Unit-Level Nonlinear Models for Small Area Estimation in Informative Samples

Ralph E. Folsom[1], Akhil K. Vaish (avaish@rti.org)
RTI International, 3040 E. Cornwallis Road, P.O. Box 12194
Research Triangle Park, NC 27709

Abstract

Unit-level logistic and exponential models with additive random coefficients (ARCs) are developed for binary and Poisson count data. These ARC models incorporate multivariate small area random effect vectors to accommodate correlated effects for area-level demographic and temporal domain statistics. A survey design-weighted total vector of domain-specific residuals relative to the ARC model's marginal means is used to form an empirical Bayes-type small area estimator (SAE) for the domain total vector. This SAE vector is a linear combination of synthetic universe totals based on the ARC model's marginal means and survey design-consistent generalized regression estimators (GREG) for the domain totals. A version of these SAE vectors is benchmarked to the national sample GREG estimator. A conditional mean squared error (MSE) matrix is developed for the SAE vectors that accounts for all the complex potentially informative features of the sample design, including stratification, clustering, unequal selection probabilities, and weight calibration. Markov Chain Monte Carlo steps were developed to account for the MSE matrix contributions deriving from the uncertainty in the fixed model regression coefficients and random effect variance component matrix.

To demonstrate the ARC model solution's robustness to informative sampling, binary population data were generated from a logistic mixed model for 100 small areas. Both informative and noninformative cluster samples were then drawn from each small area. The ARC model's small area estimates were compared with those from pure logistic mixed model-based hierarchical Bayes (HB) solutions and survey-weighted pseudo-hierarchical Bayes (PHB) estimates for the logistic mixed model based on a previously developed methodology. The ARC model and PHB solutions clearly outperformed the pure logistic mixed model results in the informative sample case. For the noninformative samples, the ARC and PHB model results were reasonably comparable with those from the optimal (HB) mixed model solution. All three solutions suffered from over-shrinkage symptoms in small areas at the upper and lower ends of their population mean ranges. The ARC model solution suffered the least of the three.

Key Words: Small are estimation, Additive random coefficients (ARC); Multivariate random effects; Unit-level models; Benchmarking; Bayes-assisted; Markov Chain Monte Carlo (MCMC); Informative cluster sample simulation.

[1] Dr. Ralph Folsom, former chief scientist at RTI International and ASA Fellow, passed away on December 14, 2022.

1. Introduction

Unit-level small area models, such as the familiar nested error regression (NER) model of Battese, Harter, and Fuller (1988), are potentially more efficient than their area-level Fay-Herriot (1979) counterparts. This potential reduction in small area estimate mean squared errors (MSEs) results from the more efficient outcome-specific regression estimator used by the NER model for its nonsynthetic contribution compared with the possibly weight-calibrated expansion estimator used by the Fay-Herriot (FH) solution. To achieve this gain in efficiency, one has to employ fixed predictors that achieve a notable reduction in the within-area residual variance relative to the total outcome variance. Hidiroglou and You (2016) presented a design-based simulation showing that their survey-weighted pseudo-EBLUP version of the unit-level NER model was robust against highly informative samples and outperformed its FH area-level competitors in terms of root MSE, average absolute bias, and interval coverage rates.

In this paper, we develop nonlinear extensions of the unit-level NER model that allow for a multivariate area-level random effect to simultaneously model statistics for multiple demographic and temporal domains. To minimize the computational burden, we have adopted an additive random coefficient (ARC) version of the generalized linear mixed model (GLMM). This approach is patterned after Singh and Verret's (2006) area-level generalized linear model with additive random components (GLMARC). Our nonlinear ARC superpopulation model has the form,

$$E(y_{dk} \mid \boldsymbol{\eta}_d) = f(\boldsymbol{X}_{dk}\boldsymbol{\beta}) + [\partial f(\boldsymbol{X}_{dk}\boldsymbol{\beta})]\boldsymbol{Z}_{dk}\boldsymbol{\eta}_d \equiv \mu_{dk} \tag{1.1}$$

