All the Small Things:
An Introduction to Small Area Estimation

Flavio Morelli

November 8, 2021

About me

- M.Sc. in Statistics at Humboldt University of Berlin
- Wrote my thesis on poverty estimation with Bayesian methods
- Co-organizer of the Berlin Bayesians Meetup (@BBayesians)
- Machine Learning Intern @ Bayer Pharmaceuticals

Disclaimer

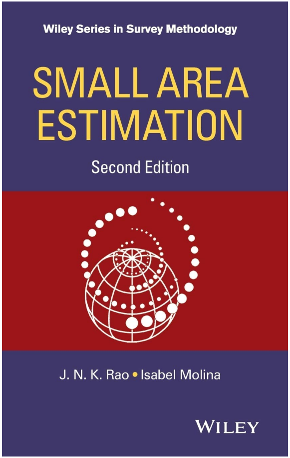
- SAE is a deeply frequentist field...
- ... and I am more of a Bayesian!

Contents

1. Overview of SAE
2. Design-Based Estimators
3. Model-Based Estimators
4. Use Case: Poverty Estimation

Contents

1. Overview of SAE
2. Design-Based Estimators
3. Model-Based Estimators
4. Use Case: Poverty Estimation



Statistical Science
2013, Vol. 28, No. 1, 40–68
DOI: 10.1214/12-STS395

New Important Developments in Small Area Estimation

Danny Pfeffermann

Abstract. The problem of small area estimation (SAE) is how to produce reliable estimates of characteristics of interest such as means, counts, quantiles, etc., for areas or domains for which only small samples or no samples are available, and how to assess their precision. The purpose of this paper is to review and discuss some of the new important developments in small area estimation methods. Rao [*Small Area Estimation* (2003)] wrote a very comprehensive book, which covers all the main developments in this topic until that time. A few review papers have been written after 2003, but they are limited in scope. Hence, the focus of this review is on new developments in the last 7–8 years, but to make the review more self-contained, I also mention shortly some of the older developments. The review covers both design-based and model-dependent methods, with the latter methods further classified into frequentist and Bayesian methods. The style of the paper is similar to the style of my previous review on SAE published in 2002, explaining the new problems investigated and describing the proposed solutions, but without dwelling on theoretical details, which can be found in the original articles. I hope that this paper will be useful both to researchers who like to learn more on the research carried out in SAE and to practitioners who might be interested in the application of the new methods.

Key words and phrases: Benchmarking, calibration, design-based methods, empirical likelihood, informative sampling, matching priors, measurement errors, model checking, M-quantile, ordered means, outliers, poverty mapping, prediction intervals, prediction MSE, spline regression, two-part model.

1. PREFACE

The problem of small area estimation (SAE) is how to produce reliable estimates of characteristics of inter-

The great importance of SAE stems from that many new programs, such as fund alloc needed areas, new educational or health progr environmental planning rely heavily on th

What is Small Area Estimation?

- *Aim:* produce reliable estimates for small areas
- *Small area:* subdivision with few ($\leq$30) or no available observations, not necessarily geographical
- *Estimates:* means, count, quantiles, etc.
- *Reliable:* point estimator + prediction error for each area

SAE is mainly a **prediction** task!

Types of SAE Methods

The two main type of models are:

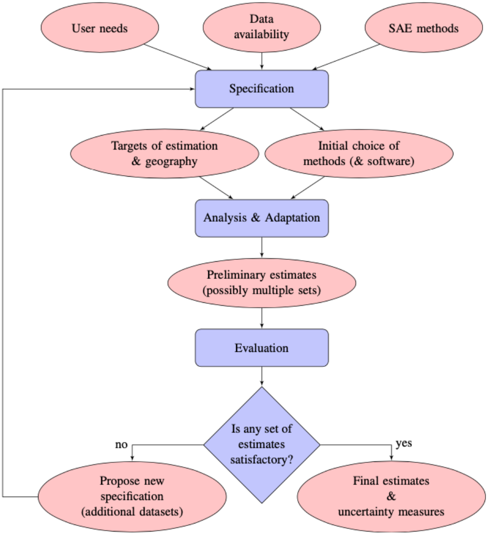
- Design-based
- *Model-based* (frequentist, empirical Bayesian, full Bayesian)

Both methods have in common the use of **auxiliary data** such as censuses, registers or larger surveys.

