Producing high quality small area estimates of poverty

Small Area Estimation at the World Bank

Poverty and Equity Global Practice

February 1, 2021

WORLD BANK GROUP

Paul Corral
pcorralrodas@worldbank.org

Based on work done with:
Isabel Molina, Alexandru Cojocaru, and Sandra Segovia

Based on work from the recently completed poverty mapping guidelines

Corral, P., Molina, I., Cojocaru, A. and Segovia, S. **Guidelines to Poverty Mapping.** *Forthcoming*

The latest Stata sae package can be obtained from:
https://github.com/pcorralrodas/SAE-Stata-Package

WORLD BANK GROUP

Introduction

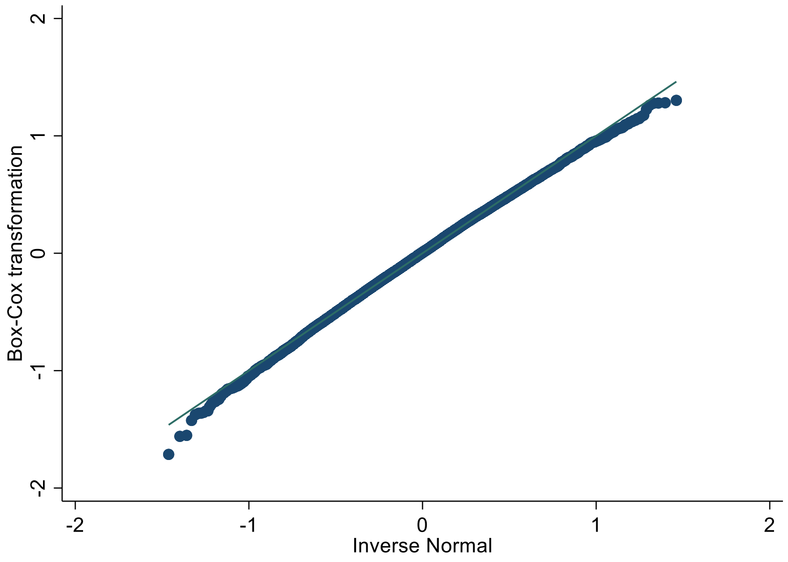
1. Poverty is a non-linear parameter
2. Under the traditional unit-level method (ELL 2003 or Molina and Rao 2010) the goal is to simulate the welfare distribution and from this welfare distribution obtain poverty estimates for each area
3. When this is not possible, it is not advisable to attempt to model household level welfare in the absence of household characteristics
4. In instances where the model used for simulating the welfare distribution is unrealistic (e.g. assuming a model without household level characteristics) or we lack a contemporaneous census we want to rely on area-level models
5. The poverty mapping guidelines build upon previous work and illustrate pros and cons of some of the most often applied methods for poverty mapping

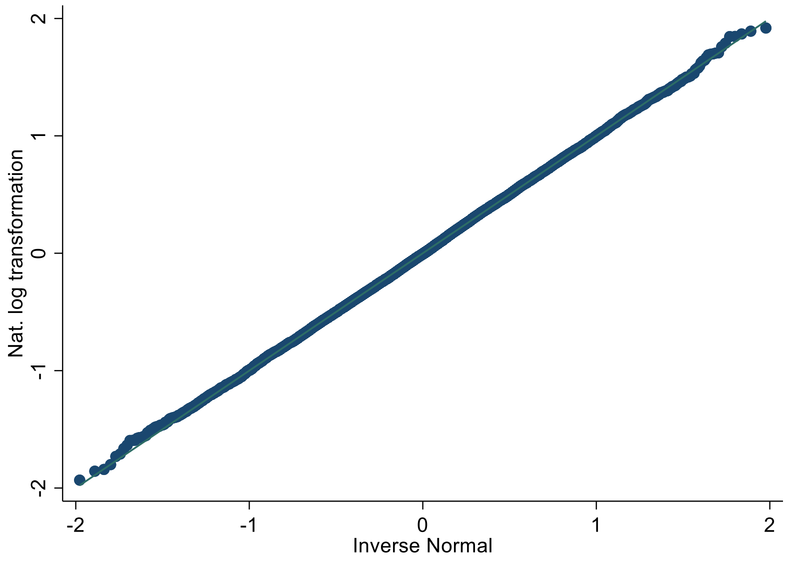
The latest Stata sae package can be obtained from: https://github.com/pcorralrodas/SAE-Stata-Package

