On Measuring Uncertainty Associated with Model-based Small Area Estimation: an Appraisal

J. N. K. Rao

Carleton University, Ottawa, Canada

Invited paper presented at 9e Colloque Francophone sur les Sondages, October 11 -14, 2016, Gatineau, Quebec, Canada

Uncertainty measures for direct estimators

- Sample s, design weights $d_k, k \in s$

  Basic design-based estimator of population total Y:

$$\hat{Y} = \sum_{k \in s} d_k y_k = \hat{Y}(y)$$

- Variance estimator: $v(\hat{Y}) = v(y) = s^2(\hat{Y})$

- CV$(\hat{Y}) = s(\hat{Y}) / \hat{Y}$

- Normal theory confidence intervals:

$$[\hat{Y} - z_{\alpha/2}s(\hat{Y}), \hat{Y} + z_{\alpha/2}s(\hat{Y})]$$

- Neyman (1934): If the method of sampling and estimation allows us to ascribe to every possible sample s a confidence interval such that the frequency of errors in the confidence statement does not exceed $1-\alpha$ prescribed in advance, *whatever the unknown properties of the population,* I shall call the method of sampling *representative* and the method of estimation *consistent*.

Variance estimator is not unique:

GREG estimator: $\hat{Y}_G = \sum_{k \in s} (d_k g_k) y_k = \hat{Y} + (X - \hat{X})^T \hat{B}$

Variance estimators: $v(\hat{Y}_G) = v(e)$ and $v_1(\hat{Y}_G) = v(ge)$

Residuals: $e_k = y_k - x_k' \hat{B} = y_k - \hat{y}_k$

GREG weights: $w_k = g_k d_k \ , k \in s$

Example: Ratio estimator $\hat{Y}_R = (\hat{Y} / \hat{X}) X = \hat{R} X$

$g_k = X / \hat{X}$ and $e_k = y_k - \hat{R} x_k$

Which variance estimator to choose?

Ratio estimator is design consistent and also model-unbiased under the ratio model

$$E_m(y_k) = \beta x_k,\ V_m(y_k) = \sigma^2 x_k$$

- Unlike $v(\hat{Y}_R)$, $v_1(\hat{Y}_R)$ is design consistent and also approximately unbiased for the model variance of $\hat{Y}_R$.

What about confidence interval coverage? Is Neyman correct?

- Pivotal $(\hat{Y}_R - Y) / s(\hat{Y}_R) \approx \hat{Y}(e) / s(e)$

- If the deviations from the ratio model are not large, then the skewness in the residuals e_k will be small even if y and x are highly skewed and normal CI perform well. Hence, population structure does matter in design-based inferences contrary to traditional belief.

Example (Dorfman 1994, Rao et al. 2003)

- Two phase SRS, linear regression estimator
  $\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x})$

- True model: $y_i = 8x_i^2 + \varepsilon_i$ and skewness of linear regression residuals 6.40. For first phase sample size $n' = 80$ and second phase sample size $n = 40$, CI coverage is 61% .

- Model-assisted quadratic regression estimator: CI coverage is 91%.

Direct domain estimation

$s(i)$ = sample belonging to domain i, a_{ik} domain indicator

Unbiased estimator: $\hat{Y}_i = \sum_{k \in s(i)} d_k y_k$

GREG estimator:

$$\hat{Y}_{Ri} = \sum_{k \in s(i)} w_k y_k = (X / \hat{X})\hat{Y}_i$$

$\hat{Y}_{Ri}$ gives no gain in efficiency over $\hat{Y}_i$.

Model-based SAE: Basic area level FH model

- Parameters of interest are area (domain) means $\bar{Y}_i$ and associated area level covariates z_i are available.

- Direct estimators $\hat{\bar{Y}}_i$ are not reliable if area sample sizes are small. Necessary to borrow strength across areas through linking models making use of z_i.