Depending on the type of auxiliary data they can be further classified into:

- Area-level methods
- Unit-level methods

SAE Framework



Source: Tzavidis et al. (2018)

Contents

1. Overview of SAE
2. Design-Based Estimators
3. Model-Based Estimators
4. Use Case: Poverty Estimation

Design-based estimators

- Bias, variance and other properties are evaluated wrt the design-based distribution
- However: may use a model for the construction of the estimators
- Two types: **direct** and **synthetic**

Horvitz-Thompson direct estimator for mean

$$\hat{\theta}_d^{HT} = \frac{\sum_i y_{di} w_{di}}{N_d} = \theta_d + e_d.$$

- y_{di} is the variable of interest for area d
- w_{di} are the survey weights for area d
- N_d is the size of area d
- θ_d is the expectation for variable y_d in region d
- $e_d \overset{\text{ind}}{\sim} (0, \psi_d)$ is the **sampling error**
- ψ_d is taken as given
- Note: $E[\hat{\theta}_d^{HT}] = E[\theta_d + e_d] = \theta_d$.

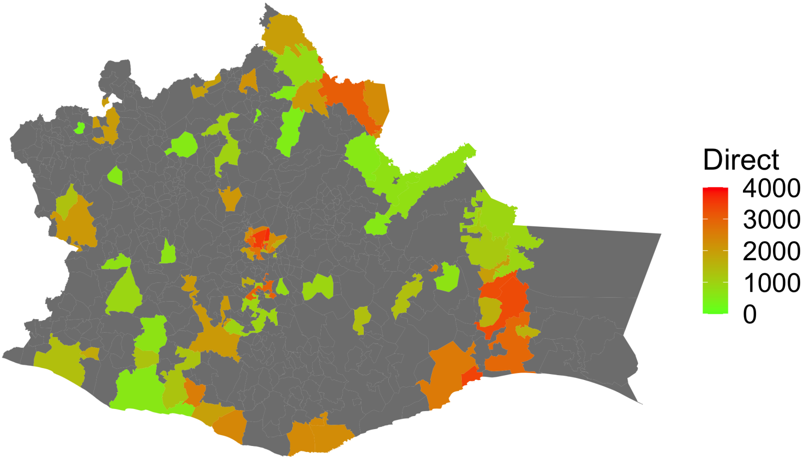
More direct estimators

Hajek direct mean estimator (small bias):

$$\hat{\theta}_d^{Hajek} = \frac{\sum_i y_{di} w_{di}}{\sum_i w_{di}}$$

- Not very useful when there are missing values
- However, it is commonly used for benchmarking because it is (nearly) unbiased

Example: income in Oaxaca



Synthetic estimators

- Main idea: regress y_{di} on a set of auxiliary variables x_{di}
- $\hat{\theta}_d^{syn} = \frac{1}{N_d} \sum_{i=1}^{N_d} x'_{di} \hat{\beta}$
- Rarely used by themselves due to sensitivity to regression coefficients.

Contents

1 Overview of SAE

2 Design-Based Estimators

3 Model-Based Estimators

4 Use Case: Poverty Estimation

Area-level: Fay-Herriot Model

- Assume $\theta_d = \mathbf{x_d}'\boldsymbol{\beta} + u_d, \quad u_d \overset{\text{iid}}{\sim} (0, \sigma_u^2)$ (Rao & Molina, 2015)
- u_d is the area-specific effect
- Combine with the direct estimate to get Fay-Herriot **model**:
$$\hat{\theta}_d^{direct} = \theta_d + e_d = \mathbf{x_d'}\boldsymbol{\beta} + u_d + e_d$$
- Define interclass correlation $\hat{\gamma}_d = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_d}$
- Fay-Herriot **estimator** is given by
$$\hat{\theta}_d^{FH} = \hat{\gamma}_d \hat{\theta}_d^{direct} + (1 - \hat{\gamma}_d)\mathbf{x}_d'\hat{\boldsymbol{\beta}}$$

Bayesian Fay-Herriot model

- Rewrite the Fay-Herriot model as a hierarchical Bayesian model (Rao & Molina, 2015):

$$
\begin{aligned}
\hat{\theta}_d^{direct}|\theta_d, \boldsymbol{\beta}, \sigma_u^2 &\sim \mathcal{N}(\theta_d, \psi_d), \quad i = 1, ..., D, \\
\theta_d|\boldsymbol{\beta}, \sigma_u^2 &\sim \mathcal{N}(\mathbf{x}_d'\boldsymbol{\beta}, \sigma_u^2) \quad i = 1, ..., D, \\
\pi(\boldsymbol{\beta}, \sigma_u^2) &\propto g(\boldsymbol{\beta}, \sigma_u^2).
\end{aligned}
$$