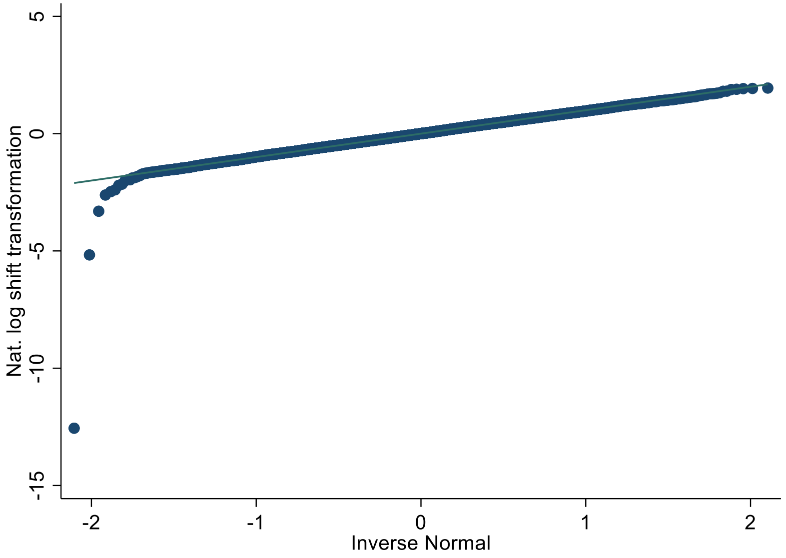
Poverty is non-linear and under unit-level models we need to take care in replicating the entire distribution otherwise we risk ending up with biased estimates

This is done to ensure we can obtain accurate estimates regardless of the poverty line

