Inferential Issues in Small Area Estimation (SAE): Some History and Selective Recent Developments

J. N. K. Rao

Carleton University, Ottawa, Canada

International Association of Survey Statisticians (IASS)

Webinar, April 30, 2025

Pfeffermann (2013)

- " In 2002, I published a review paper with a similar title. In that year, small area estimation (SAE) was flourishing both in research and applications, but my own feeling then was that the topic has been more or less exhausted in terms of research and that it will just turn into a routine application in sample survey practice. As the past 9 years show, I was completely wrong; not only is the research in this area accelerating, but it now involves some of the best- known statisticians, who otherwise are not involved in survey sampling theory. The diversity of new problems investigated is overwhelming, and the solutions proposed are not only elegant and innovative, but also very practical."

Some SAE history

- National Institute on Drug Abuse (1979). Synthetic estimation.
- International Symposia on SAE: Ottawa, Canada (1985), Warsaw, Poland (1993), Riga, Latvia (1999).
- Series of international SAE conferences starting in 2005: Jyvaskyla, Finland (2005), Pisa, Italy (2007), Elche, Spain (2009), Trier, Germany (2011), Bangkok , Thailand (2013), Poznan, Poland (2014), Santiago, Chile (2015), Maastricht, Netherlands I2016), Paris, France (2017), Shanghai, China (2018), Singapore (2019), Naples, Italy (2021), Washington, DC, USA (2022), Lima, Peru (2024).
- Rhine River Cruise (2009) to celebrate 30th anniversary of Fay-Herriot (1979) paper (organized by Ralf Munnich and Partha Lahiri).

SAE books/monographs

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation, 2nd ed*. (3911 GS citations, two editions together).

Morales, D., Esteban, M. D., Perez, A. and Hobza, T. (2021). *A Course on Small Area Estimation and Mixed Models*. Springer.

Jiang, J. (2017). *Asymptotic Analysis of Mixed Effects Models. Theory, Applications and Open Problems*. CRC Press.

Jiang, J. and Rao, J. S. (2025). *Robust Small Area Estimation: Methods, Theory, Applications and Open Problems*. CRC Press (to appear).

Pratesi, M. (2016) (Editor). *Analysis of Poverty Data by Small Area Estimation, Wiley*.

Sugasawa, S. and Kubokawa, T. (2023). *Mixed Effects Models and Small Area Estimation*. Springer.

Some SAE review papers

Ghosh, M. and Rao, J. N. K. (1994). *Statistical Science,* 55-93 (1306 GS citations).

Jiang, J. and Lahiri, P. (2006). *Test*, 1-96.

Datta, G. S. (2009). *Handbook of Statistics, vol. 29B, North-Holland*, 251-288.

Pfeffermann, D. (2013). *Statistical Science*, 40-68 (GS citations).

Sugasawa, S. and Kubokawa, T. (2020). *Japanese Journal of Statistics and Data Science*.

Ghosh, M. (2020). *Statistics in Transition*, 1-22 (Morris Hansen Lecture).

Molina, I. and Rao, J. N. K. (2023). *Survey Statistician,* 21-24.

Rico. F. O., Pinerez, C. F. and R-Vargas (2024). *Revista Columbiana de Estadistica- Applied Statistics,* 407-422 (Review of SAE in Columbia).

Basic parametric Fay-Herriot (FH) area-level model

- **Notation: m areas out of M sampled. Associated direct estimators $\hat{\theta}_i = g(\bar{y}_{iw})$ of the parameters $\theta_i = g(\bar{Y}_i)$. FH use $\theta_i = \log(\bar{Y}_i)$ and $\bar{Y}_i$ is mean income. We focus on the special case $\theta_i = \bar{Y}_i$.**

- **Linking model: $\theta_i = z_i'\beta + v_i, v_i \sim_{iid} N(0, \sigma_v^2)$.**

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

- **Advantages: (1) Takes account of survey design through direct estimators. (2) Only area-level covariates are needed.**

Model-based estimation of θ_i

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

- For sampled areas, empirical best (EB) estimator under normality: $\hat{\theta}_i^{EB} = \tilde{\theta}_i^B(\hat{\beta}, \hat{\sigma}_v^2)$, where $\tilde{\theta}_i^B(\beta, \sigma_v^2) = \gamma_i \hat{\theta}_i + (1-\gamma_i) z_i'\beta$ with $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$ and $(\hat{\beta}, \hat{\sigma}_v^2)$ are estimators of model parameters: ML, REML, AML. Moment estimators of Prasad and Rao (1990, 1158 GS citations) and Fay and Herriot (1979) do not need normality. EB estimator is identical to EBLUP of Henderson, $\hat{\theta}_i^H$, without normality assumption.

