# Small Areas, Benchmarking, and Political Battles: Today's Novel Demands in Small-Area Estimation

Rebecca C. Steorts

Department of Statistics
Carnegie Mellon University
joint with Malay Ghosh, Gauri Datta, and Jerry Maples

September 4, 2013

*Small area estimation is about disaggregating surveys to small noisy subgroups.*

An area $i$ is small if the sample size is not large enough to support direct estimates $\hat{\theta}_i$ of adequate precision.

- An "area" could be geographic, demographic, etc.
- Borrow strength from related areas.
- Hierarchical and Empirical Bayes methods.

*Many applications have multiple levels of resolution that call for aggregating estimates.*

- Model-based estimates for small areas often do not aggregate to the direct estimates for larger areas.
- Having model-based estimates that *do* aggregate properly is often a political necessity.

**Benchmarking**

*Benchmarking* is adjusting model-based estimates such that they aggregate to direct estimates for larger areas.

Helps deal with possible model misspecification and overshrinkage.

*Goals: Develop general class of benchmarked Bayes estimators and explore effects on the MSE.*

In Datta et al. (2011), we extend Wang et al. (2008), developing a general class of benchmarked Bayes estimators.

- No distributional assumptions.
- Linear or nonlinear estimators.
- Benchmark the weighted mean and/or weighted variability.
- Multivariate version.
- Includes many previously proposed estimators as special cases.

## Objective

Minimizing a posterior risk

$$\min_{\boldsymbol{\delta}} \sum_{i=1}^{m} \phi_i E[(\delta_i - \theta_i)^2 | \hat{\boldsymbol{\theta}}]$$

subject to the benchmarking constraint(s)

$$\sum_{i=1}^{m} w_i \delta_i = t \text{ and possibly } \sum_{i=1}^{m} w_i (\delta_i - t)^2 = h \ .$$

- Derive the benchmarked Bayes estimators $\hat{\boldsymbol{\theta}}^{BM}$ in closed form.
- $\hat{\boldsymbol{\theta}}^{BM}$ = Bayes estimator $\hat{\boldsymbol{\theta}}^{B}$ plus a correction factor.

*How does benchmarking affect the errors of the estimates?*

Using Fay-Herriot model and standard benchmarking constraint:

- Theoretically compare $MSE[\hat{\theta}^{EB}]$ and $MSE[\hat{\theta}^{EBM}]$.
  - Builds off Prasad and Rao (1990) and Wang et al. (2008); Ugarte et al. (2009).
- Derive two estimators of $MSE[\hat{\theta}^{EBM}]$ (asymptotically unbiased and parametric bootstrap).
- Evaluate methods using Small Area Income and Poverty Estimate Program (U.S. Census Bureau).

[Steorts and Ghosh (2013)]

*With m small areas, the increase in MSE due to benchmarking is* $O(m^{-1})$.

*This is shown via a second-order asymptotic expansion.*

Consider the area-level effects model of Fay and Herriot (1979):

$$\hat{\theta}_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i)$$
$$\theta_i|\boldsymbol{\beta}, \sigma_u^2 \overset{ind}{\sim} N(x_i'\boldsymbol{\beta}, \sigma_u^2), \qquad i = 1, \ldots, m.$$

Assume $D_i$ is known and $\sigma_u^2$ and $\boldsymbol{\beta}$ are unknown.

- Estimate $\sigma_u^2$ by moment estimator $\tilde{\sigma}_u^2$. Then $\hat{\sigma}_u^2 = \max\{\tilde{\sigma}_u^2, 0\}$.
- Estimate $\boldsymbol{\beta}$ by a GLS-type estimator.
- Derive the benchmarked empirical Bayes estimator $\hat{\theta}^{EBM}$.