- Linking model:

$$\theta_i = g(\bar{Y}_i) = z_i'\beta + v_i,\ v_i \sim_{iid} N(0, \sigma_v^2)$$

- Examples of linking models: $g(P_i) = \sin^{-1}(\sqrt{P_i})$ for proportions $\bar{Y}_i = P_i$ (Casas-Cordero et al. 2016) and $g(\bar{Y}_i) = \log(\bar{Y}_i)$ for income of small places (Fay and Herriot 1987).

- "Matching" sampling model:
  $\hat{\theta}_i = \theta_i + e_i,\ e_i \mid \theta_i \sim_{ind} N(0, \psi_i)$, known ψ_i

- Combined model: $\hat{\theta}_i = z_i'\beta + v_i + e_i$ (linear mixed model)

- Unmatched model uses $\hat{\bar{Y}}_i = \bar{Y}_i + f_i$ with $E(f_i) = 0$.

- Data: $\{(\hat{\theta}_i, z_i), i = 1, \ldots, m\}$. Estimators of model parameters (β, σ_v^2): ML, REML, FH method of moments.

- Empirical Best (EB) estimator of $\theta_i = \bar{Y}_i$:

$$\hat{\theta}_i^{EB} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) z_i' \hat{\beta} \ , \ \hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \psi_i)$$

- Unmatched models: Hierarchical Bayes (HB) estimator of area mean $\bar{Y}_i$ (You and Rao 2002).

Unconditional MSE estimation:

- Second order unbiased MSE estimator for REML:

$$\text{mse}(\hat{\theta}_i^{EB}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2)$$

- First term $g_{1i}(\hat{\sigma}_v^2) = \hat{\gamma}_i \psi_i$. Second term due to estimating β and last term due to estimating σ_v^2 are of lower order $O(m^{-1})$.

Parametric bootstrap

- Step 1: Generate θ_i^* from $N(z_i'\hat{\beta}, \hat{\sigma}_v^2)$.

  Step 2: Generate $\hat{\theta}_i^*$ from $N(\theta_i^*, \psi_i)$.

  Step 3: Apply EB to bootstrap data $(\hat{\theta}_i^*, z_i), i = 1, \ldots, m$ to get EB estimates $\hat{\theta}_i^{EB*}$.

  Step 4: Repeat the steps 1-3 B times.

- Bootstrap MSE estimator:

$$mse_B(\hat{\theta}_i^{EB}) = B^{-1}\sum\nolimits_b [\hat{\theta}_i^{EB*}(b) - \theta_i^*(b)]^2$$

- $\text{mse}_B(\hat{\theta}_i^{EB}) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2)$

- Bias-adjusted estimator tracking $mse(\hat{\theta}_i^{EB})$ can be obtained but requires double bootstrap.

Design properties of $\text{mse}(\hat{\theta}_i^{EB})$

- Assume all parameters are known: $\text{mse}(\hat{\theta}_i^{EB}) = g_{1i}(\sigma_v^2)$

- Design expectation of the average of MSE estimators converges in model probability to the average of the design MSE of EB estimators as $m \to \infty$.

Design unbiased MSE estimator

- Express $\hat{\theta}_i^{EB}$ as $\hat{\theta}_i + h_i(\theta)$ where $\theta = (\theta_1, \ldots \theta_m)'$.

- $$\text{mse}_d(\hat{\theta}_i^{EB}) = \psi_i + 2\partial h_i(\hat{\theta}) / \partial \hat{\theta}_i + h_i^2(\theta)$$

  is unbiased with respect to the sampling model. It can take **negative** values.

- In the special case of known model parameters, we have

$$mse_d(\hat{\theta}_i^B) = \gamma_i \psi_i + (1-\gamma_i)[(\hat{\theta}_i - z_i^T \beta)^2 - (\sigma_v^2 + \psi_i)]$$

- CV of the design MSE estimator can be very large, especially for areas with large sampling variances (Datta et al. 2011). Note that for large ψ_i we have $\gamma_i \approx 0$ and

$$mse_d(\hat{\theta}_i^B) \approx (\hat{\theta}_i - z_i^T \beta)^2 - (\psi_i + \sigma_v^2)$$

which is unstable and can take negative values.

