Hybrid BRR and Parametric-Bootstrap Variance Estimates for Small Domains in Large Surveys

Eric V. Slud, U.S. Census Bureau, CSRM
& Univ. of Maryland, Math. Dept.

Joint Statistical Meetings, August 2017

Joint work with Robert Ashmead

Disclaimer

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are the author's and not necessarily the Census Bureau's.

Voting Rights Act Section 203(b) Mandate

States & political subdivisions (Counties, MCDs, American Indian Areas) must provide voting materials in a language other than English for members of Language Minority Groups (LMGs) according to specific rules based on population fractions.

Census Bureau Director makes the determinations based on (Decennial Census & Amer. Commun. Survey) data, 263 in 2016

Press release Dec. 5, 2016 links to data files and determinations:
https://www.census.gov/newsroom/press-releases/2016/cb16-205.html
https://www.census.gov/rdo/pdf/1_FRN_201628969.pdf

Terminology

US Voting Age persons (2014 5-yr American Community Survey)

partitioned into States; also Jurisdictions (Counties & MCDs)

American Indian Alaska Native pop'n partitioned into AIAs

68 Language Minority Groups: Race self-ID and National Origin (16 Asian, 51 American Indian Alaska Native, plus Hispanic)

CIT : Voting Age Citizen population

LEP : CIT & Limited English-Proficient (Foreign Language spoken in home, English spoken not 'Very well')

ILLIT : LEP CIT with $<$ 5th grade education

Scope of Data Example in this Talk

Restrict to Hispanic LMG within ($\approx$ 7600) Jurisdictions

Rule for Determinations: $\{$ LEP (Jur,Hisp) count $> 10,000$ `or`
(Jur,Hisp) count LEP $>$ 5% (Jur,Hisp) CIT count $\}$
and (Jur,Hisp) ILLIT / (Jur,Hisp) LEP $>$ national illit. rate

Must estimate $N_j^C,\ N_j^L,\ N_j^I,\quad N_j^L/N_{j,\text{tot}}^C,\quad N_j^I/N_j^L$

Direct Survey Estimates and Ratios available: ACS 'special tabulation' from sample sizes n_j and person-weights w_i, e.g.,

$$\hat{N}_j^C = \sum_{i \in (j,\text{Hisp})} w_i \cdot I_{[i \in \text{CIT}]}$$

Abstract Setting

Sample survey with sampled subjects $i \in \mathcal{S}$, weights w_i

Disjoint subdomains C_{jk}, $k = 1, \ldots, K$, partitioning
larger areas D_j, $j = 1, \ldots, m$: $D_j = \cup_{k=1}^{K} C_{jk}$

Survey-weighted counts $\hat{N}_j = \sum_{i \in \mathcal{S} \cap D_j} w_i$, $\hat{N}_{jk} = \sum_{i \in \mathcal{S} \cap C_{jk}} w_i$

Regard $n_j = |\mathcal{S} \cap D_j|$ as fixed, along with
design-based estimators $\hat{N}_j$, $j = 1, \ldots, m$, of N_j

Target proportions: $\underline{\pi}_j = (\pi_{jk},\ k = 1, \ldots, K)$, $\pi_{jk} = N_{jk}/N_j$

Two-Level Small Area Estimation Model

Define $\quad \underline{Y}_j \equiv (Y_{jk},\ k = 1, \ldots, K)\ , \quad Y_{jk} \equiv n_j \, \hat{N}_{jk} / \hat{N}_j$

Data Model: $\quad \underline{Y}_j \sim \text{Multinom}(n_j,\ \underline{\pi}_j)$

Parametric Linking Model: $\quad \underline{\pi}_j \sim f(\underline{\pi},\ \theta,\ \mathbf{X}_j)$

Assume $\quad \{\hat{N}_j\}_{j=1}^m \quad$ indep. of $\quad \{\underline{\pi}_j,\ \underline{Y}_j\}_{j=1}^m$

Shared parameter θ allows borrowing strength

Estimate θ by combined MLE $\hat{\theta}$: how to estimate Variances of Predictors $\quad \tilde{N}_{jk} = Y_{jk} + (\hat{N}_j - n_j) \cdot E_\theta(\pi_{jk} \,|\, \underline{Y}_j)\Big|_{\theta=\hat{\theta}}$?