- The hierarchical Bayes (HB) estimate $\hat{\theta}_d^{HB}$ is calculated similarly to the frequentist counterpart, but with the posterior parameters

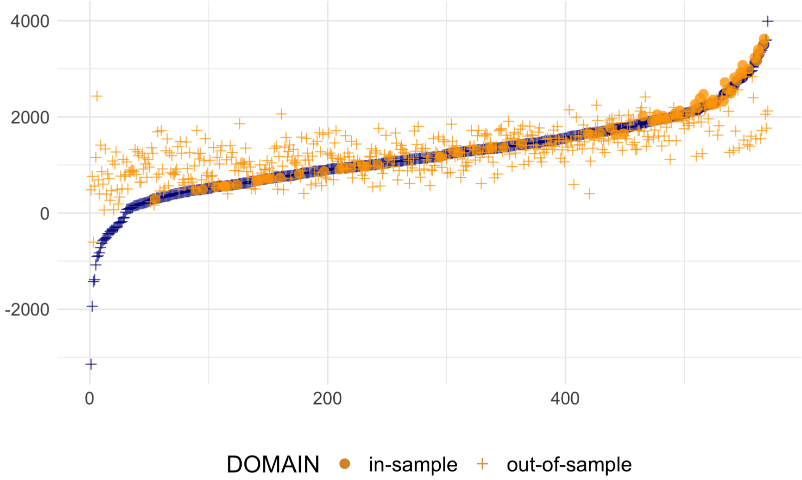
Hierarchical Bayes estimator

For $s = 1, ..., S$

1. Sample $\hat{\boldsymbol{\beta}}^{(\mathbf{s})}, \hat{\sigma}_u^{(s)}, \hat{\theta}^{(s)}$ from the posterior distribution.
2. Sample $\tilde{\theta}_i^{(s)}|\theta$ from the posterior predictive distribution. There are two cases:
    1. If municipality i is in-sample, then calculate $\hat{\gamma}_i^{(s)} = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \psi_i)$ and estimate $\tilde{\theta}_i^{(s)}|\theta = \hat{\gamma}_i^{(s)} + (1 - \hat{\gamma}_i^{(s)})\hat{\theta}_i^{(s)}$.
    2. If municipality i is out-of-sample, generate $\tilde{\theta}_i^{(s)}|\theta$ from $\mathcal{N}(\mathbf{x'}_i\hat{\boldsymbol{\beta}}^{(\mathbf{s})}, \hat{\sigma}_u^{(s)})$.
3. Finally, $\hat{\theta}_i^{HB} = \dfrac{1}{S}\sum_{s=1}^{S}\tilde{\theta}_i^{(s)}$ and

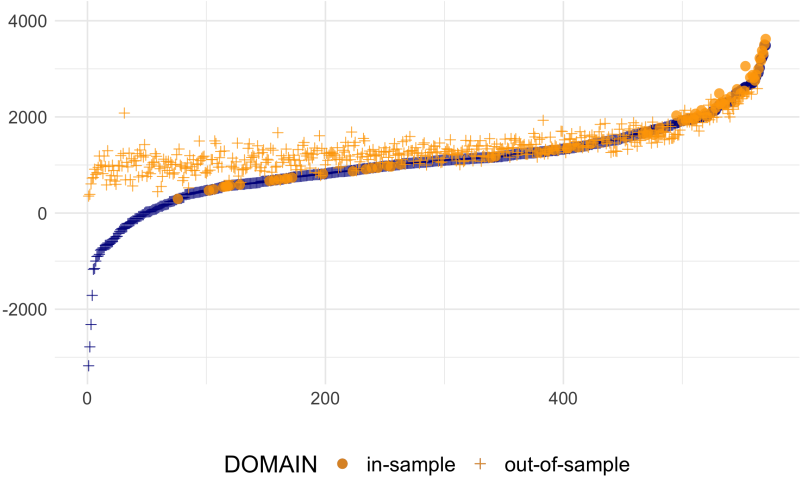
$$\hat{\sigma}_i^{HB} = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(\tilde{\theta}_i^{(s)} - \hat{\theta}_i^{HB}\right)^2}.$$

Example: FH income estimates for Oaxaca (large model)



Blue: frequentist. **Orange**: HB

Example: FH income estimates for Oaxaca (small model)



Blue: frequentist. **Orange**: HB

Unit-level models: Battese-Harter-Fuller model

Key Concept:
Include random area-specific effects to account for between-area variation/ unexplained variability between the small areas.

