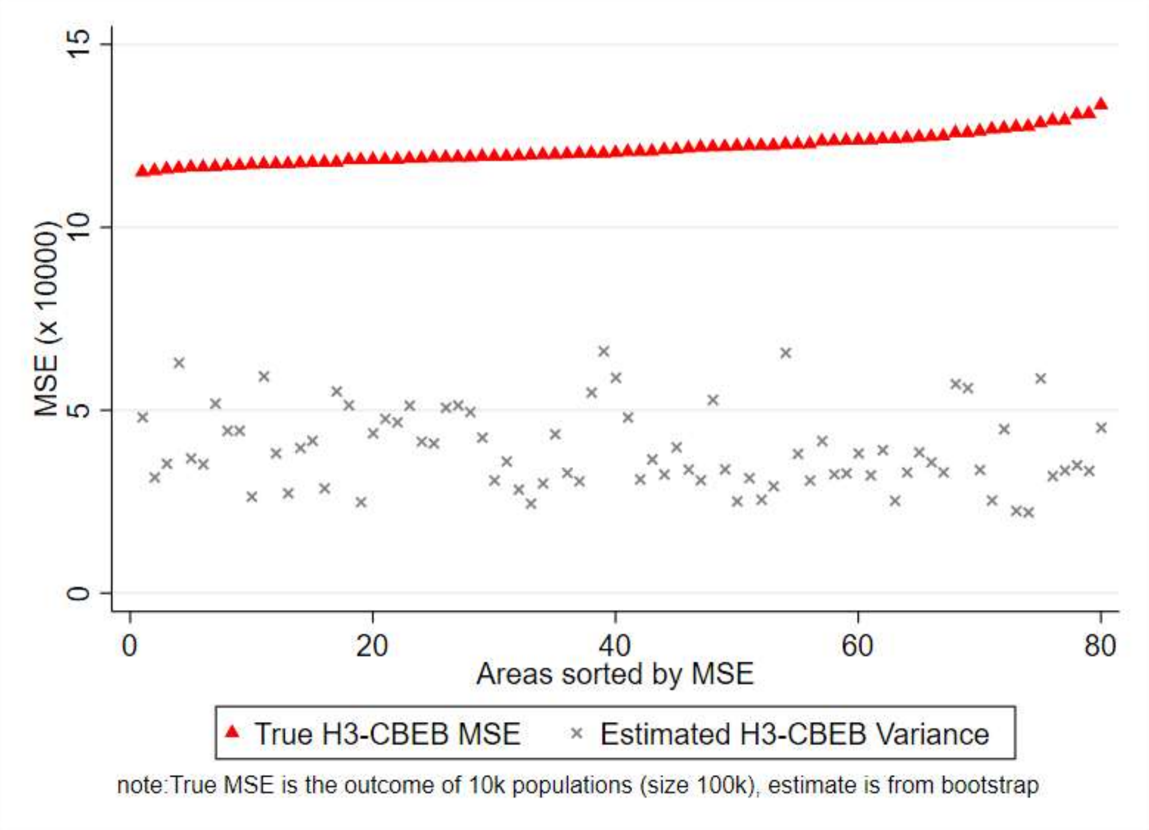
Source: Corral, Molina, and Nguyen (2021)

- Notice how the EB implementation in PovMap (H3-CBEB) markers fall consistently above 0 (left)
- ELL and direct estimates are scattered around 0 (left)
- MSE for ELL and H3-CBEB is higher than that of direct estimates (granted, this a large sample fraction, but this should not be the case)

Additionally, it allowed us to test the variance measure used in the past (PovMap & original Stata sae)



note:True MSE is the outcome of 10k populations (size 100k), estimate is from bootstrap



note:True MSE is the outcome of 10k populations (size 100k), estimate is from bootstrap

Source: Corral, Molina, and Nguyen (2021)

- The methods produce higher noise estimates, but at the same time underestimate the true noise

Assumed model is the same in ELL (2003) and Molina and Rao (2010)

The nested error model used for unit level small area estimation comes from Battese, Harter and Fuller (1988):

$$y_{ch} = x_{ch}\beta + \eta_c + e_{ch}; \quad h = 1, \dots, N_c; c = 1, \dots, C$$

where $\eta_c \sim N(0, \sigma_\eta^2)$ and $e_{ch} \sim N(0, \sigma_e^2)$

- C is the number of locations, N_c is the number of observations in location c
- The model was originally used to produce county-level corn and soybean crop area estimates for Iowa, U.S
- The model assumes normally distributed error terms