### Theorem

$MSE[\hat{\theta}_i^{EBM}] = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) + g_4(\sigma_u^2) + o(m^{-1}),$
where

$$g_{1i}(\sigma_u^2) = \frac{D_i \sigma_u^2}{D_i + \sigma_u^2} = O(1),$$

$$g_{2i}(\sigma_u^2) \approx \text{diagonal of hat matrix } h_{ii}^V = O(m^{-1}),$$

$$g_{3i}(\sigma_u^2) \approx \text{noise in estimating } \sigma_u^2 = O(m^{-1}),$$

$$g_4(\sigma_u^2) \approx \text{avg. variance specific to each } \hat{\theta}_i = O(m^{-1}).$$

- Note: $MSE[\hat{\theta}_i^{EB}] = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) + o(m^{-1})$.
- The difference in MSEs is $g_4(\sigma_u^2)$.

We extend the method of Butar and Lahiri (2003) to derive a parametric bootstrap estimator $V_i^{\text{B-BOOT}}$ of $MSE[\hat{\theta}_i^{EBM}]$.

- Use parametric bootstrapping from Fay-Herriot model to correct plug-in estimates of $g_{1i}(\sigma_u^2), g_{2i}(\sigma_u^2)$, and $g_4(\sigma_u^2)$.
- Use the same bootstrap to estimate $g_{3i}(\sigma_u^2)$ directly.
- Combination is asymptotically unbiased:

$$E[V_i^{\text{B-BOOT}}] = MSE[\hat{\theta}_i^{\text{EBM}}] + o(m^{-1}).$$

*How does benchmarking perform in applications?*

- Small Area Income and Poverty Estimates (SAIPE) program (U.S. Census Bureau): model-based estimates of the number of poor children (aged 5–17).
- Model-based state estimates were benchmarked to a direct estimate of national child poverty by raking.
- Direct estimates came from from the Annual Social and Economic (ASEC) Supplement of the Current Population Survey (CPS) and the American Community Survey (ACS).
- Weights $w_i \propto$ estimated number of children in each state.

Recall the model of Fay and Herriot (1979):

$$\hat{\theta}_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i)$$
$$\theta_i|\beta, \sigma_u^2 \overset{ind}{\sim} N(x_i'\boldsymbol{\beta}, \sigma_u^2), \qquad i = 1, \ldots, m$$

- where $D_i > 0$ are known,
- $\theta_i$ are the true state level poverty rates,
- $\hat{\theta}_i$ are the direct state estimates.

Employ EB on unknown $\boldsymbol{\beta}$ and $\sigma_u^2$.

- We consider data from 1997 and 2000.
- The data from 2000 behaves as our theory indicates: $\text{MSE}[\hat{\theta}^{EBM}]$ are slightly larger than $\text{MSE}[\hat{\theta}^{EB}]$.
- The same is true when we bootstrap.