Random effects model:
Notation: (d =domain, i =individual)

$$y_{di} = \mathbf{x}_{di}^T \boldsymbol{\beta} + u_d + e_{di}, i = 1, ..., N_d, d = 1, ..., D$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}.$$

- Random effects $u_d \sim N(0, \sigma_u^2)$
- Error term $e_{di} \sim N(0, \sigma_e^2)$
- Sample size in area d is denoted by n_d
- $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with $\mathbf{V} = \sigma_u^2 \mathbf{Z}\mathbf{Z}^T + \sigma_e^2 \mathbf{I}_n$.

Unit-level models: Battese-Harter-Fuller model

Empirical Best Linear Unbiased Predictor (EBLUP) of $\bar{y}_d$ is

$$\hat{\theta}_d^{BHF} = \hat{\bar{y}}_d = N_d^{-1}\Big\{ \sum_{i \in s_d} y_{di} + \sum_{i \in r_d} \hat{y}_{di} \Big\} = N_d^{-1}\Big\{ \sum_{i \in s_d} y_{di} + \sum_{i \in r_d} (\mathbf{x}_{di}^T \hat{\boldsymbol{\beta}} + \hat{u}_d) \Big\}$$

where

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{y} \\ \hat{\mathbf{u}} &= \hat{\sigma}_u^2 \mathbf{Z}^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \hat{\mathbf{V}} &= \hat{\sigma}_u^2 \mathbf{Z}\mathbf{Z}^T + \hat{\sigma}_e^2 \mathbf{I}_n \end{aligned}$$

The variance components are estimated by ML or REML.
Problem: Not useful for non-linear estimators like poverty!

Poverty Indicators: FGT

$$F_d(\alpha, t) = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{t - y_{di}}{t} \right)^{\alpha} I(y_{di} \leq t), \qquad \alpha = 0, 1, 2.$$

- FGT-indicators (Foster, Greer, & Thorbecke, 1984)
- t is the poverty line set at 60% of median income of the state
- y_{di} is the income of the i-th person in municipality d
- $I(\cdot)$ is the indicator function
- $\alpha = 0$ is the head count ratio (HCR)
- $\alpha = 1$ is the poverty gap (PGAP)
- $\alpha = 2$ is the poverty severity
- Hard to estimate with regression only

Unit-level models: the EBP approach

Point of departure: Random effects model

$$y_{di} = \mathbf{x}_{di}^T\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, n_d, \quad d = 1, \dots, D,$$

Estimation process:

1. Use sample data to estimate $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{u}_d$ and $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d}}$.
2. For $s = 1, ..., S$
   - Generate $e_{di}^* \sim N(0, \hat{\sigma}_e^2)$ and $u_d^* \sim N(0, \hat{\sigma}_u^2 \cdot (1 - \hat{\gamma}_d))$ and obtain a pseudo-population

     $$y_{di}{}^{*(s)} = \mathbf{x}_{di}^T\hat{\beta} + \hat{u}_d + u_d^* + e_{di}^*$$

   - Calculate the poverty measures of interest $\theta_d^{(s)}$.
3. Obtain $\hat{\theta}_d^{EBP} = 1/S \sum_{s=1}^{S} \hat{\theta}_d^{(s)}$ for each region d.

Parametric Bootstrap: MSE Estimation

- Fit the random effects model to the original sample
- Generate $u_d^* \sim N(0, \hat{\sigma}_u^2)$, $e_{di}^* \sim N(0, \hat{\sigma}_e^2)$
- Construct B bootstrap populations

$$y_{di}^* = \mathbf{x}_{di}^T \hat{\boldsymbol{\beta}} + u_d^* + e_{di}^*$$

- For each b population compute the population value θ_d^{*b}
- From each bootstrap population select a bootstrap sample
- Implement the EBP with the bootstrap sample, get $\hat{\theta}_d^{*b}$

$$\widehat{MSE}(\hat{\theta}_d) = B^{-1} \sum_{b=1}^{B} (\hat{\theta}_d^{*b} - \theta_d^{*b})^2$$