Composite MSE estimator

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

- Less weight is given to design MSE estimator when sampling variance is large and this controls its CV.

- Composite MSE estimator has always smaller design bias than the model based MSE estimator.

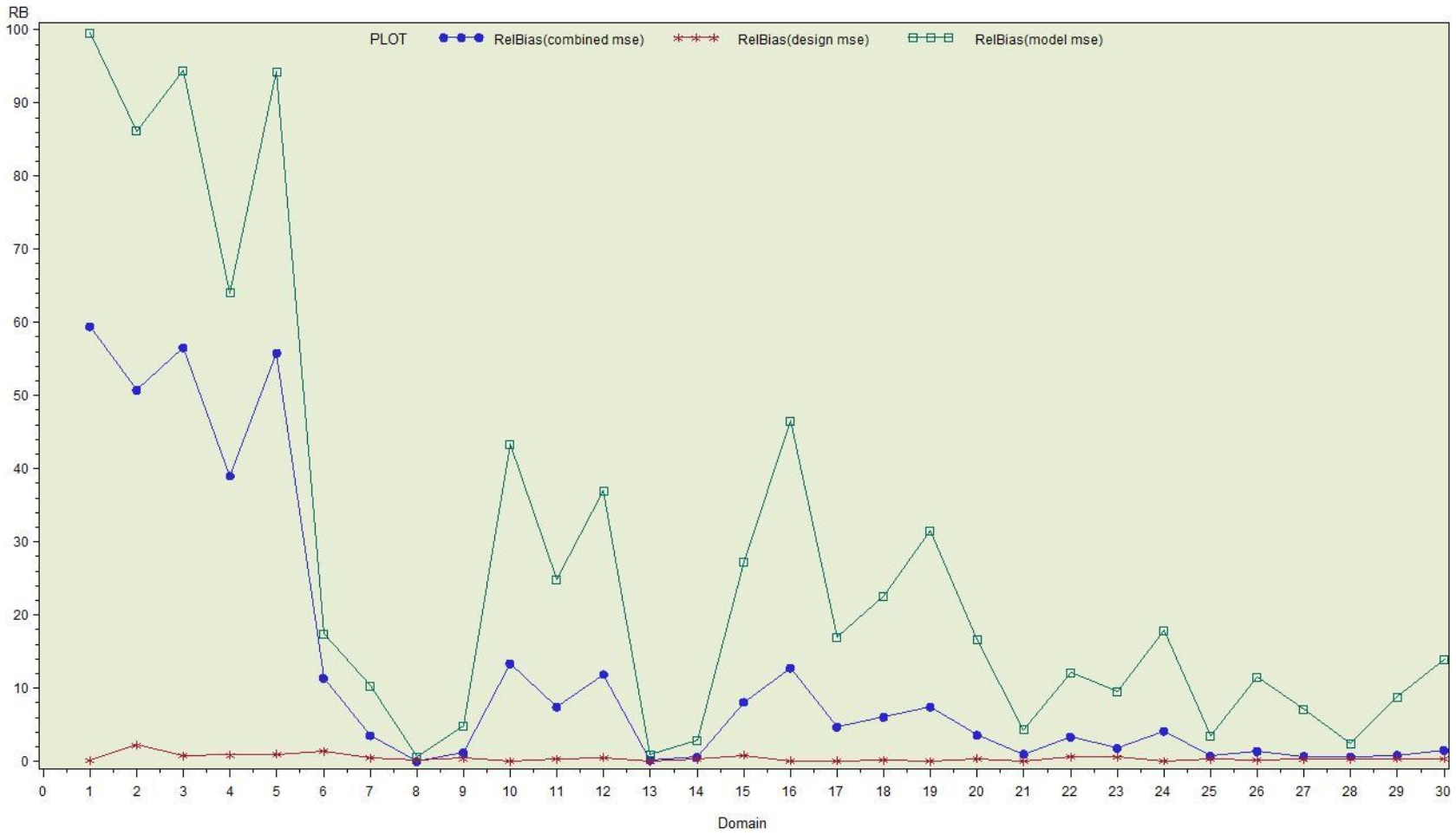
Simulation results

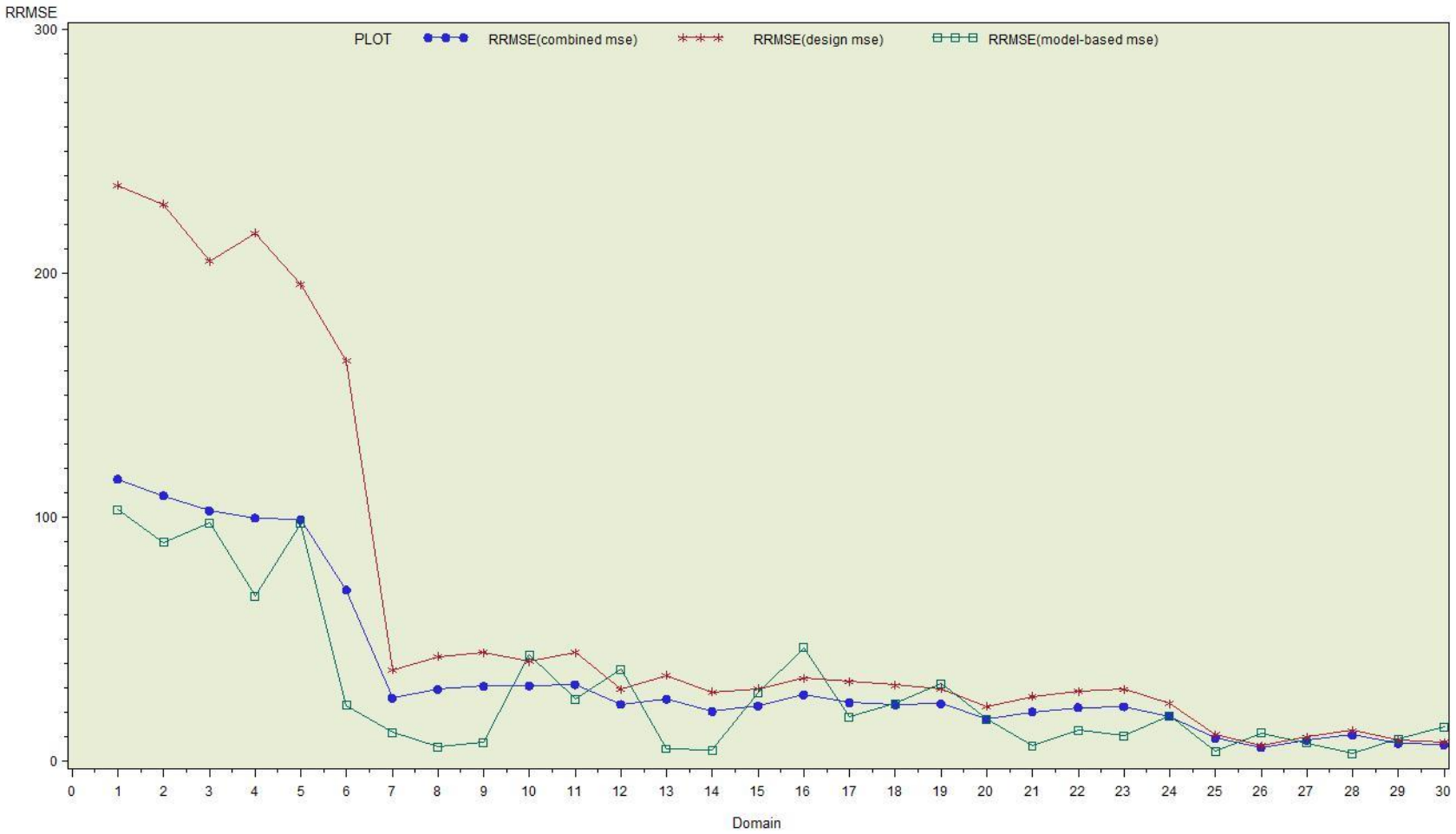
- $m = 30$ areas with ψ_i pattern (2,0.6,0.5,0.4,0.2) each given to six areas. Generate $v_i \sim N(0,1)$ and hold it fixed and calculate $\theta_i = z_i^T \beta + v_i$.