Why Small-Area Estimation ?

Many samples & VOTAG counts $n_j, \hat{N}_j$ are small ($<$ 10)
only 3671 j's with $\hat{N}_j^L > 0$ and 6837 with $\hat{N}_j > 0$

Most, direct-method (SDR) CV's are very large.

Fractions of 3671 Jur's with sampled LEP citizens in which direct- and model-based estimates of CV fall within ranges.

Range	0–.2	.2–.3	.3–.4	.4–.5	.5–.61	.61–1	$>$1
Direct	.168	.104	.089	.089	.093	.317	.140
Model	.234	.151	.141	.111	.108	.221	.034

64% Direct CV's $>$ 0.4, vs. 47% Hybrid BRR-Model-based

Remarks about Model & Estimation

n_j Hisp voting-age persons deemed fixed (analyses conditional)

$\widehat{N}_j$, $\widehat{N}^C_{j,\mathrm{tot}}$ design-based: random selected weights within Jur. j

Ratios $\underline{\omega}_j = (N_j - N_j^C,\, N_j^C - N_j^L,\, N_j^L - N_j^I,\, N_j^I)/N_j$
modelled as Dirichlet-distributed target Jur random effects.

Data n_j^A, $\widehat{N}_j^A$ (for $A = C, L, I$) on occurrence of CIT, LEP and ILLIT persons within Hispanic VOTAG in j : modelled only through ratio-scaled sample sizes $Y_j^A = n_j\,(\widehat{N}_j^A \,/\, \widehat{N}_j)$

Predictive Covariates X for SAE Model

Proportion Foreign-Born – within Jur or (Jur, Hisp)
Average years in US for Foreign-Born
Proportion $<$ High-School Educ – within Jur or (Jur, Hisp)
State proportion CIT (within Hisp)
State proportion LEP (within Hisp)

Model selection and validation discussed in companion paper
Ashmead & Slud, JSM 2017 earlier in this session.

Model for Borrowing Strength within LMG

Treat $\hat{N}_j$, $\hat{N}^C_{j,\text{tot}}$ via direct survey-weighted ACS estimators

Dirichlet-Multinomial model for C,L,I proportions and counts

$$\underline{\omega}_j \;\equiv\; \frac{1}{N_j}\Big(N_j - N_j^C,\; N_j^C - N_j^L,\; N_j^L - N_j^I,\; N_j^I\Big)$$

$$\sim \text{Dirichlet}(\tau\,\sqrt{n_j},\; (1-\mu_j, \mu_j(1-\nu_j),\; \mu_j\nu_j(1-\rho),\; \mu_j\,\nu_j\,\rho))$$

CIT-rate $\mu_j \;=\; \dfrac{\exp(\beta'\,\mathbf{X}_j)}{1+\exp(\beta'\,\mathbf{X}_j)}$, LEP-rate $\nu_j \;=\; \dfrac{\exp(\gamma'\,\mathbf{X}_j)}{1+\exp(\gamma'\,\mathbf{X}_j)}$

$(n_j - Y_j^C, Y_j^C - Y_j^L, Y_j^L - Y_j^I, Y_j^I) \;\sim\;$ Multinom$(n_j,\; \underline{\omega}_j)$

where $\quad Y_j^A \;=\; (n_j/\hat{N}_j)\hat{N}_j^A\,, \quad A = C,\, L,\, I$

Predictor Formulas (like Beta-Binomial)