- Use $\widehat{MSE}(\hat{\theta}_d)$ to compute estimated coefficients of variation (CVs)

Unit-level models: BHF HB model

Original HB model (Molina, Nandram, & Rao, 2014)

$$
\begin{aligned}
y_{di}|\boldsymbol{\beta}, u_d, \sigma_e &\sim \mathcal{N}(\mathbf{x'}_{di}\boldsymbol{\beta} + u_d, \sigma_e), \\
u_d|\sigma_u &\sim \mathcal{N}(0, \sigma_u), \\
p(\boldsymbol{\beta}, \sigma_u, \sigma_e) &= p(\boldsymbol{\beta})p(\sigma_u)p(\sigma_e) \propto p(\sigma_u)p(\sigma_e).
\end{aligned}
$$

Issues with the model:

- Sticks to the assumption of normal errors
- Flat priors on most parameters
- Parametrized to avoid MCMC

Contents

1. Overview of SAE
2. Design-Based Estimators
3. Model-Based Estimators
4. Use Case: Poverty Estimation

Poverty Indicators: FGT

$$F_d(\alpha, t) = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{t - y_{di}}{t} \right)^{\alpha} I(y_{di} \leq t), \qquad \alpha = 0, 1, 2.$$

- FGT-indicators (Foster et al., 1984)
- t is the poverty line set at 60% of median income of the state
- y_{di} is the income of the i-th person in municipality d
- $I(\cdot)$ is the indicator function
- $\alpha = 0$ is the head count ratio (HCR)
- $\alpha = 1$ is the poverty gap (PGAP)
- $\alpha = 2$ is the poverty severity
- Hard to estimate with regression only

Modified HB model

$$
\begin{aligned}
p(y^*_{di}|\boldsymbol{\beta}, u_d, \sigma_e, \nu) &= \text{Student}(\log(y_{di} + \lambda)|\mathbf{x'}_{di}\boldsymbol{\beta} + u_d,\ \sigma_e\ , \nu) \cdot \frac{1}{(y_{di} + \lambda)}, \\
u_d|\sigma_u &\sim \mathcal{N}(0, \sigma_u),\ \ d = 1, ..., D \\
\beta_0 &\sim \mathcal{N}(0, 5), \\
\beta_k &\sim \mathcal{N}(0, 0.2),\ \ k = 1, ..., K \\
\tilde{\nu} &\sim Ga(2, 0.1), \\
\nu &= \tilde{\nu} + 2, \\
\sigma_u &\sim Ga(2, 7), \\
\sigma &\sim Ga(2, 7), \\
\sigma_e &= \sigma\sqrt{\frac{\nu - 2}{\nu}}, \\
S(y^*) &\sim \mathcal{N}(0, 0.01)
\end{aligned}
$$

Algorithm 1: Estimate FGT-indicators with HB model

Input: A model $p(\theta, y)$, some data y and $\alpha \in \{0, 1, 2\}$
Output: $\hat{F}_d^{HB}, \hat{\sigma}_d^{HB}$, for $d = 1, ..., D$

for $s \in \{1, ..., S\}$ **do**
  for $d \in \{1, ..., D\}$ **do**
    $\tilde{y}_d^{(s)} | y = (\tilde{y}_{d1}^{(s)}, ..., \tilde{y}_{dN_d}^{(s)})'$;
    Sample $\hat{\boldsymbol{\beta}}^{\mathbf{(s)}}, \hat{u}_d^{(s)}, \hat{\sigma}_e^{(s)}, \hat{\sigma}_u^{(s)}, \hat{\nu}^{(s)}$ from $p(\beta, u_d, \sigma_e, \sigma_u, \nu | y)$;
    if *d is in-sample* **then**
      Sample $\tilde{y}_d^{(s)} | y$ from Student($\mathbf{x'}_d \hat{\boldsymbol{\beta}}^{\mathbf{(s)}} + \hat{u}_d^{(s)}, \hat{\sigma}_e^{(s)}, \hat{\nu}^{(s)}$)
    else
      if *d is out-of-sample* **then**
        Sample $\tilde{u}_d^{(s)}$ from $\mathcal{N}(0, \hat{\sigma}_u^{(s)})$;
        Sample $\tilde{y}_d^{(s)} | y$ from Student($\mathbf{x'}_d \hat{\boldsymbol{\beta}}^{\mathbf{(s)}} + \tilde{u}_d^{(s)}, \hat{\sigma}_e^{(s)}, \hat{\nu}^{(s)}$)
      end
    end
  end
  $\tilde{y}^{(s)} = (y_1^{(s)}, ..., y_D^{(s)})'$;
  $t^{(s)} = 0.6 \cdot median(\tilde{y}^{(s)})$;
  for $d \in \{1, ..., D\}$ **do**