- Generate $\{\hat{\theta}_i^{(r)}, i = 1,...,30; \mathrm{r} = 1,...,\mathrm{R}\}$ from the sampling model $\hat{\theta}_i = \theta_i + e_i$ with $e_i \sim N(0, \psi_i)$.

- From each simulation run estimate model parameters and calculate EB estimators and their MSE estimators using $R = 30,000$. True MSE calculated using $R = 100,000.$

- Relative bias (RB) of MSE estimators and relative root mean squared error (RRMSE) are computed for each area ordered by sample size.


% Relative Bias of MSE estimators
Pattern b, based on 30K samples
RB
PLOT
RelBias(combined mse)
RelBias(design mse)
RelBias(model mse)
100
90
80
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Domain


%RRMSE of MSE estimators
Pattern b, based on 30K samples
RRMSE
PLOT
RRMSE(combined mse)
RRMSE(design mse)
RRMSE(model-based mse)
300
200
100
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Domain

Conditional MSE given $\hat{\theta}_i$

- $MSE_c(\hat{\theta}_i^{EB}) = E[(\hat{\theta}_i^{EB} - \theta_i)^2 \mid \hat{\theta}_i]$

- Datta et al. (2011) obtained a second order unbiased estimator of conditional MSE. Leading term however is unchanged: $g_{1i}(\hat{\sigma}_v^2)$. Only the g_{3i} term is affected and it depends on $\hat{\theta}_i$.

- Conditional MSE estimator is similar to the unconditional MSE estimator in terms of RB and CV.

Confidence Intervals

Pivotal: $t_i = (\hat{\theta}_i^{EB} - \theta_i) / \{g_{1i}(\hat{\sigma}_v^2)\}^{1/2}$

Bootstrap pivotal: $t_i^* = (\hat{\theta}_i^{*EB} - \hat{\theta}_i^{EB}) / \{g_{1i}(\hat{\sigma}_v^{*2})\}^{1/2}$

Procedure: Generate B bootstrap pivotals $t_i^*(1),...,t_i^*(B)$ and determine lower and upper points q_1 and q_2 such that the area between lower and upper points of the empirical bootstrap distribution is equal to specified level $1-\alpha$.

- Bootstrap calibrated interval on θ_i obtained from $q_1 \le t_i \le q_2$ as $c_{1i} \le \theta_i \le c_{2i}$.

- This interval is second order correct provided normality holds (Chatterjee et al. 2008). Interval on a one to one function $h(\theta_i) = g^{-1}(\theta_i)$ is obtained as $h(c_{1i}) \le h(\theta_i) \le h(c_{2i})$.

- Casas-Cordero et al (2014) used bootstrap intervals for poverty rates in Chilean Communas. In their case $\bar{Y}_i = P_i = h(\theta_i) = \sin^2(\theta_i)$.

Basic unit level model

- Unit level data $\{(y_{ij}, x_{ij}),\ j = 1,\ldots,n_i; i = 1,\ldots,m\}$ and population area means $\bar{X}_i$ are available.

- Population nested error linear regression model:

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij},\ j = 1,\ldots,N_i; i = 1,\ldots,m$$

- Population model holds for the sample: no sample selection bias.

- EB estimator $\hat{\bar{Y}}_i^{EB}$ and its model-based MSE estimator are available (Rao and Molina 20015, chapter 7).

- Design-unbiased estimator of $MSE_d(\hat{\bar{Y}}_i^{EB})$ under simple random sampling within areas is derived, assuming model parameters are known.

- Associated composite MSE estimator is obtained as a weighted average of the two MSE estimators.