$\hat{\theta} \; = \; (\hat{\beta}, \; \hat{\gamma}, \; \hat{\tau}, \; \hat{\rho}) \qquad \text{MLE} \quad \text{from} \quad \{n_j, Y_j^C, Y_j^L, Y_j^I\}_j$

$$\begin{pmatrix} \hat{\omega}_{1j} \\ \hat{\omega}_{2j} \\ \hat{\omega}_{3j} \\ \hat{\omega}_{4j} \end{pmatrix} = \frac{1}{n_j + \hat{\tau}_j} \begin{pmatrix} n_j - Y_j^C \\ Y_j^C - Y_j^L \\ Y_j^L - Y_j^I \\ Y_j^I \end{pmatrix} + \frac{\hat{\tau}_j}{n_j + \hat{\tau}_j} \begin{pmatrix} 1 - \hat{\mu}_j \\ \hat{\mu}_j (1 - \hat{\nu}_j) \\ \hat{\mu}_j \hat{\nu}_j (1 - \hat{\rho}) \\ \hat{\mu}_j \, \hat{\nu}_j \, \hat{\rho} \end{pmatrix}$$

these are the ‘random-effect’ or target predictors

Predictors, continued

Then the pop-count predictors N_j^A (within Hispanic LMG) are

$$\begin{pmatrix} \tilde{N}_j^C \\ \tilde{N}_j^L \\ \tilde{N}_j^I \end{pmatrix} = \begin{pmatrix} Y_j^C \\ Y_j^L \\ Y_j^I \end{pmatrix} + (\hat{N}_j - n_j) \begin{pmatrix} 1 - \widehat{\omega}_{1j} \\ \widehat{\omega}_{3j} + \widehat{\omega}_{4j} \\ \widehat{\omega}_{4j} \end{pmatrix}$$

where

$$\widehat{\tau}_j = \widehat{\tau}\sqrt{n_j} \ , \quad \widehat{\mu}_j = \frac{\exp(\widehat{\beta}' \mathbf{X}_j)}{1 + \exp(\widehat{\beta}' \mathbf{X}_j)} \ , \quad \widehat{\nu}_j = \frac{\exp(\widehat{\gamma}' \mathbf{X}_j)}{1 + \exp(\widehat{\gamma}' \mathbf{X}_j)}$$

Rate predictors: $\tilde{N}_j^L \Big/ \hat{N}_{j,\mathrm{tot}}^C$ and $\tilde{N}_j^I \Big/ \tilde{N}_j^L$

Variance Estimation

- for survey-weighted estimators, via Balanced Repeated Replication (BRR) – Successive Difference Replication (SDR) in ACS

- for functions of model parameters, via Parametric Bootstrap

- treat sample sizes n_j as fixed, and within j
  model for N_j^A/N_j, $n_j\, \widehat{N}_j^A/\widehat{N}_j$ has same form regardless of $\widehat{N}_j$,
  so parametric bootstrap loops nest inside BRR replicates

Balanced Repeated Replication (BRR)

In internal or public-use files for ACS (and other large surveys), responder weights w_i are provided along with replicate weights $w_i^{(r)} = w_i \cdot f_{i,r}, \;\; r = 1, \ldots, R$ (40 sampled r's)

- $f_{i,r}$ constant over i in pseudo-strata defined in ACS by sort order mod R with respect to specific variables ($R = 80$)
- $R^{-1} \sum_{r=1}^{R} f_{i,r} \approx 1$, $(4/R) \sum_{r=1}^{R} (f_{i,r} - 1)^2 \approx 1$
- $\mathrm{Var}(\sum_i w_i \, z_i) \approx (4/R) \sum_{r=1}^{R} \Big(\sum_i w_i^{(r)} \, z_i - \sum_i w_i \, z_i \Big)^2$

Parametric Bootstrap

Within all Jur's j, for fixed $n_j,\, \widehat{N}_j,\, \mathbf{X}_j$ in each

Variance of function $q(n_j,\, \widehat{N}_j,\, Y_j^C,\, Y_j^L,\, Y_j^I,\, \widehat{\theta})$ estimated by

- generating many indep. replicate samples $\underline{\omega}_j^{*(b)},\ Y_j^{*(b)A}$ across all $j = 1, \ldots, m,\ b = 1, \ldots, B$ (B=30)
- estimating MLE $\widehat{\theta}^{*(b)}$ from data $\{Y_j^{*(b)A},\, n_j,\, \mathbf{X}_j\}_{j,A}$
- for each j, calculating sample variance across $b = 1, \ldots, B$ of $q(n_j,\, \widehat{N}_j,\, Y_j^{*(b)C},\, Y_j^{*(b)L},\, Y_j^{*(b)I},\, \widehat{\theta}^{*(b)})$