The problem is the bootstrap methods used

- The PovMap implemented ELL is based on multiple imputation (MI) literature
- The survey data is used to obtain $\hat{\theta}_0 = (\hat{\beta}_0, \hat{\sigma}^2_{\eta\,0}, \hat{\sigma}^2_{e\,0})$
- To fill in the missing vectors of welfare we simulate y^* in the Census using the model's parameters
  - In the census we have x_{ch}, but we are missing everything else

$$y^* = X_{census}\beta^* + \eta^* + e^*$$

- The original ELL obtains the necessary parameters from their estimated approximate distributions
  - $\sigma_e^{2*} \sim \widehat{\sigma_e^2}_0 (N_0 - k)/\chi^2_{N_0-k}$ $\beta^* \sim N(\hat{\beta}_0, \text{vcov}(\hat{\beta}_0))$
    - And we draw the residuals : $e^* \sim N(0, \sigma_e^{2*})$
  - $\sigma_\eta^{2*} \sim Gamma(\widehat{\sigma_\eta^2}_0, \text{var}(\widehat{\sigma_\eta^2}_0))$
    - And we draw the residuals : $\eta^* \sim N(0, \sigma_\eta^{2*})$
- **Note how $\hat{\theta}_0 = (\hat{\beta}_0, \hat{\sigma}^2_{\eta\,0}, \hat{\sigma}^2_{e\,0})$ are not used for the census simulated vectors y^***

For the EB implementation in PovMap something similar was attempted

However, van der Weide (2014) does not offer under the new method an alternative to ELL's $\sigma_\eta^{2*} \sim Gamma(\hat{\sigma}^2_{\eta\,0}, \mathrm{var}(\hat{\sigma}^2_{\eta\,0}))$

- Thus, bootstrap samples of clusters of the data are taken to maintain a similar algorithm we call it a clustered-bootstrap EB (**CBEB**)
- Under each bootstrap sample we obtain $\theta^* = (\beta^*, \sigma_\eta^{2*}, \sigma_e^{2*})$

- To fill in the missing vectors of welfare we simulate y^* in the Census using the model's parameters

$$y^* = X_{census}\beta^* + \eta^* + e^*$$

- **Note how $\hat{\theta}_0 = (\hat{\beta}_0, \hat{\sigma}^2_{\eta\,0}, \hat{\sigma}^2_{e\,0})$, i.e those from the sample at hand are not used for the census simulated vectors**
- Fitting method has advantages such as allows for sampling weights and heteroskedasticity

This led to the revision of the EB method we used. We wanted to make it like Molina and Rao's (2010) approach but use the fitting method from van der Weide (2014)

1. The survey data is used to obtain $\hat{\theta}_0 = (\hat{\beta}_0, \hat{\sigma}_{\eta\,0}^2, \hat{\sigma}_{e\,0}^2)$ using Henderson's Method III
2. Use parameters from step 1 to simulate M vectors of welfare in the census data

$$y_{ch}^{*} = x_{ch}\hat{\beta}_0 + \eta_c^{*} + e_{ch}^{*}$$

Notice how the $\hat{\beta}_0$ is kept fixed across the vectors.

$\boldsymbol{\eta_c^{*} \sim N(\hat{\eta}_{c_0}, \widehat{\mathrm{var}}[\hat{\eta}_{c_0}])}$ for areas in the sample, otherwise $\boldsymbol{\eta_c^{*} \sim N(0, \hat{\sigma}_{\eta_0}^2)}$

- $\widehat{\mathrm{var}}[\eta_c^{*}] = \hat{\sigma}_{\eta\,0}^2 - \hat{\gamma}_c^2 \left(\hat{\sigma}_{\eta\,0}^2 + \sum_h \left(\frac{w_{ch}}{\hat{\sigma}_{e_{ch0}}^2} \right)^2 \hat{\sigma}_{e_{ch0}}^2 \right)$
- $\hat{\eta}_{c_0} = \hat{\gamma}_c \left(\sum_h \left(\frac{w_{ch}}{\hat{\sigma}_{e_0}^2} \right) \hat{e}_{ch} \right) \left(\sum_h \left(\frac{w_{ch}}{\hat{\sigma}_{e_0}^2} \right) \right)^{-1}$
- $\hat{\gamma}_c = \frac{\hat{\sigma}_{\eta\,0}^2}{\hat{\sigma}_{\eta\,0}^2 + \sum_h w_{ch}^2 \left(\sum_h w_{ch} \sum_h \frac{w_{ch}}{\hat{\sigma}_{e_{ch0}}^2} \right)^{-1}}$