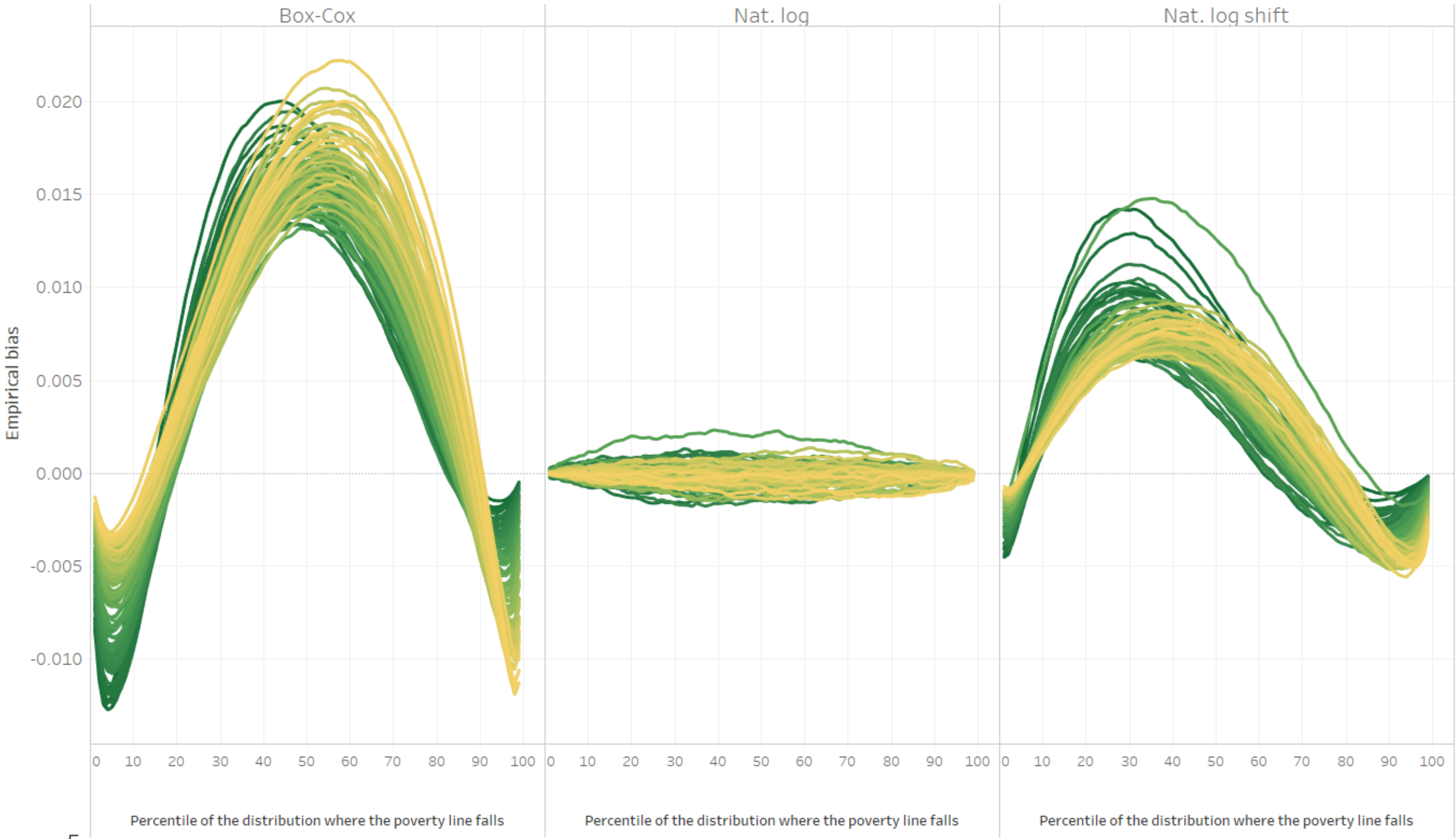
- It entails ensuring the embedded model assumptions hold: mainly normally distributed residuals under Molina and Rao's EB approach
- **Deviations from the embedded model assumptions will lead to biased and noisy estimates for the areas of interest**