Hybrid BRR & Parametric Bootstrap

To estimate Mean Squared Prediction Error

$$\text{MSPE} \;=\; E(\tilde{N}_j^A - \omega_j^A\, N_j)^2 \qquad \text{for} \quad \text{A=C,L,I}$$

decompose using independence of $\hat{N}_j$ and $(\underline{Y}_j,\, \hat{\theta})$:

$$\text{MSPE} \;=\; E(\tilde{N}_j^A - \hat{N}_j\, \omega_j^A)^2 + \text{Var}(\hat{N}_j)\, E\Big((\hat{\omega}_j^A)^2 - (\omega_j^A - \hat{\omega}_j^A)^2\Big)$$

Get MSPE from squared residuals nested b within r :

$$\tilde{N}_j^{*(r,b)A} - \hat{N}_j^{(r)}\, \omega_j^{*(r,b)A} \;, \quad \hat{\omega}_j^{*(r,b)A} - \omega_j^{*(r,b)A}$$

Components of MSPE

Decompose mean-square residuals (of 2 types) into pieces

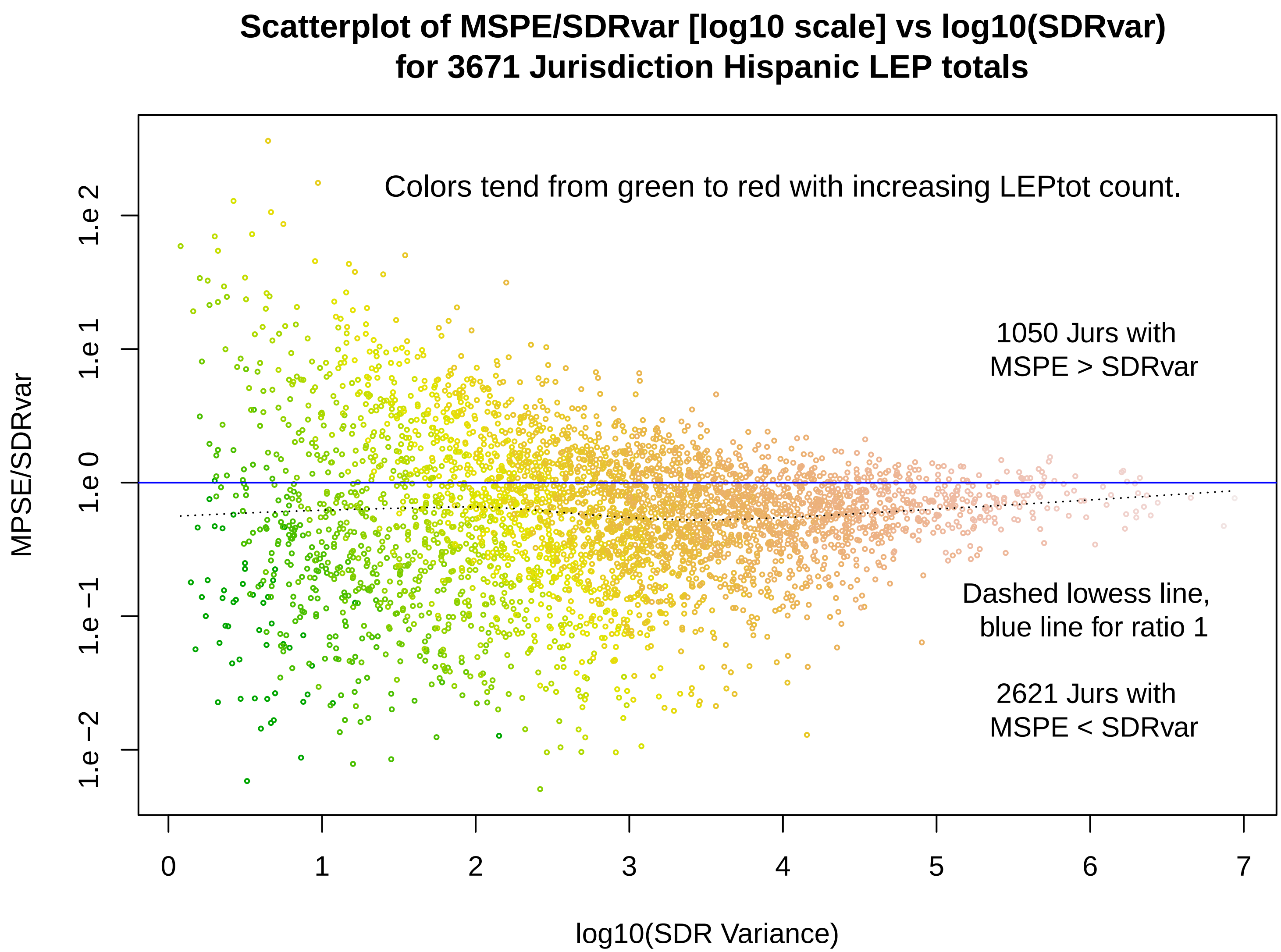
$$E(e_j)^2 = E(e_j - E(e_j \,|\, \hat{N}_j))^2 + E\Big(E(e_j \,|\, \hat{N}_j) - E(e_j)\Big)^2 + (E(e_j))^2$$