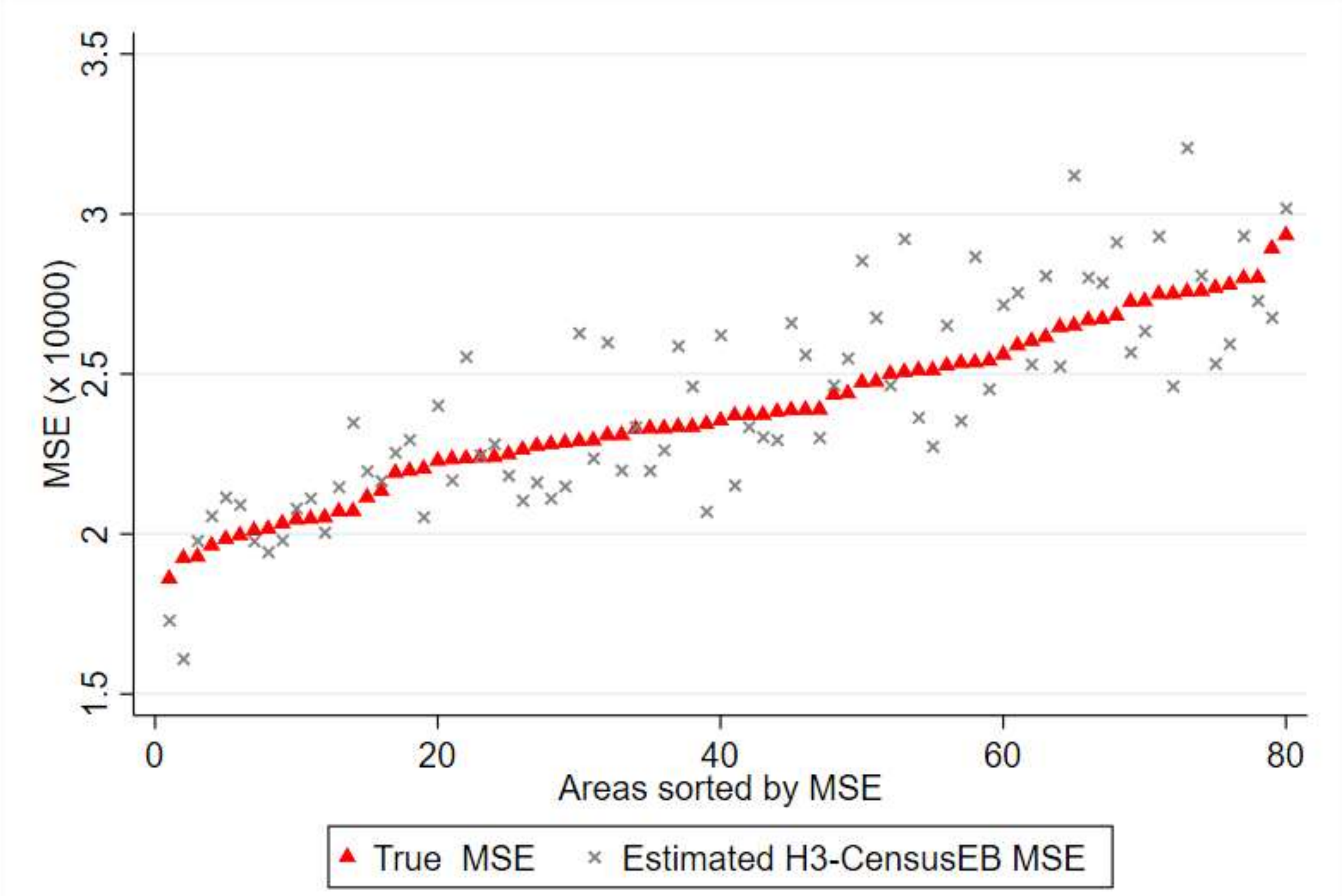
Household specific errors are drawn from $e_{ch}^{*} \sim N\left(0, \hat{\sigma}_{e_{ch0}}^2\right)$

All of this comes directly from van der Weide (2014) the only difference is in how it is applied.

In the absence of weights and heteroskedasticity all of these will equal those from Molina and Rao (2010)

More details in Corral, Molina and Nguyen (2021)

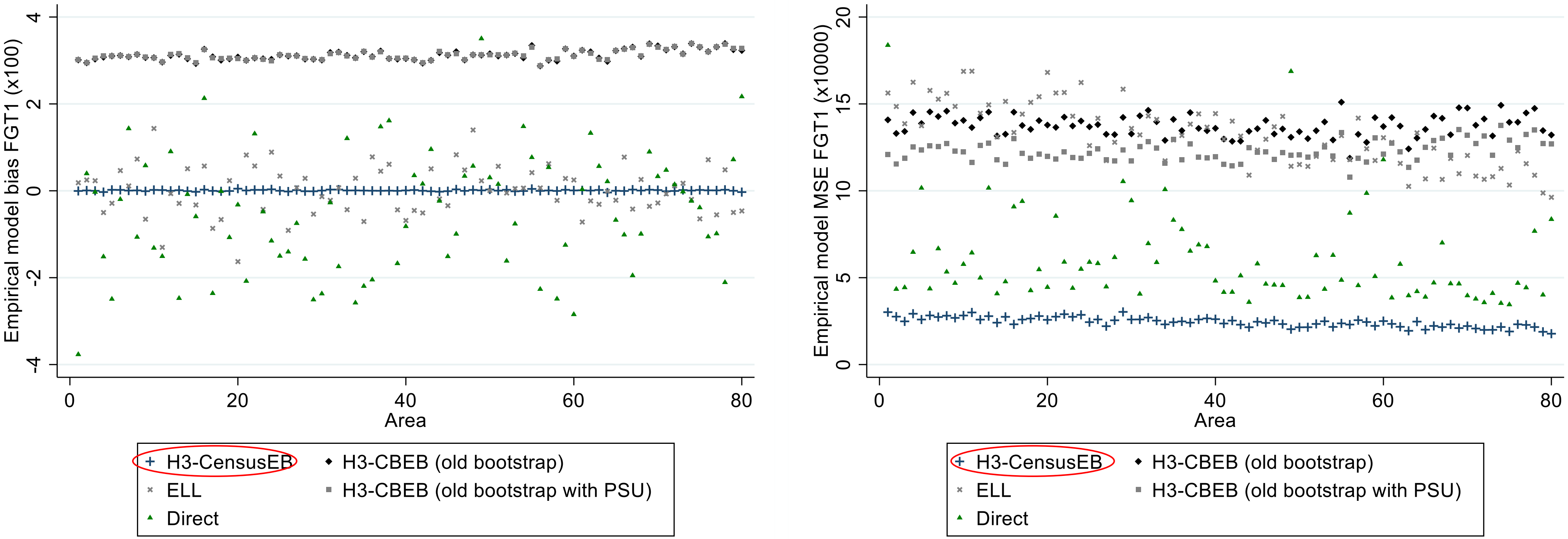
Now the MSE estimate is aligned to the empirical MSE…