Table: Table of estimates for 1997

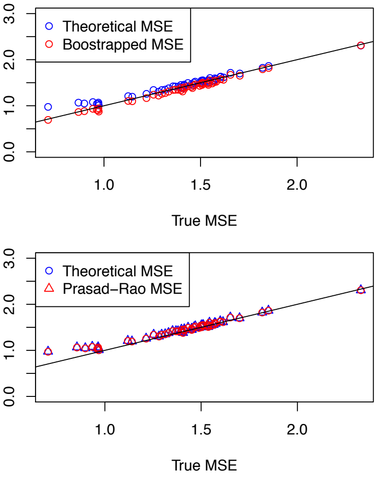
| | Estimates | | | MSEs | | | Bootstrap | |
|---|---|---|---|---|---|---|---|---|
| $i$ | $\hat{\theta}_i$ | $\hat{\theta}_i^{EB}$ | $\hat{\theta}_i^{EBM1}$ | $\hat{\theta}_i$ | $\hat{\theta}_i^{EB}$ | $\hat{\theta}_i^{EBM1}$ | $\hat{\theta}_i^{EB}$ | $\hat{\theta}_i^{EBM1}$ |
| 12 | 18.98 | 13.72 | 13.89 | 20.87 | 2.45 | 2.48 | 1.24 | 1.26 |
| 13 | 17.56 | 13.64 | 13.82 | 12.38 | 1.70 | 1.73 | 0.23 | 0.25 |
| 14 | 14.57 | 15.72 | 15.89 | 3.56 | 3.45 | 3.47 | −0.06 | −0.05 |
| 15 | 11.07 | 12.53 | 12.70 | 7.58 | 1.84 | 1.86 | −0.23 | −0.22 |
| 16 | 11.09 | 11.21 | 11.38 | 8.49 | 1.74 | 1.76 | −0.24 | −0.22 |
| 17 | 11.01 | 13.48 | 13.65 | 9.34 | 1.61 | 1.63 | −0.15 | −0.14 |
| 18 | 23.12 | 20.78 | 20.95 | 13.98 | 1.37 | 1.40 | −0.12 | −0.11 |
| 19 | 21.08 | 24.15 | 24.32 | 15.19 | 1.80 | 1.82 | 0.40 | 0.42 |
| 20 | 13.18 | 12.44 | 12.61 | 13.63 | 2.09 | 2.11 | 0.56 | 0.57 |
| 21 | 9.90 | 13.16 | 13.33 | 9.28 | 1.65 | 1.67 | −0.03 | −0.01 |
| 22 | 19.66 | 14.38 | 14.56 | 7.66 | 2.46 | 2.48 | 1.02 | 1.04 |
| 23 | 13.78 | 16.86 | 17.03 | 4.04 | 3.11 | 3.13 | 0.38 | 0.39 |

- Strange behavior for 1997; problem occurs when $\hat{\sigma}_u^2$ is 0.
- Note that

$$V_i^{\text{B-BOOT}} = g_{1i}(\hat{\sigma}_u^2) + \{g_{1i}(\hat{\sigma}_u^2) - E_*[g_{1i}(\hat{\sigma}_u^{*2})]\} + O(m^{-1}).$$

- $g_{1i}(\hat{\sigma}_u^2) = D_i\hat{\sigma}_u^2(D_i + \hat{\sigma}_u^2)^{-1} = O(1)$.
  - For 1997 dataset this term is 0.
  - This causes many of the bootstrap estimates of the MSE of the benchmarked estimators to be negative.
- Theoretical (asymptotic) MSE escapes problem since

$$P(\tilde{\sigma}_u^2 \leq 0) = O(m^{-r}) \; \forall \; r > 0.$$




Simulation study for 1997

- Unified framework for one-stage benchmarking.
- The increase in MSE due to benchmarking is negligible.
- Derived two estimators of our MSE (asymptotically unbiased and parametric bootstrap).
- Recommend use of estimator of the MSE of the benchmarked EB estimator.
  - Fast calculation.
  - Parametric bootstrap yields undesirable results.

- Spatial and temporal smoothing for SAE and benchmarking.
- Application to high dimensional dataset (both in covariates and parameter space) and more standard applications in SAE.
- Comparing to frequentists benchmarks under MSE comparisons (under bootstrapping).
- Validations under CV and model-checking.

Questions: beka@cmu.edu

Thank you to Malay Ghosh: mentor, inspiration, and friend.

This research has been supported by the U.S. Census Bureau Dissertation Fellowship Program and the NSF. The views expressed reflect those of the authors and not of the United States Census Bureau or NSF.

Butar, F. and Lahiri, P. (2003). On measures of uncertainty of empirical bayes small area estimators. *J. Statist. Plann. Inference*, **112** 63–76.

Datta, G., Ghosh, M., Steorts, R. and Maples, J. (2011). Bayesian benchmarking with applications to small area estimation. *TEST*, **20** 574–588.

Fay, R. and Herriot, R. (1979). Estimates of income from small places: an application of James-Stein procedures to census data. *Journal of the American Stastical Association*, **74** 269–277.

Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Stastical Association*, **85** 163–171.

