"Small Area Estimation: It's Evolution in Five Decades" by Malay Ghosh

DISCUSSION

J. N. K. Rao

Carleton University, Ottawa, Canada

28th Annual Morris Hansen Lecture

October 30, 2019, Washington DC

Inference from survey data: Hansen

- **Design unbiasedness not insisted upon because "it often results in much larger MSE than necessary". Instead, design consistency deemed necessary for large samples.**

- **Model dependent strategies perform poorly in large samples even under small model misspecifications.**

- **Substantial advantage in small samples if model is appropriate. Sampling plan need not be a probability sampling plan. Relax the model by including additional variables but may not be adequate.**

Motivation for SAE

- **Demand for reliable local or small area statistics has greatly increased. Direct area-specific estimates are inadequate due to small domain sample sizes or even zero sample sizes.**

- **Necessary to "borrow strength" across related areas through linking models.**

- **Opposition to models has been overcome by the demand for small area estimation (Kalton 2018).**

Basic area-level model

- **Notation:** m areas out of M are sampled. Associated parameters θ_i and direct estimators $\hat{\theta}_i, i = 1,\ldots,m$.

- **Sampling model:** $\hat{\theta}_i = \theta_i + e_i$ with $e_i \sim_{ind} N(0,\psi_i)$ and known sampling variance ψ_i ($i = 1,\ldots,m$).

- **Matched linking model:** $\theta_i = z_i'\beta + v_i$ with $v_i \sim_{iid} N(0,\sigma_v^2)$ and area-level covariates z_i.

- Fay and Herriot (1979): $\theta_i = \log(\bar{Y}_i)$ with mean income $\bar{Y}_i$

- For sampled areas, empirical best (EB) estimator of θ_i is given by $\hat{\theta}_i^{EB} = \tilde{\theta}_i^{B}(\hat{\beta}, \hat{\sigma}_v^2)$, where $\tilde{\theta}_i^{B}(\beta, \sigma_v^2) = \gamma_i \hat{\theta}_i + (1-\gamma_i)(z_i' \beta)$ is the best estimator, $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$ and ($\hat{\beta}, \hat{\sigma}_v^2$) estimators of model parameters (β, σ_v^2): REML or FH moment estimators.

- For non-sampled areas, use synthetic estimator $\hat{\theta}_i^{S} = z_i' \hat{\beta}$

- Tacitly assumed that the population linking model holds for sampled and non-sampled areas separately: non-informative sampling of areas. Most of the literature assumes all areas are sampled: $m = M$.

Demonstrating merits of model-based SAE

- $MSE(\hat{\theta}_i^{EB}) \approx g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2)$. Leading term $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$ is much smaller than ψ_i , the variance of $\hat{\theta}_i$, if γ_i is small. Second term is due to estimating β and third term due to estimating σ_v^2

- Design MSE of EB estimator is not necessarily smaller than the variance of $\hat{\theta}_i$ for **every** area. Some averaging of MSEs needed (James-Stein 1961).

- External evaluation (Canadian experience): Census areas (CAs) are small areas. Direct estimate is unemployment rate from LFS and area-level covariate is EI beneficiary rate. Much larger survey (NHS) estimates treated as gold standard or true values (Hidiroglou et al. 2019)

- Average absolute relative error (ARE) over all areas: LFS direct estimates give 33.9% while EB estimates give 14.7%.

- For the 28 smallest areas reduction in ARE more pronounced: LFS give 70.4% and EB give 17.7%.

MSE estimation

- **Model MSE estimator** (Prasad and Rao, 1990)

  $mse_{PR}(\hat{\theta}_i^{EB}) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2)$.

- **Pfeffermann (2017): National Statistical Agencies prefer estimates of design MSE of EB, similar to design MSE estimate of $\hat{\theta}_i^{EB}$, conditional on $\theta = (\theta_1, \ldots, \theta_m)'$.**

- **Exact design-unbiased MSE estimator** can be highly unstable and can take negative values often when ψ_i is large relative to σ_v^2 (Datta et al. 2011)

- **Composite** MSE estimator based on mse_d and mse_{PR}:

$$mse_c(\hat{\theta}_i^{EB}) = \hat{\gamma}_i mse_d(\hat{\theta}_i^{EB}) + (1-\hat{\gamma}_i) mse_{PR}(\hat{\theta}_i^{EB})$$

  Alternative MSE estimator uses $\sqrt{\hat{\gamma}_i}$ and $1-\sqrt{\hat{\gamma}_i}$: More weight to mse_d.

Simulation study (Rao et al. 2019)

- $m = 30$ areas divided into five groups each of size six with equal ψ_i values: 2.0,0.6,0.5,0.4,0.2 and $\sigma_v^2 = 1$.

Simulation results

- Average probability of getting negative mse_d is large (46%) for group 1 with large sampling variance. Modification leads to large ARB (94% for group 1). Probability is zero for mse_c across all areas.

- For group 1, ARB of mse_c is smaller relative to mse_{PR} at the expense of increase in RRMSE. For other areas, ARB of mse_{PR} persists unlike mse_c and RRMSE values are similar.

- Serious under-coverage rates for group 1.