note:True MSE is the outcome of 10k populations (size 100k), estimate is from bootstrap

Source: Corral, Molina, and Nguyen (2021)

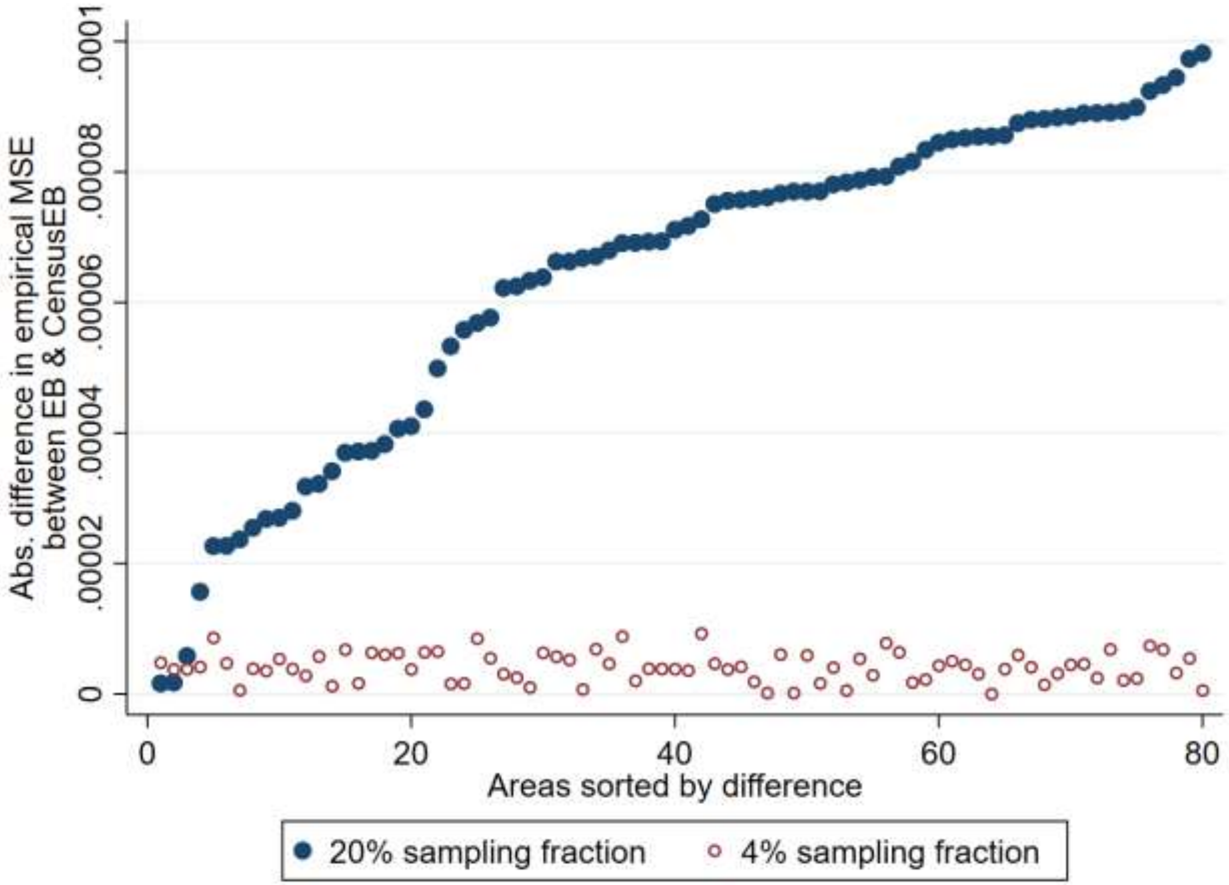
…and bias is near 0 for all areas and the empirical MSE is much lower



Source: Corral, Molina, and Nguyen (2021)

Updated H3-CensusEB presents less bias and lower MSE

…and CensusEB approximates EB as sample shares by area decrease



Source: Corral, Molina, and Nguyen (2021)

- EB estimator from Molina and Rao (2010) requires linking households, CensusEB doesn't link these

Okay, so how does this perform with real data?