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

Second-order unbiased MSE estimation

- **Using the Prasad and Rao (PR) moments estimator of σ_v^2 and normality:**

$$mse_{PR}(\hat{\theta}_i^H) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2),$$

where $g_{1i}(\sigma_v^2) = \gamma_i \psi_i \ll \psi_i$ when γ_i is small. This MSE estimator is based on Taylor linearization. Similar MSE estimators under ML, REML, AML and FH methods of estimating σ_v^2. The last two terms are due to estimating β and σ_v^2 respectively.

- **Jackknife MSE estimator (Jiang, Lahiri and Wan, 2002): Applicable under normality to all methods of estimating model parameters.**

- **(1) Hybrid bootstrap (Butar and Lahiri, 2003) and (2) Double bootstrap under normality. (1) is computationally simpler than (2).**

MSE estimation under a semi-parametric FH model

- We consider two semi-parametric FH area-level models: (1) Distribution of v_i unspecified but the distribution of e_i remains normal. (2) Both distributions are unspecified.

- We consider second -order correct MSE estimation of EBLUP based on PR and FH moment estimators of σ_v^2.

- Using PR moment estimator for case 1, Lahiri and Rao (1995) proved that the normality-based PR MSE estimator remains second order correct, thus establishing some robustness. This however is not true under the FH moment estimator (Chen, Lahiri and Rao 2025), using the corresponding normality-based MSE estimator.

Semi-parametric FH model under case 2

- We assume that the kurtosis of e_i is non-zero and known or obtained from smoothing, similar to smoothed estimator of ψ_i.

- We use Poisson sampling to estimate both the variance and fourth moment of e_i. Resulting estimated kurtosis values are smoothed using a method similar to GVF.

- For case 2, a second order correct MSE estimator is derived. It is shown that it depends on the kurtosis of e_i, but does not depend on the kurtosis of the random area effects v_i.

Linking model unspecified: OBP

- Linking model: $\theta_i = \mu_i + v_i$ and the form of μ_i is unknown.
- Let $\hat{\theta}_i^B$ be the best estimator under working FH model with $\mu_i = z_i'\beta$. It is a function of (β, σ_v^2). Minimize an estimator of total design MSE $E_p \sum_{i=1}^{m} (\hat{\theta}_i^B - \theta_i)^2$. Resulting estimator of β is called best predictive estimator (BPE). It is a WLS estimator with weights $w_i^{bpe} \propto [\psi_i / (\sigma_v^2 + \psi_i)]^2$ in contrast to $w_i^{mle} \propto 1/(\sigma_v^2 + \psi_i)$ under the working FH model. Substituting BPE of (β, σ_v^2) into $\hat{\theta}_i^B$ leads to observed best predictor (OBP) of θ_i (Jiang, Nguyen and Rao, 2011). Compromise weights proposed by Henderson, Varadhan and Louis (2023).
- PR estimator of MSE as estimator of MSE of OBP is robust in the sense it remains first-order unbiased under model misspecification (Liu, Ma and Jiang, 2022).

Random forests (RF) area-level model

- Machine learning (ML) methods are now used for non-parametric regression, in particular RF. Beaumont, Bocci, Bosa and Sambo (2024) applied RF to the FH model with mean function $\mu_i = m(z_i)$ for unspecified $m(.)$.

- RF regression of $\hat{\theta}_i$ on z_i leads to the estimator $\hat{\mu}_i^{RF}$. Similarly, RF smoothed estimators of ψ_i are obtained using RF regression of direct estimators $\log(\hat{\psi}_i)$ on some other covariates x_i, leading to smoothed estimators $\hat{\psi}_i^{RF}$ treated as true ψ_i. RF requires large m.

- Now apply the standard FH estimation to $(\hat{\theta}_i, \hat{\mu}_i^{RF}; i = 1, \ldots, m)$, ignoring the measurement error in $\hat{\mu}_i^{RF}$, to get EB estimator of θ_i.