for unit-k in area-d, where $\boldsymbol{X}_{dk}$ is a vector of predictors linked to unit-k and $\boldsymbol{Z}_{dk}$ is a vector of the indicator variables for the demographic and temporal domains of interest. The *p* element $\boldsymbol{\beta}$ vector depicts fixed regression coefficients, and the *q* element $\boldsymbol{\eta}_d$ vector specifies the area-d specific random effects for the domain statistics of interest. We assume that the $\boldsymbol{\eta}_d$ are independent and identically distributed (i.i.d.) $N_q(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$, where a general $(q \times q)$ covariance matrix $\boldsymbol{\Sigma}_\eta$ is allowed. We choose to allow for between-domain covariances in $\boldsymbol{\Sigma}_\eta$ because such covariances are clearly called for when contrasting time period domains. In our experience, covariances between domain random effects can also be important for comparing youth and parent age group statistics and male and female gender statistics.

In the ARC model, the conditional mean μ_{dk} of y_{dk} given $\boldsymbol{\eta}_d$ is comprised of a marginal mean contribution $f(\boldsymbol{X}_{dk}\boldsymbol{\beta})$ and an ARC contribution $(\partial f)_{dk}\eta_{da}$ for members of domain-a. We consider ARC models for three specific marginal mean functions: namely, the logistic function for binary outcomes y_{dk}, the exponential function for Poisson count data, and the linear version for more continuous, approximately normal outcomes. The logistic and exponential ARC models can be obtained as first order Taylor approximations to the corresponding GLMMs, where $\partial f(\boldsymbol{X}_{dk}\boldsymbol{\beta})$ is $f_{dk}(1-f_{dk})$ for the binary/logistic version and f_{dk} for the Poisson/exponential model. We treat these ARC models as the true superpopulation models and specify MSE estimators for our small area statistics directed at the ARC superpopulation model expectation of sample design-based MSEs. The ARC model is particularly amenable to MSE formulations that fully incorporate possibly informative within-area sample design features, such as stratification, clustering, unequal selection probabilities, and weight adjustments/calibration.

In order to capture all the within-area sample design features in our MSEs, we resort to a Bayes-assisted frequentist approach akin to Singh (2013). Although our small area estimates can be motivated from frequentist considerations, we take advantage of the simple Bayes prescription for sampling general $\boldsymbol{\Sigma}_\eta$ matrices from the inverse Wishart distribution, and we also use a Markov

Chain Monte Carlo (MCMC) approach to incorporate into our MSEs the variance contribution from estimating the $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_\eta$ parameters.

In section 2, we specify an approximate Bayes conditional posterior distribution for our $\boldsymbol{\eta}_d$ random effect vectors, including our scheme for estimating the sample design-based residual covariance matrix involved. In section 3, we use the estimated $\boldsymbol{\eta}_d$ vectors to specify the vector of ARC model small area estimates $\hat{\boldsymbol{\mu}}_d^{ARC}$ for our q area-d domain means. We also present the Bayes conditional posterior covariance matrix for $\hat{\boldsymbol{\mu}}_d^{ARC}$, which is equivalent to the ARC model expected value of the sampling MSE matrix for $\hat{\boldsymbol{\mu}}_d^{ARC}$ given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_\eta$. In section 4, a version of the $\hat{\boldsymbol{\mu}}_d^{ARC}$ vector is developed that is benchmarked to the total sample generalized regression (GREG) estimator $\boldsymbol{\mu}^{GREG}$. Section 5 covers our fixed parameter estimation, and section 6 outlines our MCMC algorithm for adding the MSE matrix contributions from the $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_\eta$ estimation. Section 7 specifies the setup for our sample design-based simulation study involving repeated informative and noninformative cluster samples within each of 100 areas-d. Results are discussed in section 8, and concluding remarks are found in section 9.