Mexican Intra-censal data presents an opportunity to test models under real world scenarios

Survey is fielded by INEGI and consists of a sample of 5.9 million households

- Contains a measure of income at the household level
- After cleaning data we end up with a census of 3.9 million households (remove 90 percent of households reporting 0 income and all municipalities with less than 500 households)
  - End up with 1,865 municipalities and 16,297 PSUs
- Draw 500 survey samples following a similar sampling approach to LSMS surveys
  - Representative at the State level (32 states)
  - Total sample size of ~23,500
  - Number of municipalities included ranges from 951 to 1,020
  - The median municipality in the sample is represented by a single PSU

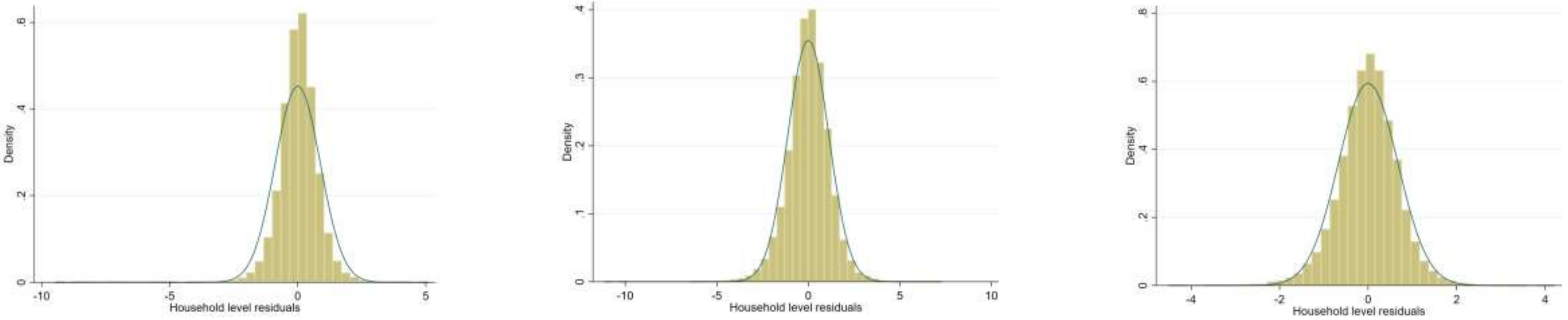
We test the following methods to obtain indicators at municipality level:

- CensusEB with **one**fold nested error model, random location effects specified at municipality level (CEB_a)
- CensusEB with **one**fold nested error model, random location effects specified at PSU level (CEB_c)
- CensusEB with **two**fold nested error model, random location effects specified at municipality and PSU level (CEB_{ac}) – Marhuenda et al. (2017)
- CensusEB with **two**fold nested error model, random location effects specified at state and municipality level (CEB_{sa}) – Marhuenda et al. (2017)
- ELL with **one**fold nested error model, random location effects specified at PSU level (ELL_c)

Results from Corral, Himelein, McGee, and Molina (2021)

Twofold nested error models are available in Stata's sae package

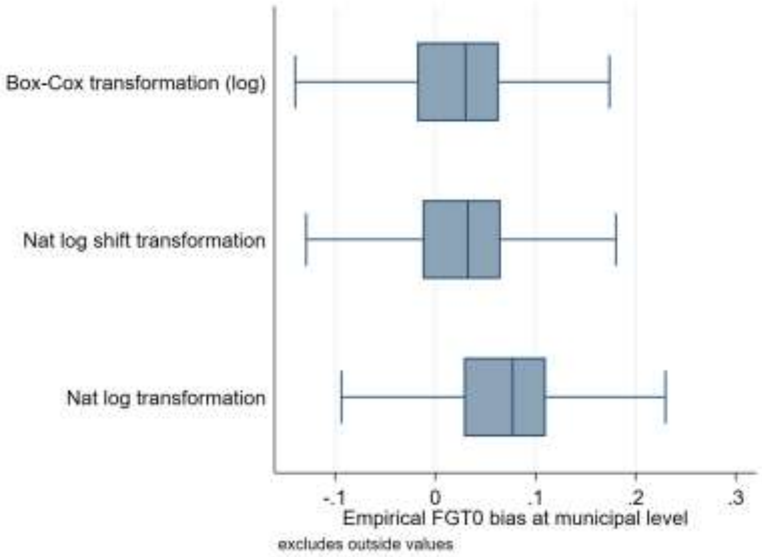
Transformation is necessary to achieve an approximately normal distribution…and it may lead to improved results