Rao, J. (2003). *Small Area Estimation*. Wiley, New York.

Ugarte, M., Goicoa, T. and Militino, A. (2009). Benchmarked estimates in small areas using linear mixed models with restrictions. *TEST*, **18** 342–364.

WANG, J., FULLER, W. and QU, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, **34** 29–36.

We benchmark a weighted mean or both a weighted mean and variability.

- $\hat{\theta}_1, \ldots, \hat{\theta}_m =$ direct estimators of the $m$ small area means $\theta_1, \ldots, \theta_m$.
- Find the benchmarked Bayes estimator

$$\hat{\boldsymbol{\theta}}^{BM1} = (\hat{\theta}_1^{BM1}, \ldots, \hat{\theta}_m^{BM1})$$

  of $\boldsymbol{\theta}$ such that $\sum_{i=1}^m w_i \hat{\theta}_i^{BM1} = t$, where $t$ is prespecified from some other source or $t = \sum_{i=1}^m w_i \hat{\theta}_i$.
- The $w_i$ are known weights, where $\sum_{i=1}^m w_i = 1$.

- Goal:

$$\min_{\boldsymbol{\delta}} \sum_{i=1}^{m} \phi_i E[(\delta_i - \theta_i)^2 | \hat{\boldsymbol{\theta}}]$$

  such that the $\delta_i$'s satisfy $\bar{\delta}_w = \sum_{i=1}^m w_i \delta_i = t$.
- $\hat{\theta}_i^B$ = posterior mean of $\theta_i$ under a particular prior.
- $\bar{\hat{\theta}}_w^B = \sum_{i=1}^m w_i \hat{\theta}_i^B$.
- $r = (r_1, \ldots, r_m)'$ where $r_i = w_i/\phi_i$, and define $s = \sum_{i=1}^m w_i^2/\phi_i$.

## Theorem 1

$$\hat{\boldsymbol{\theta}}^{BM1} = \hat{\boldsymbol{\theta}}^{B} + s^{-1}(t - \bar{\hat{\theta}}_w^B)r.$$

minimizes $\sum_{i=1}^{m} \phi_i E[(\delta_i - \theta_i)^2 | \hat{\boldsymbol{\theta}}]$ subject to $\bar{\delta}_w = t$.

(The theorem extends to a multivariate setting)

- We can also benchmark using (i) $\sum_i w_i \hat{\theta}_i^{BM2} = t$ and (ii) $\sum_i w_i (\hat{\theta}_i^{BM2} - t)^2 = H$, where H is defined below. Maybe we think our estimates are too close together, for example.
- This can be extended to a multivariate setting.

## Theorem 2

Subject to (i) and (ii), the benchmarked Bayes estimators of $\theta_i$ are given by

$$\hat{\theta}_i^{BM2} = \hat{\theta}_i^B + (t - \bar{\hat{\theta}}_w^B) + (a_{CB} - 1)(\hat{\theta}_i^B - \bar{\hat{\theta}}_w^B),$$

where $a_{CB} = H / \sum_{i=1}^m w_i (\hat{\theta}_i^B - \bar{\hat{\theta}}_w^B)^2$. Note that $a_{CB} \geq 1$ when $H = \sum_{i=1}^m w_i E[(\theta_i - \bar{\theta}_w)^2 | \hat{\boldsymbol{\theta}}]$.

Consider the area-level effects model of Fay and Herriot (1979):

$$\hat{\theta}_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i)$$
$$\theta_i|\beta, \sigma_u^2 \overset{ind}{\sim} N(x_i'\boldsymbol{\beta}, \sigma_u^2), \qquad i = 1, \ldots, m$$

Assume $D_i$ is known and $\sigma_u^2$ and $\beta$ are unknown.