Hierarchical Bayes (HB) method

- HB approach is straightforward, although it can be complex. It provides inferences that are clear-cut and "exact". Requires specification of a subjective prior $f(\beta, \sigma_v^2)$. Assume prior independence and $f(\beta) \propto 1$. Different choices of prior on σ_v^2 are available, including matching priors. Under this set up, samples are generated from the posterior distribution of $(\theta_1, \ldots, \theta_m)$.

- Nandram, Cruze and Erciulescu (2023) studied inequality constraints $c_i \leq \theta_i, i = 1, \ldots, m$ and benchmarking constraint $\sum_{i=1}^{m} \theta_i < a$ for specified c_i and a. Ghosh, Ghosh, Maples and Tang (2022) assumed $v_i \sim_{ind} N(0, \lambda_i^2 \sigma_v^2)$ and specified priors on the global parameter σ_v^2 and local parameter λ_i^2.

Extensions of FH model

- Bivariate FH models (Fay, 1987).
- Time series and cross-sectional models (Fay and Diallo, 2018).
- Spatial FH models (Pratesi, Marchetti, Giusti and Salvati, 2023; Torabi and Jiang, 2020; Chung and Datta. 2022).
- Use of big data (for example GPS data: Marchetti et al. ,2015).
- Binomial or multinomial logistic linear mixed models, using effective sample sizes and HB (Gutierrez, 2024).
- Two-fold subarea models using HB approach (Erciulescu, Cruze and Nandram, 2019) with application to county crop estimation.

Battese-Harter-Fuller (BHF) unit-level model

- **Requires unit-level sample data** $\{(y_{ij}, x_{ij}),\ j = 1,\ldots,n_i; i = 1,,,.m\}$. **Area population means** $\bar{X}_i$ **assumed known.**

- **Population model assumed to hold for the sample:**

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij}; v_i \sim_{iid} N(0,\sigma_v^2), e_{ij} \sim_{iid} N(0,\sigma_e^2)$$

- **EBLUP (without normality) and EB of area mean** $\bar{Y}_i \approx \bar{X}'_i\beta + v_i$ **are equivalent: Weighted combination of sample regression estimator and regression synthetic estimator with weights** $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i)$ **and** $1 - \hat{\gamma}_i$.

MSE estimation

- **Second-order correct MSE estimators: PR, jackknife, double bootstrap (without normality) for EBLUP of area mean.**

- **Basic unit level model is a special case of general linear mixed model. Second-order MSE estimation proposed by Das, Jiang and Rao (2004) received considerable attention (347 GS citations).**

Comparison of FH and BHF under unit level data

- Pseudo-EBLUP (You and Rao 2022) under unit level population model takes account of survey weights. It uses an aggregated model which may be written in terms of a sample regression estimator (SREG) $\tilde{y}_{iw} = \bar{y}_{iw} + (\bar{X}_i - \bar{x}_{iw})'\beta$ as

$$\tilde{y}_{iw} = \bar{X}_i'\beta + v_i + \bar{e}_{iw}, \bar{e}_{iw} \sim N(0, \sigma_e^2 \delta_i), \delta_i = \sum_j w_{ij}^2 / (\sum_j w_{ij})^2$$

- Fay (2018) suggested using SREG $\bar{y}_{iw}^{SR} = \bar{y}_{iw} + (\bar{X}_i - \bar{x}_{iw})'\hat{B}_w$ as the direct estimator in the FH model. This leads to modified FH model

$$\bar{y}_{iw}^{SR} = \bar{X}_i'\beta + v_i + \bar{e}_{iw}^{SR}, \bar{e}_{iw}^{SR} \sim N(0, \psi_i^{SR}).$$

- Two methods lead to similar results in terms of MSE. FH estimator using $\bar{y}_{iw}$ as direct estimator performed poorly relative to pseudo-EBLUP in terms of MSE for small m (Hidiroglou and You, 2016).