Nat. log

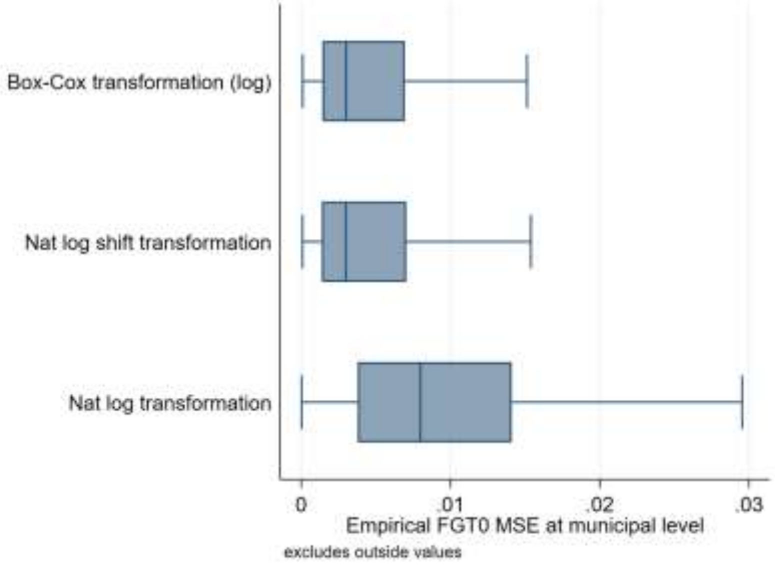
Box-Cox of Nat. log

Log shift

Bias (CEB_a)

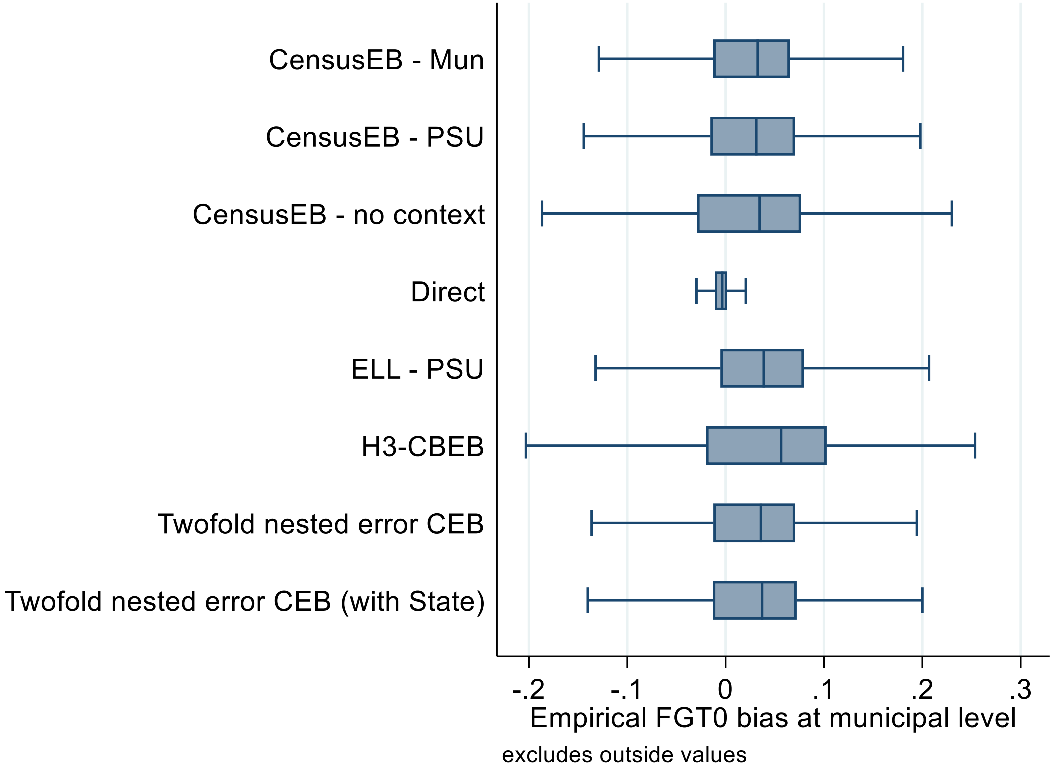


MSE (CEB_a)