Unit-level models rely on being able to replicate the full distribution of welfare and the wrong transformation can lead to biased estimates



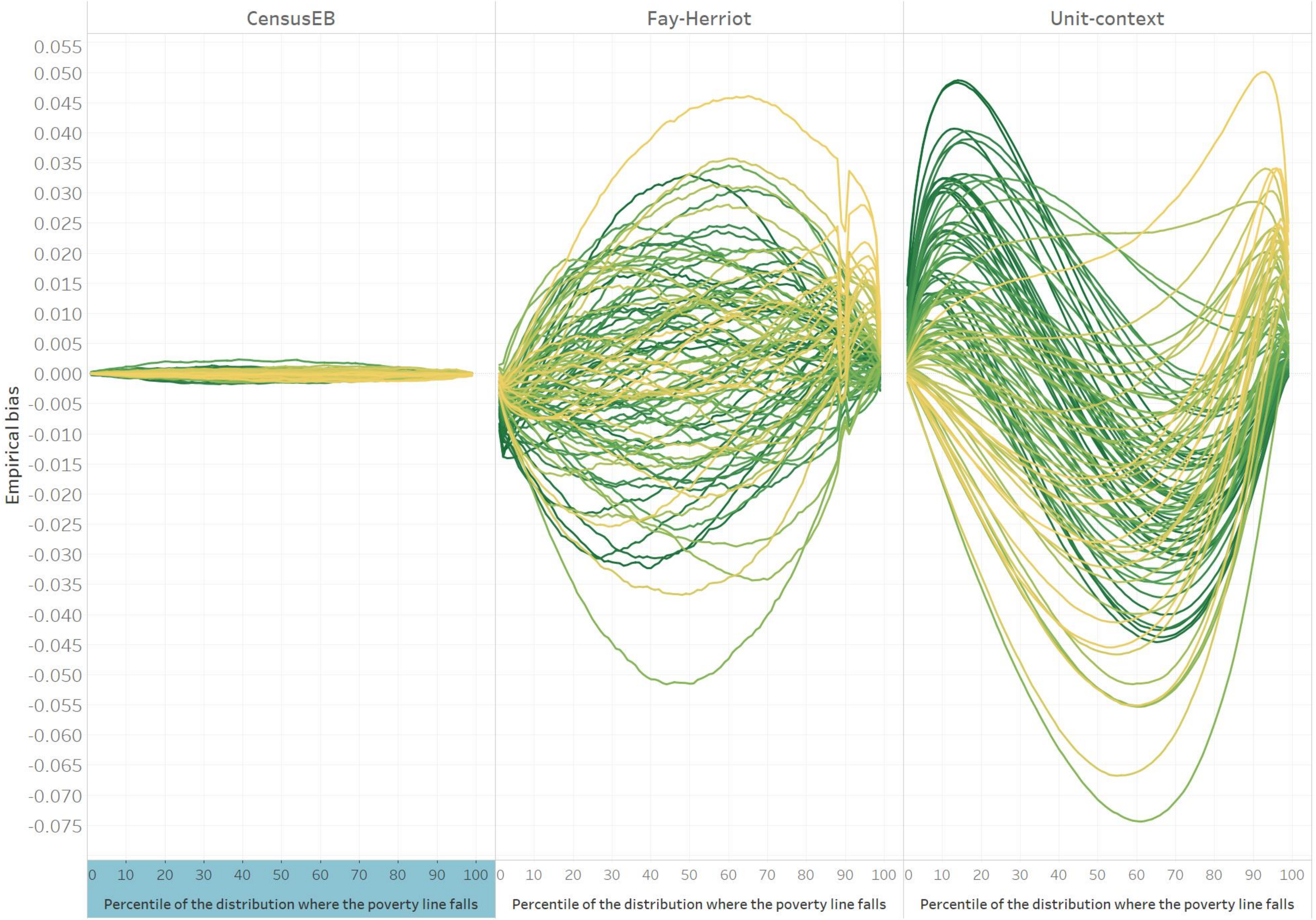
- In this scenario, the natural logarithm is the correct transformation for the data at hand
    - For all areas and all poverty thresholds it yields the smallest bias
    - Every line represents an area
    - The x-axis is the poverty line
- Note how the wrong transformation may work for some thresholds – even a broken watch will give the right time of day twice a day