MSE estimation after preliminary model testing

- Test $H_0: \sigma_v^2 = 0$ at level α. For small m, Datta et al. (2011) proposed PT estimator : Use synthetic estimator $z_i'\hat{\beta}_{PT}$ if H_0 is not rejected and retain $\hat{\theta}_i^{EB}$ otherwise. In the PT literature, $\alpha = 0.2$ is recommended. In this case, MSE of PT and EB estimators practically the same.

- Molina et al. (2015): Use $mse_{PT}(\hat{\theta}_i^{EB}) = g_{2i}(0)$ if H_0 not rejected **or** $\hat{\sigma}_v^2 = 0$, and PR MSE estimator if H_0 rejected **and** $\hat{\sigma}_v^2 > 0$. Performed well in simulations in terms of RB. Avoid zero $\hat{\sigma}_v^2$ (Yoshimori and Lahiri 2014): AML.

Misspecified linking model

- Best estimator of area mean under “working” FH linking model. Only sampling model assumed to be correct.

- Minimizing estimator of total design MSE of best estimators w.r.t. β and σ_v^2 gives best predictive estimators (BPE) of β and σ_v^2. Resulting EB estimator is observed best predictor (OBP). Performed well under linking model misspecification (Jiang et al. 2011)

- MSE estimation of OBP (Chen et al. 2019): One-bring-one-Route (OBOR)

Unmatched or mismatched models

- Linking model: $h(\theta_i) = z_i'\beta + v_i$ with specified function $h(.)$ and sampling model $\hat{\theta}_i = \theta_i + e_i$, where $\hat{\theta}_i$ is unbiased or approximately unbiased. Sugasawa et al. (2018): EB estimation and associated MSE estimation.

- HB estimation under unknown link function $h(.)$ using P-spline mixed model formulation (Sugasawa et al., 2018).

Big data as covariates

- Marchetti et al. (2015): GPS data on car mobility used to create mobility index related to poverty rate and household income in Italy. Advantage: GPS data also available for non-sampled local areas.

- Schmidt et al. (2017): Mobile phone data as covariate to estimate literacy level at the commune level in Senegal. Direct estimates obtained from a probability sample.

Two-fold area level models

- Sampling model $\hat{\theta}_{ij} = \theta_{ij} + e_{ij}$ for sampled sub-area j within area i. Torabi and Rao (2014) studied EB estimation of area means and sub-area means under matched linking model: $\theta_{ij} = z_{ij}'\beta + v_i + e_{ij}$. Advantage: Efficient estimators for non-sampled sub- areas.

- Erciulescu et al. (2017) used HB for county crop estimation satisfying benchmarking.

- PIAAC project of Westat: Three-fold area level model using HB.
- Mohadjer et al. (2012) extended the two-fold matched model to unmatched case using HB to get county-level adult literacy estimates using NAAL data.
- Cai et al. (2019): EB estimation for two-fold unmatched model.

Unit level models

- **Basic unit level model:** $y_{ij} = x'_{ij}\beta + v_i + e_{ij}$ **with** $v_i \sim_{iid} N(0,\sigma_v^2)$ **and independent of** $e_{ij} \sim N(0,\sigma_e^2)$**, see Rao and Molina (2015, ch. 7) for estimation of area means and MSE estimation.**

- **Robust estimation using semi-parametric spline models (Rao, Sinha and Dumitrescu, 2013).**

- **Bias-corrected outlier robust estimators and associated MSE estimation (Chambers et al. 2014).**

Informative sampling within areas

- Model design weights within areas and develop bias adjusted EB estimator (Pfeffermann and Sverchkov 2011). Extends to sampling of areas.

- Augmented unit level models with specified function of within area selection probability as augmenting variable and $m = M$ (Verret et al. 2015).

- Augmented unit level models with unspecified function approximated by P-spline (Cai et al. 2017).

SAE using record linkage with big data

- Unit level covariates x_{ij} obtained from external source and matched to sample y_{ij}. Estimation under linkage errors studied by Han and Lahiri (2017) and Chambers et al. (2019), assuming non-informative sampling within areas.

Regression tree methods for SAE

- Lohr (2008) and Toth and McConville (2019)

Some extensions

- **Estimation of complex small area parameters: Poverty indicators using EB or HB estimation**

- **Bivariate area level models**

- **Time series models and spatio-temporal models**

- **SAE estimation after model selection**

Multilevel Regression and Poststratifcation (MRP)

- Find vector of variables X that affect the sample design, nonresponse and coverage (Gelman team).

- **Assumption:** Given X, the distribution of inclusion indicator is ignorable. Discretize the variables and cross classify to for a very large number of post strata and sampling within poststrata is SRS. Most poststrata are empty.

- Bayes estimates of poststrata means are obtained assuming a multilevel model and known poststrata counts. Small areas are unions of poststrata.

Production of small area official statistics

- “From start to finish: a framework for the production of small area official statistics” (Tzavidis et al. 2019). Parsimony and evaluation. Model-dependent methods with focus on model selection and testing, model diagnostics. Application to estimation of non-linear deprivation indicators.

- Molina and Marhuenda (2015): R package for SAE used in the book by Rao and Molina (2015).

- Software for HB: Erciulescu (2019) and Chen et al. (2019)