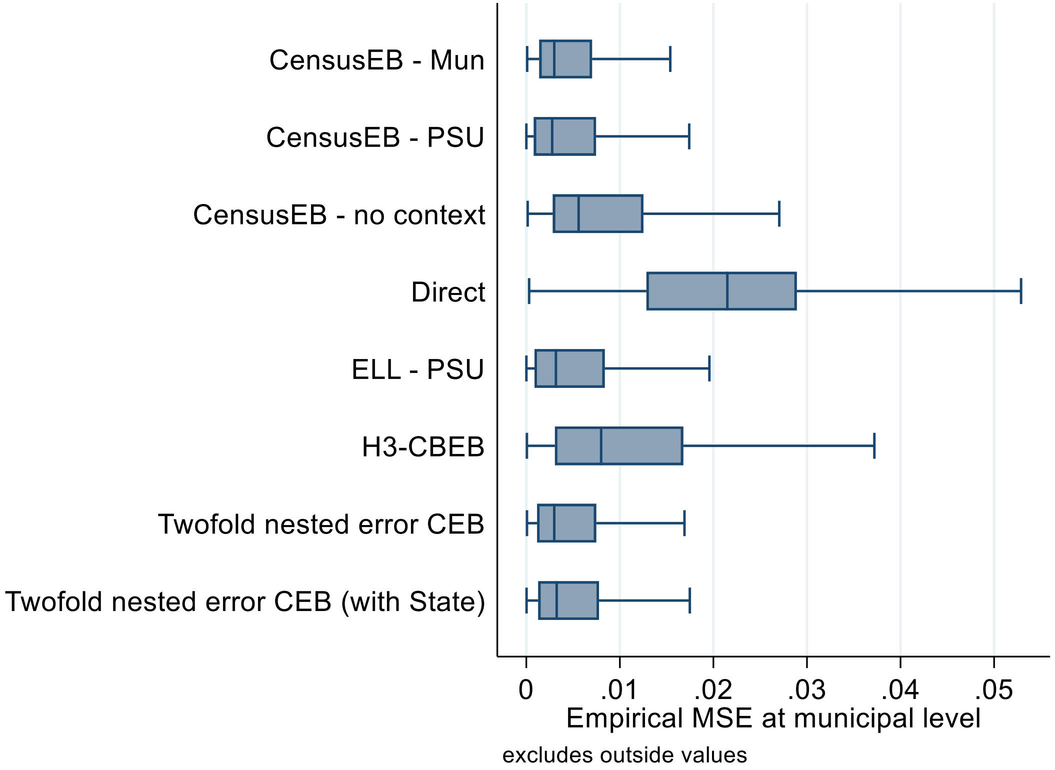


Under two-stage sampling and sample sizes like those observed in real world scenarios most methods appear to perform better than direct estimates in MSE

Bias (mun. level for 1,865 mun.)



MSE (mun. level for 1,865 mun.)

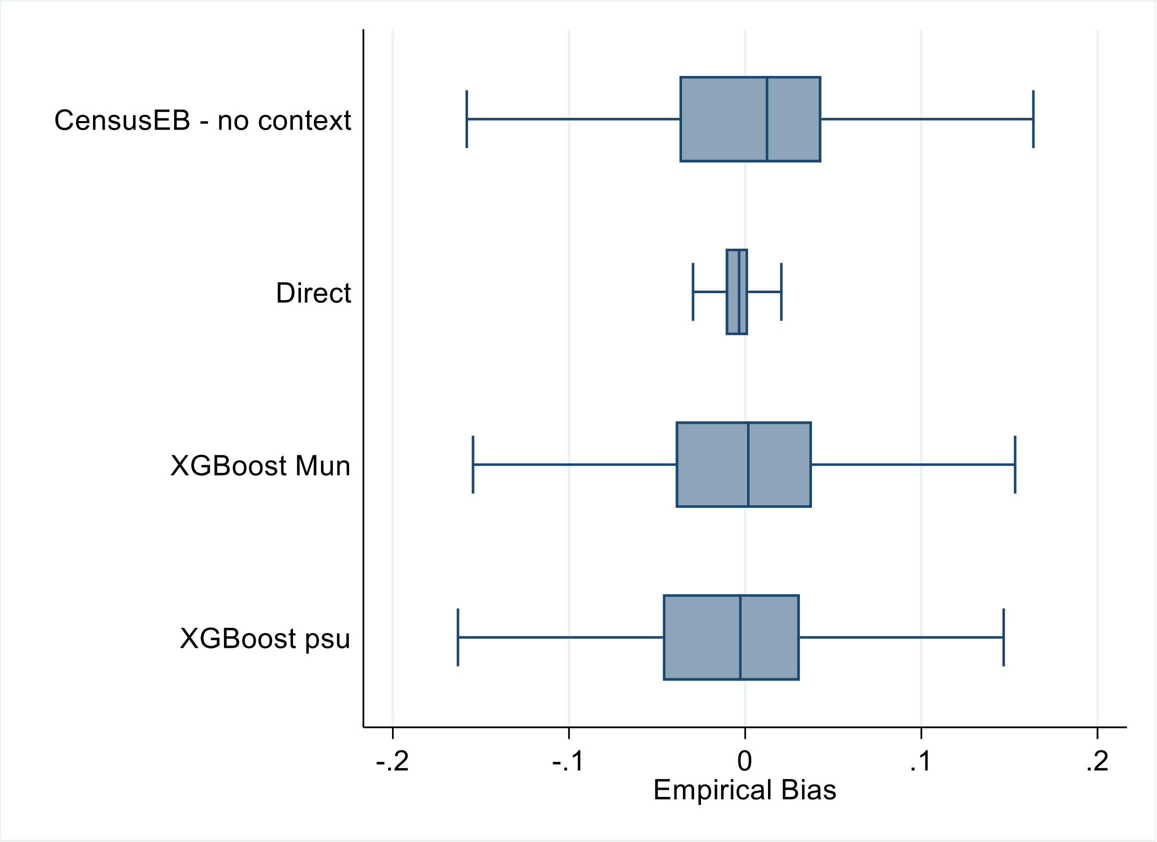


- Modeling random effects at the mun. level only when the true model follows a twofold nested error entails virtually no loss in efficiency (Marhuenda et al., 2017)