WORLD BANK GROUP

When census and survey are not from the same year, small area estimates based on unit level models may result in biased estimates

- De facto approach is through area level models (Fay and Herriot, 1979) or a twofold variant by Torabi and Rao (2014)
  - Fay-Herriot models offer small gains over direct estimates of FGT0 and FGT1, Torabi and Rao's may offer larger gains – no current software implementation
- **Alternative proposed by Nguyen (2012)** relies on an ELL approach using aggregated covariates in the model for household level welfare. Afterwards, **Masaki et al. (2020)** updates this to Molina and Rao's EB.
  - The method is not recommended as it will not replicate the welfare distribution:
    - » **Yields biased estimators**
    - » **Parametric bootstrap is not adequate to measure the method's noise**

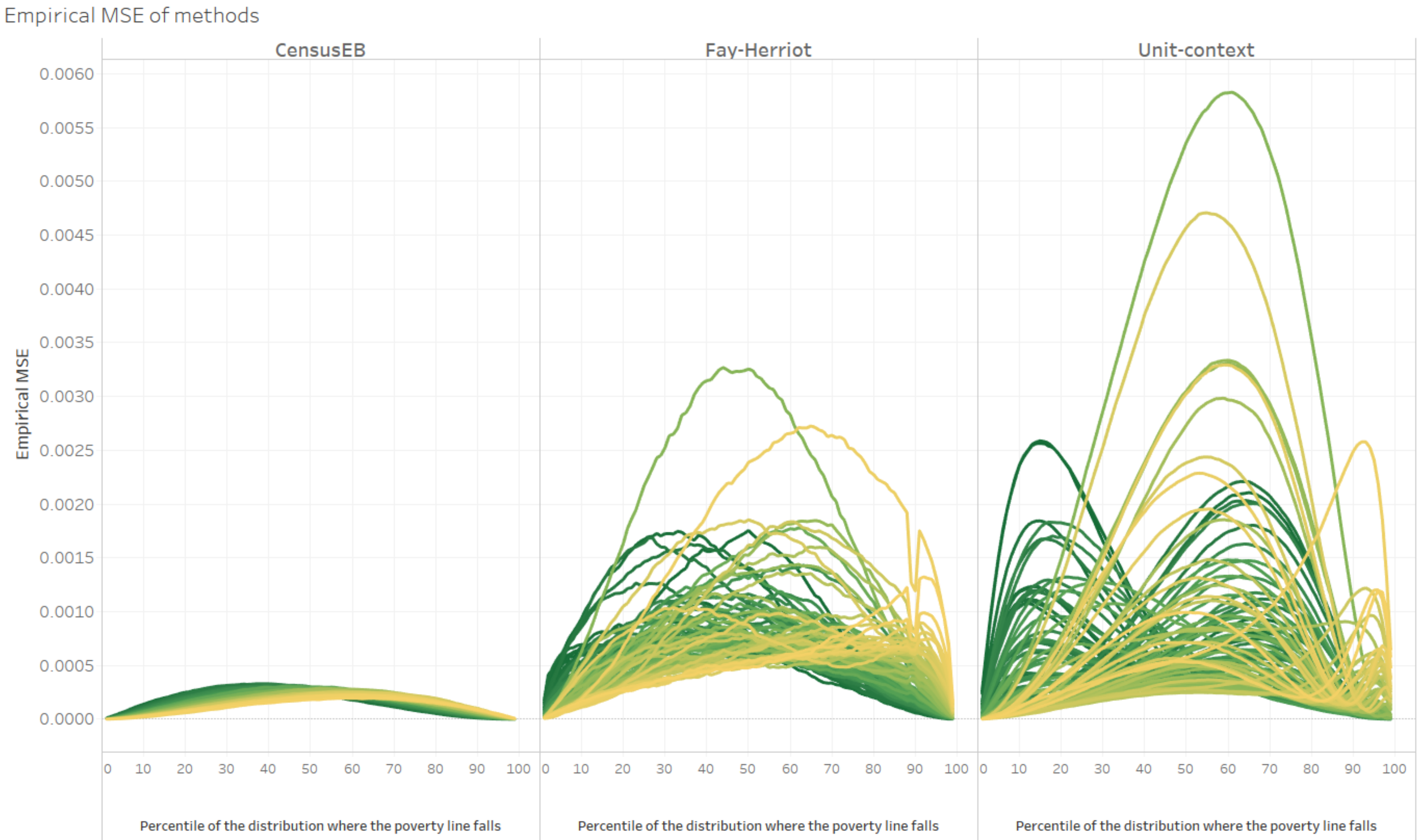
Area level models, dating back to 1979, are still the go-to method for off-census years



- The alternative is a Fay-Herriot model
  - Requires a different model for each poverty line
  - Difficult to always find suitable covariates
  - Doesn't work in areas where sampling variance of direct estimates is 0
- **But it is less biased than unit-context models like those of Masaki et al (2020) and Nguyen (2012)…**

WORLD BANK GROUP

Area level models, dating back to 1979, are still the go-to method for off-census years