$$F_d^{(s)}(\alpha, t^{(s)}) = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{t^{(s)} - \tilde{y}_{di}}{t^{(s)}} \right)^{\alpha} I(\tilde{y}_{di} \leq t^{(s)})$$

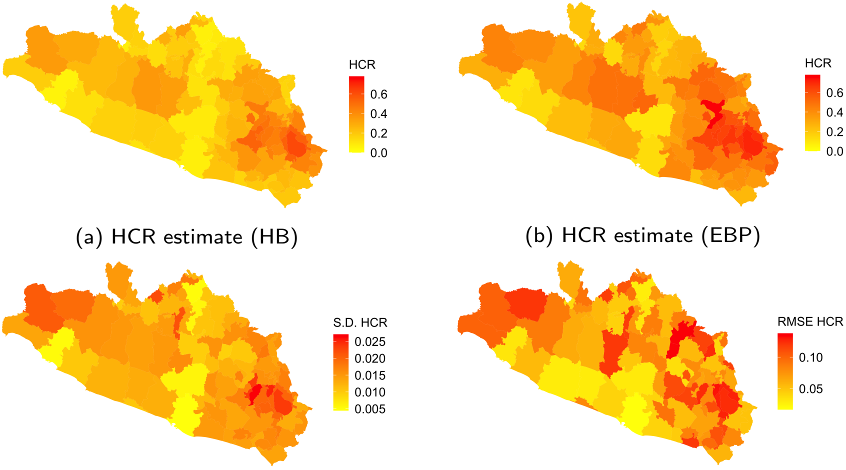
  end
end
for $d \in \{1, ..., D\}$ **do**

$$\hat{F}_d^{HB} = \frac{1}{S} \sum_{s=1}^{S} F_d^{(s)};$$

$$\hat{\sigma}_d^{HB} = \sqrt{\frac{1}{S-1} \sum_{s=1}^{S} \left(F_d^{(s)} - \hat{F}_d^{HB} \right)^2}$$

end

HCR estimates



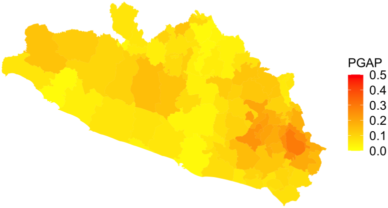
(a) HCR estimate (HB)

(b) HCR estimate (EBP)

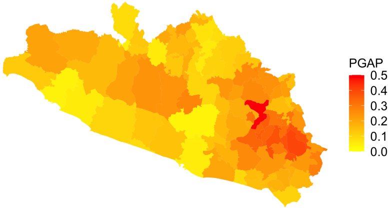
(c) HCR standard deviation (HB)

(d) HCR RMSE (EBP)

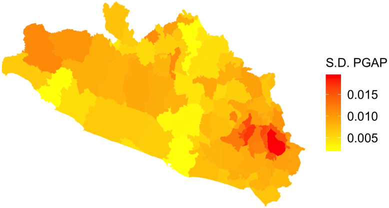
PGAP estimates



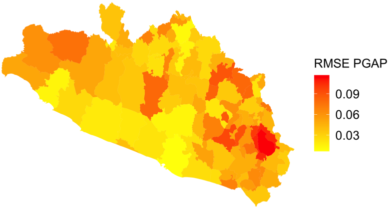
(a) PGAP estimate (HB)



(b) PGAP estimate (EBP)