Within, Between, Bias-Sq terms respectively estimated by

$$\sum_{r=1}^{R} \sum_{b=1}^{B} \frac{(e_j^{*(r,b)} - \bar{e}_j^{*(r+)})^2}{R\,(B-1)} \,, \quad \frac{4}{R} \sum_{r=1}^{R} (\bar{e}_j^{*(r+)} - \bar{e}_j^{*(0+)})^2 \,, \;\; (\bar{e}_j^{*(0+)})^2$$

where $r = 0$ denotes original sample and

$$\bar{e}_j^{*(r+)} = B^{-1} \sum_{b=1}^{B} e_j^{*(r,b)}$$

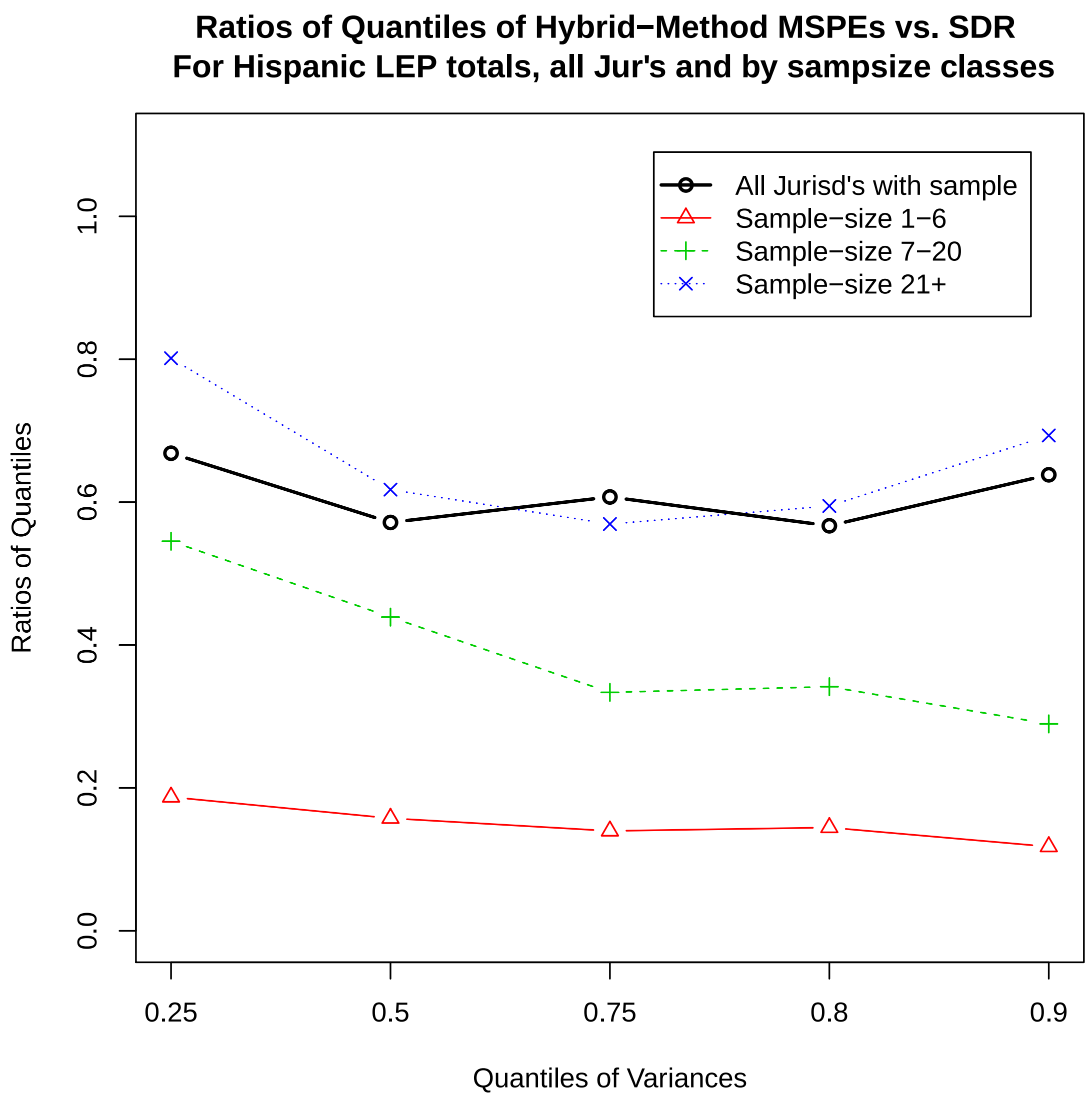

Scatterplot of MSPE/SDRvar [log10 scale] vs log10(SDRvar)
for 3671 Jurisdiction Hispanic LEP totals
Colors tend from green to red with increasing LEPtot count.
1050 Jurs with
MSPE > SDRvar
Dashed lowess line,
blue line for ratio 1
2621 Jurs with
MSPE < SDRvar
MPSE/SDRvar
1.e 2
1.e 1
1.e 0
1.e −1
1.e −2
0
1
2
3
4
5
6
7
log10(SDR Variance)

Summary of Data Results

• Jur's with large variances tend to have large estimates, with relatively smaller improvements of MSPE over SDR Variance

• Great majority of Jur's with intermediate variance have MSPEs much smaller (note log scale!) than SDR Variances, but those which also have smaller sample sizes tend to have MSPE notably worse than SDRvar.

• Jur's with very small sample size (or overall LEPcount estimate) have very small ratios of MSPE over SDR variance.


Ratios of Quantiles of Hybrid−Method MSPEs vs. SDR
For Hispanic LEP totals, all Jur's and by sampsize classes
All Jurisd's with sample
Sample−size 1−6
Sample−size 7−20
Sample−size 21+
Ratios of Quantiles
0.0
0.2
0.4
0.6
0.8
1.0
0.25
0.5
0.75
0.8
0.9
Quantiles of Variances

Further Research on this Topic

- Using independence to remove necessity for double looping (b within r)
- Generalizing model to allow dependence between modeled targets and $\hat{N}_j$ and perhaps to model $n_{jk} = \sum_{i \in \mathcal{S}} I_{[i \in C_{jk}]}$ jointly with other variables

References

Ashmead, R. and Slud, E. (2017) JSM paper on Models in Voting Rights Acts data

Shao, J. and Tu, Y. (1995) **The Bootstrap and Jackknife**

Slud, E. and Ashmead, R. (2017) JSM paper on Variances

Wolter, K. (2007) **Introduction to Variance Estimation**

Thank you !

Eric.V.Slud@census.gov