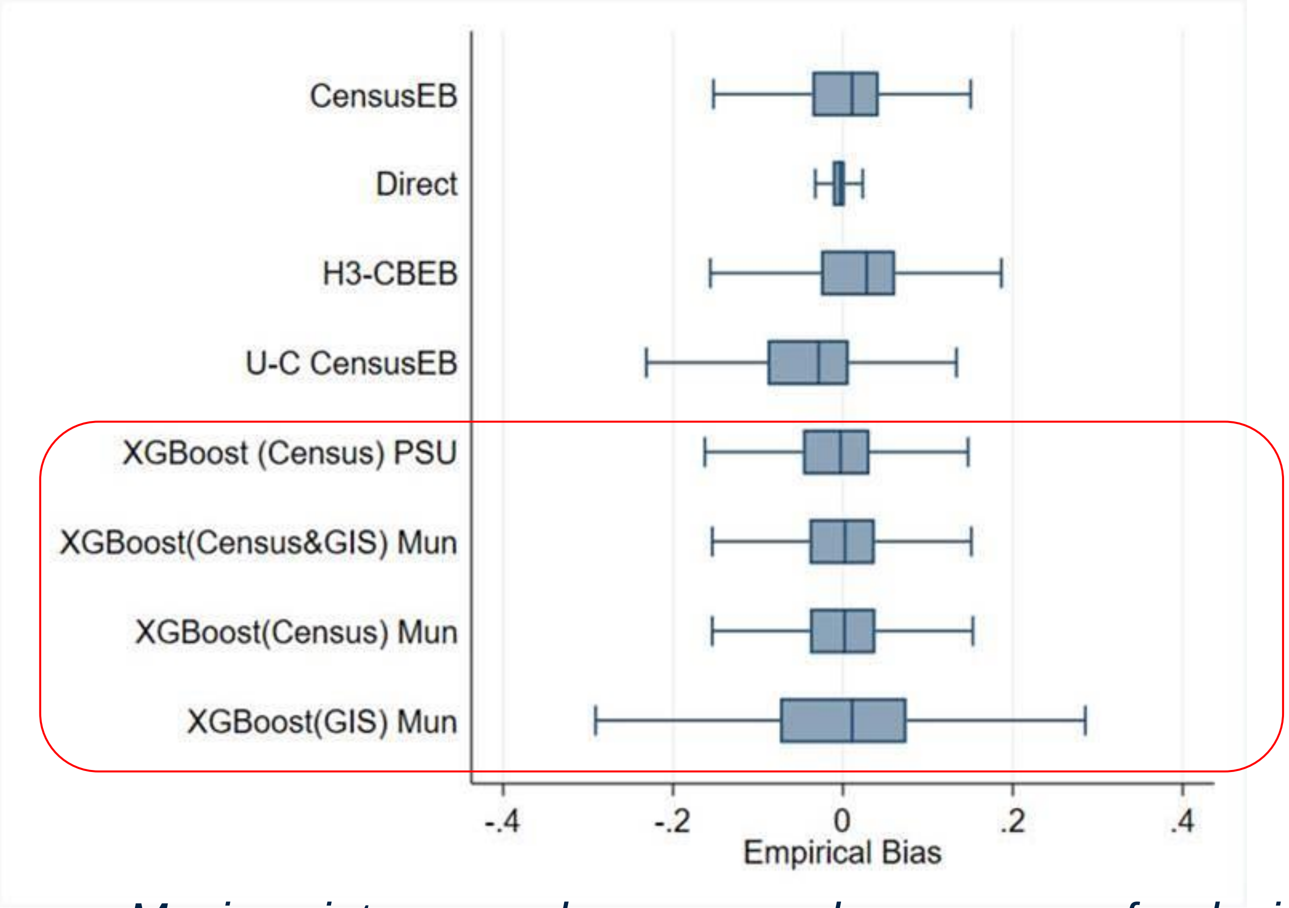


- The alternative is a Fay-Herriot model
  - Requires a different model for each poverty line
  - Difficult to always find suitable covariates
  - Doesn't work in areas where sampling variance of direct estimates is 0
- **And less noisy than unit-context models like those of Masaki et al (2020) and Nguyen (2012)…**

WORLD BANK GROUP

What about machine learning approaches? A validation with real world data...