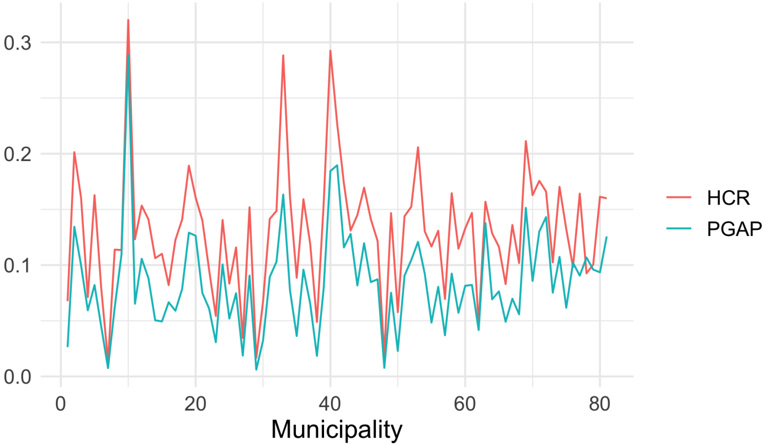


(c) PGAP standard deviation (HB)

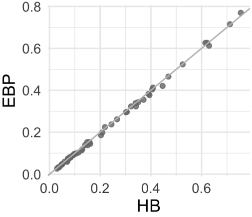


(d) PGAP RMSE (EBP)

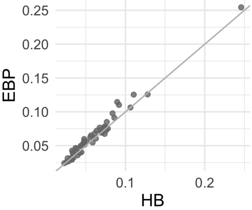
EBP vs HB estimates



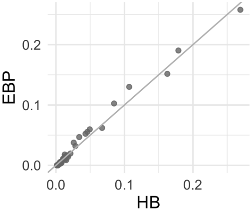
RMSE comparison HCR



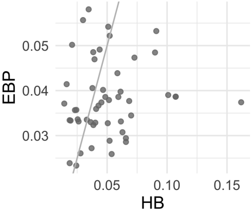
(a) HCR log-scale



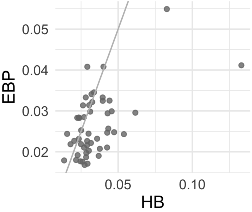
(b) HCR GB2



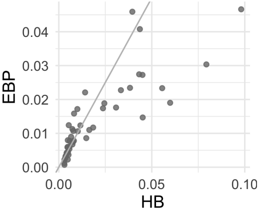
(c) HCR Pareto



(d) RMSE HCR log-scale

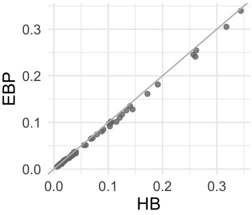


(e) RMSE HCR GB2

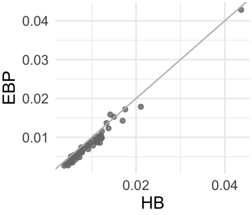


(f) RMSE HCR Pareto

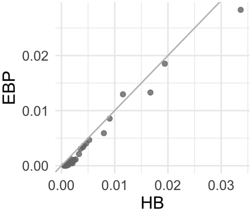
RMSE comparison PGAP



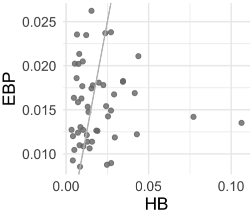
(a) PGAP log-scale



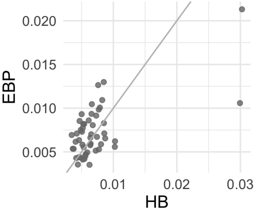
(b) PGAP GB2



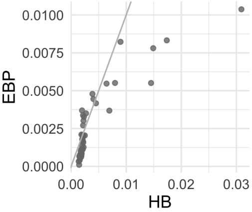
(c) PGAP Pareto



(d) RMSE PGAP log-scale



(e) RMSE PGAP GB2



(f) RMSE PGAP Pareto

Questions?

Mail: info@flaviomorelli.com

Twitter: @mexiamorelli

Bibliography I

Foster, J., Greer, J., & Thorbecke, E. (1984). A Class of Decomposable Poverty Measures. *Econometrica*, *52*(3), 761–766.

Jiang, J., & Rao, J. S. (2020). Robust Small Area Estimation: An Overview. *Annual Review of Statistics and Its Application*, *7*(1), 337–360.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators. *Journal of Statistical Software*, *91*(7).

Molina, I., Nandram, B., & Rao, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, *8*(2), 852–885.

Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation* (2nd ed.). Hoboken, NJ (USA): John Wiley & Sons, Inc.

Bibliography II

Rojas Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(1), 121–148.

Tzavidis, N., Zhang, L., 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.