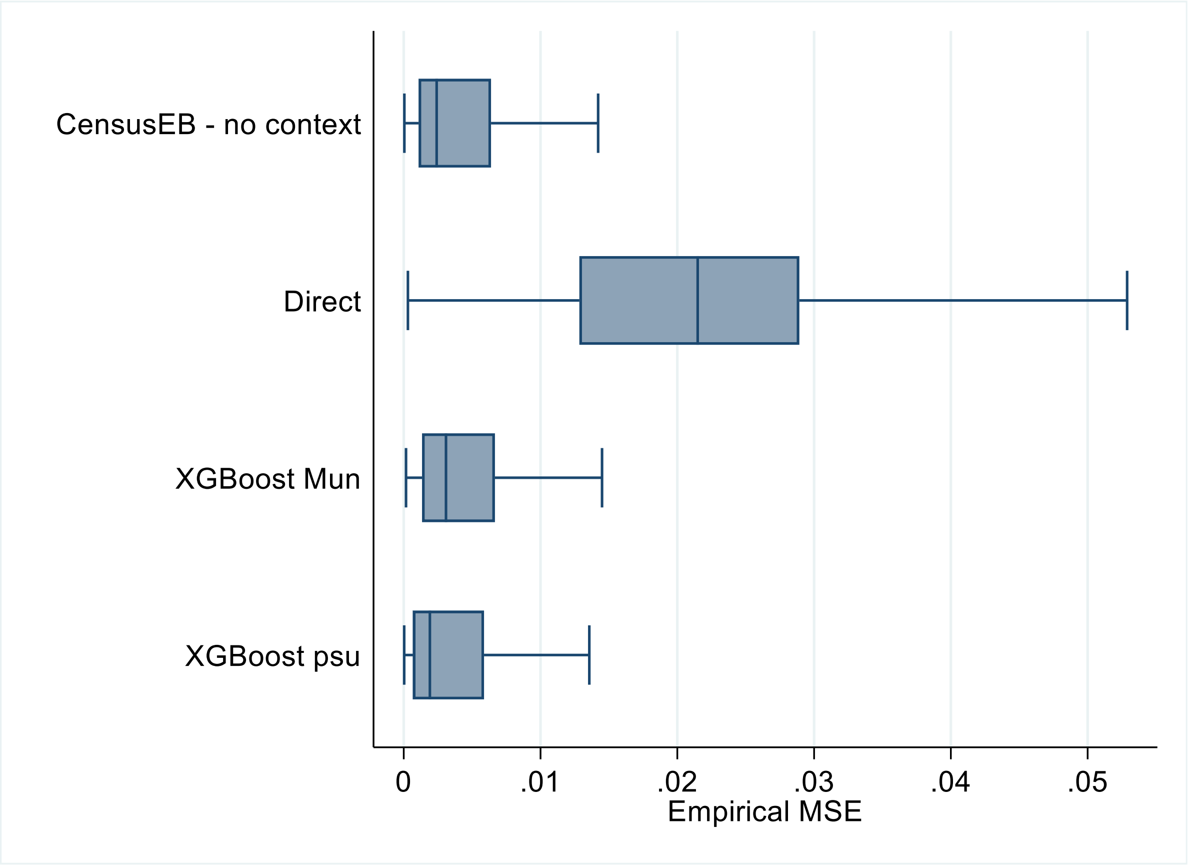
WORLD BANK GROUP

New methods could be an alternative for off census years, but more research is needed

FGT0 Bias (mun. level for 1,865 mun.)



MSE (mun. level for 1,865 mun.)



Sneak peek at upcoming work from Corral, Himelein, Rodriguez, and Segovia

Concluding remarks

- The updated EB (CensusEB) method works better than the previous method in all tested scenarios, with large or small sampling fractions, large or small population sizes, larger or smaller explanatory power of covariates, stronger or weaker location effects, even under heavier tails than normal
- The corresponding bootstrap procedure succeeds in estimating properly the true MSE, unlike the previously considered procedures based on MI methods.
- Twofold models are appropriate for two-stage sampling, but in absence of available software or for simplicity, area effects are enough. Specifying effect at lower levels and aggregating to the area is not recommended.
- Poverty mapping guidelines inspired on the presented work will lead to improved SAE done at the WB
- Research focusing on poverty maps during off census years is still nascent