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



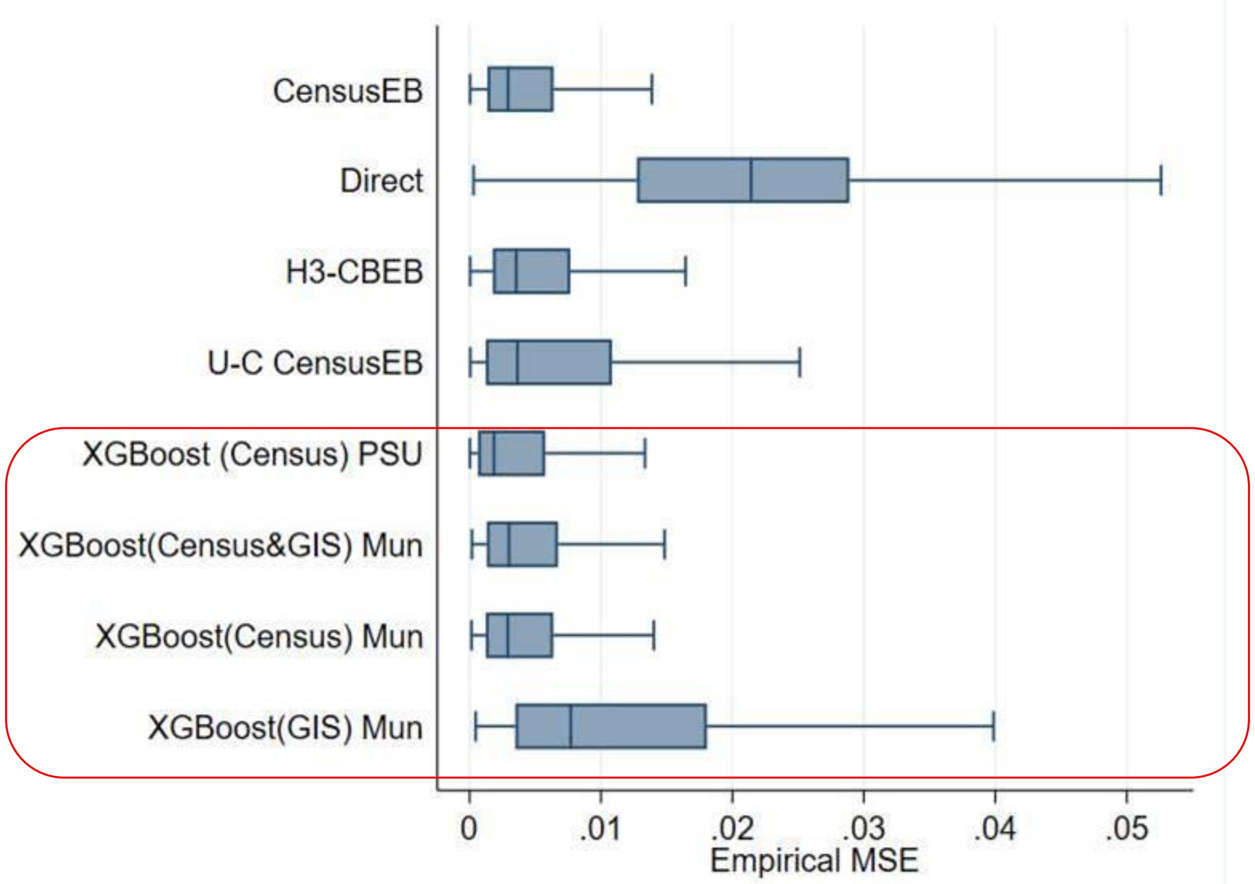
- *Still work in progress*
- All methods rely on census household and/or area aggregate covariates, unless stated otherwise
- Notice the bias of the different methods
  - Direct estimates are the least biased
  - Notice how the box-plot for unit context models, those that model household level welfare using only area level covariates, is downward biased
  - XGboost yields good results. In some cases as good as censusEB
  - Results for XGboost from models using just GIS covariates are less stellar

Mexican intra-censal survey used as a census for design-based simulation based on 500 samples.