- Estimate $\sigma_u^2$ by moment estimator $\tilde{\sigma}_u^2$. Then $\hat{\sigma}_u^2 = \max\{\tilde{\sigma}_u^2, 0\}$.
- We estimate $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}\hat{\boldsymbol{\theta}}$, where $V = \text{Diag}\{\sigma_u^2 + D_1, \ldots, \sigma_u^2 + D_m\}$.
- Benchmarked empirical Bayes estimator derived by Datta et al. (2011) is $\hat{\theta}^{EBM1} = \hat{\theta}_i^{EB} + (\bar{\hat{\theta}}_w - \bar{\hat{\theta}}_w^{EB})$.
- $\hat{\theta}_i^{EB} = (1 - \hat{B}_i)\hat{\theta}_i + \hat{B}_i x_i'\tilde{\boldsymbol{\beta}}(\hat{\sigma}_u^2)$, where $\hat{B}_i = D_i(\hat{\sigma}_u^2 + D_i)^{-1}$.

Define $h_{ij}^V = x_i'(X'V^{-1}X)^{-1}x_j$. Under some mild regularity conditions, we can find a second-order approximation of the MSE of the benchmarked empirical Bayes estimator.

## Theorem 4

$E[(\hat{\theta}_i^{EBM1} - \theta_i)^2] = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) + g_4(\sigma_u^2) + o(m^{-1})$, where

$$g_{1i}(\sigma_u^2) = B_i\sigma_u^2, \quad g_{2i}(\sigma_u^2) = B_i^2 h_{ii}^V,$$

$$g_{3i}(\sigma_u^2) = B_i^3 \text{Var}(\tilde{\sigma}_u^2),$$

$$g_4(\sigma_u^2) = \sum_{i=1}^m w_i^2 B_i^2 V_i - \sum_{i=1}^m \sum_{j=1}^m w_i w_j B_i B_j h_{ij}^V, \text{ and}$$

$$\text{Var}(\tilde{\sigma}_u^2) = 2(m-p)^{-2} \sum_{k=1}^m (\sigma_u^2 + D_k)^2 + o(m^{-1}).$$

We use the methods of Butar and Lahiri (2003) and use the following bootstrap model:

$$\hat{\theta}_i^* | u_i^* \overset{ind}{\sim} N(x_i'\beta + u_i^*, D_i)$$
$$u_i^* \overset{ind}{\sim} N(0, \hat{\sigma}_u^2).$$

We use the parametric bootstrap twice. We first use it to estimate $g_{1i}(\sigma_u^2), g_{2i}(\sigma_u^2)$, and $g_4(\sigma_u^2)$. We then use it to estimate $E[(\hat{\theta}_i^{EB} - \hat{\theta}_i^{B})^2] = g_{3i}(\sigma_u^2) + o(m^{-1})$.

Our proposed estimate of $MSE[\hat{\theta}_i^{EBM1}]$ is

$$\begin{aligned} V_i^{\text{B-BOOT}} &= 2[g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + g_4(\hat{\sigma}_u^2)] \\ &- E_* \left\{ g_{1i}(\hat{\sigma}_u^{*2}) + g_{2i}(\hat{\sigma}_u^{*2}) + g_4(\hat{\sigma}_u^{*2}) \right\} \\ &+ E_*[(\hat{\theta}_i^{EB*} - \hat{\theta}_i^{EB})^2]. \end{aligned}$$

- Our estimate $\hat{\sigma}_u^{*2}$ is the estimate of $\sigma_u^2$ that is calculated using the $\hat{\theta}_i^*$ values.
- Note that $\hat{\theta}_i^{EB*}$ is calculated using $\hat{\sigma}_u^{*2}$ and $\hat{\theta}_i$ (not $\hat{\theta}_i^*$).

We extend the methodology of Butar and Lahiri (2003) to find a parametric bootstrap estimator of the MSE of the benchmarked EB estimator. Then we can show

**Theorem 6**

$$E[V_i^{\text{B-BOOT}}] = MSE[\hat{\theta}_i^{\text{EBM1}}] + o(m^{-1}).$$