Estimation of Complex Parameters

- Parameters of interest: $\theta_i = h(y_{i1},...,y_{iN_i})$. Special case: Separable functions $\theta_i = \sum_{j=1}^{N_i} h_1(y_{ij})$. FGT poverty indictors, used by the World Bank (WB), are separable functions: Poverty rate, poverty gap and poverty severity. Here y_{ij} is a suitable one-to-one function of a welfare variable E_{ij}; for example, $y_{ij} = \log(E_{ij}).$
- Survey data $\{(E_{ij}, x_{ij}),\ j = 1,...,n_i; i = 1,...,m\}$ and census value $x_{i1},...,x_{iN_i}$ are available.

- Using the BHF unit-level model on (y_{ij}, x_{ij}), obtain estimates $(\hat{\beta}, \hat{\sigma}_v^2, \hat{\sigma}_e^2, \hat{v}_i)$.

ELL method

- Generate bootstrap census values y_{ij}^{*} as $y_{ij}^{*} = x_{ij}'\hat{\beta} + v_i^{*} + e_{ij}^{*}$, where $v_i^{*} \sim N(0, \hat{\sigma}_v^2)$ and $e_{ij}^{*} \sim N(0, \hat{\sigma}_e^2)$. Note: Normality is not necessary for ELL. Bootstrap census values can be generated from the residuals $r_{ij} = y_{ij} - x_{ij}'\hat{\beta}_{OLS}$.

- Repeat bootstrap generation for $a = 1, \ldots, A$ and calculate $\theta_i^{*(a)}$ from the bootstrap census values. Then, ELL estimator of θ_i is given by $\hat{\theta}_i^{ELL} = A^{-1}\sum_{a=1}^{A}\theta_i^{*(a)}$. The ELL method also applies to non-sampled areas. The MSE is estimated by simply taking the variability of the A bootstrap estimators, but it can lead to under-estimation.

Modified ELL (MELL)

- Diallo and Rao(2018) generated bootstrap census values y_{ij}^{*} as $y_{ij}^{*} = x_{ij}'\hat{\beta} + \hat{v}_i + e_{ij}^{*}$ for sampled areas and the ELL bootstrap for the non-sampled areas. Note: Normality is not necessary for MELL as in the case of ELL.

- A simulation study using skew normal errors e_{ij} showed large gains in efficiency of MELL over ELL for sampled areas.

- Sinha and Rao(2025) proposed using robust estimates of β, v_i, σ_e^2 (Sinha and Rao, 2009) to generate bootstrap census values, leading to robust MELL (RMELL). In simulations, it led to large efficiency gains over MELL in the presence of outliers in the error term e_{ij}.

CEB method

- In the Census EB (CEB) method under normality, bootstrap census values y_{ij}^{*} are generated as $y_{ij}^{*} = (x_{ij}'\hat{\beta} + \hat{v}_i) + u_i^{*} + e_{ij}^{*}$, where $u_i^{*} \sim N[0, \hat{\sigma}_v^2(1-\hat{\gamma}_i)]$. CEB does not require linking the sample to the census. It loses very little efficiency relative to EB (Molina and Rao, 2010; 440 GS citations) requiring the identification of non-sampled units.

- Rodas, Molina and Ngyuen (2021) gave a detailed study of EB and CEB with several refinements. World Bank added those methods to the World Bank POVMAP software version-2.5.

- Hossain, Das. Chandra and Islam (2020) provide a nice application of EB to provide district level estimates of food insecurity in Bangladesh by linking survey and five percent Census 2011 data.

Some Extensions of BHF model

- **Variable selection** for linear mixed models (Rao and Rao, 2024).

- **Linkage errors**: Two data files combined through record linkage (Han, 2018). Comparison of area level and unit level models in the presence of linkage errors (Consiglio and Tuoto, 2020).

- **Varying regression coefficients and variance components**: Lahiri and Salvati (2023).

- **RF** unit level models: Krennmair and Schmid (2021) and Lohr (2011).

Extensions of unit level models (contd.)

- **Informative sampling within areas**: (1) Extension of pseudo-EBLUP to complex parameters: pseudo-EB (Guadarrama, Molina and Rao, 2018). (2) Extension of Pfeffermann and Sverchkov (2007) method for area means to general area parameters (Cho, Guadarrama-Sanz, Molina, Eideh and Berg , 2024). (3) Comprehensive overview and extensions (Parker, Janicki and Holan, 2020).

- **Data integration by combining probability and non-probability samples**: HB approach (Nandram and Rao, 2024).