What about machine learning approaches? A validation with real world data…

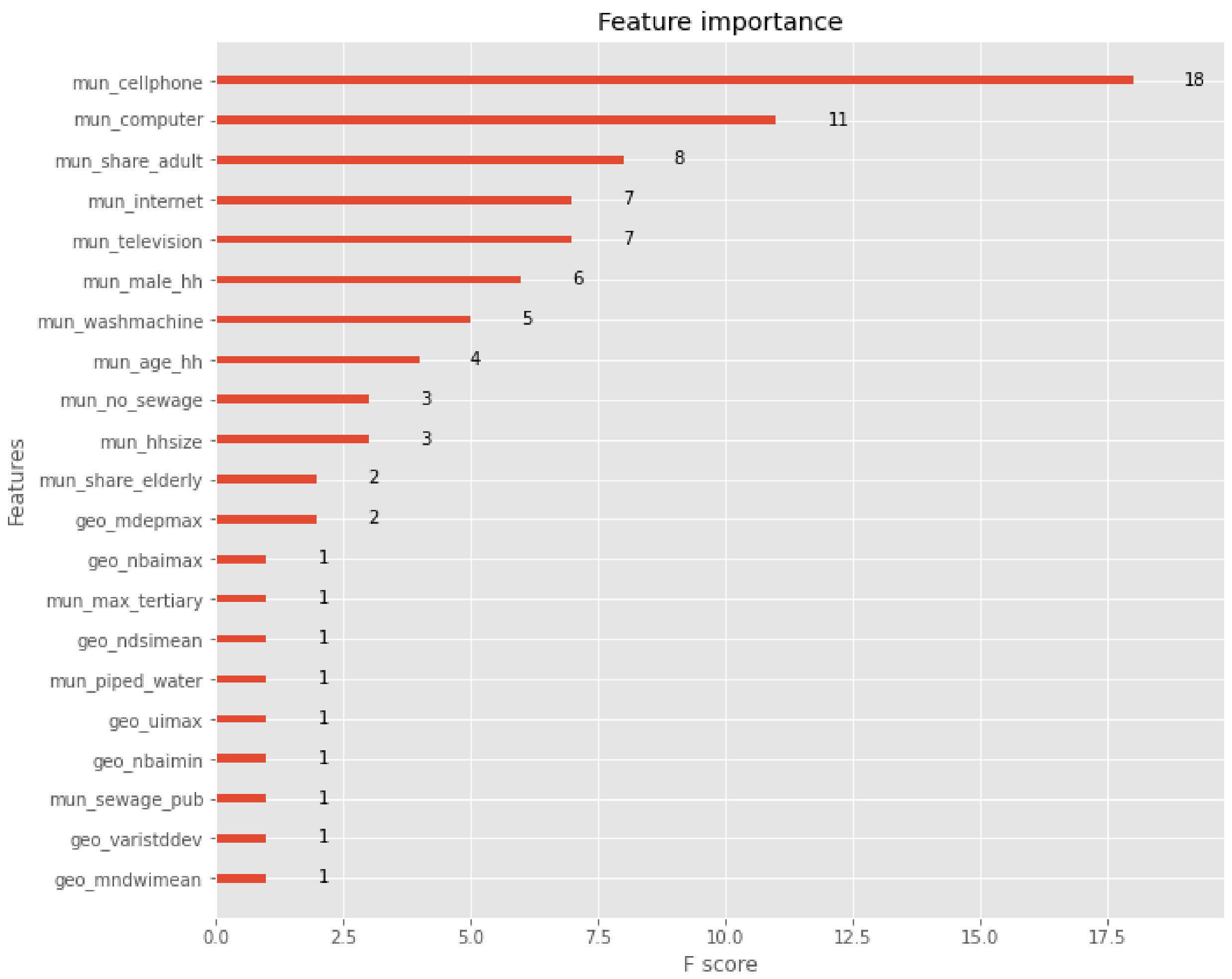
- Notice the MSE of the different methods
  - Direct estimates are the least biased, but noisiest
  - XGboost fit at the municipal level with just GIS covariates is the second noisiest
  - CensusEB, as expected, shows the smallest MSE, followed very closely by XGboost fit at the PSU level with just census aggregates
  - XGboost in this scenario outperforms unit-context models as those presented by (Masaki et al. 2020)

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



Mexican intra-censal survey used as a census for design-based simulation based on 500 samples.

Gradient boosting validation with Mexican data seems to work well but it all depends on the quality of the data you feed the model



- The model works best with census data aggregates
- When combining census aggregates and GIS covariates the method leans towards census aggregates
- The empirical MSE of the gradient boosting models is smaller when only using census aggregates
  - In this scenario it suggests that census aggregates are better than GIS at explaining the variation of area level poverty

Covariates with mun_ prefix are derived from the census, geo_ are GIS covariates…

Concluding remarks

- When conducting SAE it is necessary to ensure the model's assumptions are met
  - For poverty under unit-level models it ensures unbiased estimates as well as correctly estimated MSE
  - Under Fay-Herriot models the BLUP does not require normality assumptions for unbiasedness, but does rely on it for MSE estimation
- For off-census years more research and innovation is needed to find approaches that may yield estimates of high quality
  - Machine learning approaches still require high-quality data to obtain high-quality estimates but do seem to offer promising results in an application with Mexican data
    - We don't know how well the method may work with other data in different contexts

WORLD BANK GROUP