2. Approximate Conditional Posterior for $\boldsymbol{\eta}_d$

Consider a probability sample s_d from the area-d universe Ω_d with w_{dk} denoting the inverse selection probability weight for sample respondent $k \in s_d$. Given that $\boldsymbol{\beta}$ is known, we can calculate the survey-weighted vector of residual totals,

$$\boldsymbol{\xi}_{sd} = \sum_{k \in s_d} w_{dk} \boldsymbol{Z}_{dk}^T (y_{dk} - f_{dk}), \tag{2.1}$$

and note that under the ARC model,

$$E_y E_{s|y}(\boldsymbol{\xi}_{sd} \mid \boldsymbol{\eta}_d) = \boldsymbol{\Delta}_{\Omega d}\, \boldsymbol{\eta}_d, \tag{2.2}$$

where $\boldsymbol{\Delta}_{\Omega d} \equiv \sum_{k \in \Omega_d} (\partial f_{dk})\, \boldsymbol{Z}_{dk}^T \boldsymbol{Z}_{dk}$. We restrict our attention to nonoverlapping area-d domains where $Z_{dka} Z_{dka'} = 0$ for $a \neq a'$ and therefore

$$\boldsymbol{\Delta}_{\Omega d} = \underset{1 \le a \le q}{Diag} \left(\sum_{k \in \Omega_{da}} (\partial f_{dk}) \equiv N_{da} \bar{\delta}_{\Omega da} \right), \tag{2.3}$$

where N_{da} is the presumed known population size for domain-a in small area-d. Therefore, conditional on knowing $\boldsymbol{\beta}$, we can form $\boldsymbol{\Delta}_{\Omega d}$ by accessing our universe Ω_d file of $\boldsymbol{X}_{dk}$ predictors and $\boldsymbol{Z}_{dk}$ domain indicators. An unbiased estimator of the finite population realization for $\boldsymbol{N}_d$ given $\boldsymbol{\beta}$ is then

$$\boldsymbol{\tau}_{sd} = \boldsymbol{\Delta}_{\Omega d}^{-1} \boldsymbol{\xi}_{sd} = \underset{1 \le a \le q}{Row} \left[\left(\hat{N}_{da} \bar{\xi}_{sda} \right) / \left(N_{da} \bar{\delta}_{\Omega da} \right) \right]. \tag{2.4}$$

We will choose to poststratify the domain-a weights to our known N_{da} population counts so that $\hat{N}_{da} = \left(\sum_{k \in S_{da}} w_{dk} \right) = N_{da}$ and the τ_{sda} elements of equation (2.4) become $\left(\bar{\xi}_{sda} / \bar{\delta}_{\Omega da} \right)$. Although the poststratified residual totals $N_{da} \bar{\xi}_{sda}$ are no longer unbiased for $N_{da} \bar{\delta}_{\Omega da} \eta_{da}$, there is good reason to believe that their MSEs will be reduced by the weight poststratification. We assume further that the N_{da} are all sufficiently large so that terms of large-order $O\left(N_{da}^{-1} \right)$ are negligible. With the superpopulation models for the y_{dk} stipulating independence for $k \neq k'$ given η_d, we have the following:

$$
\begin{aligned}
& E_y E_{s|y} \left[(\tau_{sd} - \eta_d)(\tau_{sd} - \eta_d)^T \mid \eta_d \right] \\
& = E_y Cov_{s|y}(\tau_d \mid \eta_d) + Cov_y E_{s|y}(\tau_{sd}) \\
& = \Delta_{\Omega d}^{-1} E_y Cov_{s|y}(\xi_{sd} \mid \eta_d) \Delta_{\Omega d}^{-1} + 0 \left[1 / \min_{1 \le a \le q} (N_{da}) \right] \\
& = \Delta_{\Omega d}^{-1} E_y Cov_{s|y}(\xi_{sd} \mid \eta_d) \Delta_{\Omega d}^{-1}.
\end{aligned}
\tag{2.5}
$$

To provide a stable estimator for $E_y Cov_{s|y}(\xi_{sd} \mid \eta_d)$, say $\boldsymbol{C}\xi_{sd}$, we first form a $(q \times q)$ generalized design effect matrix $\boldsymbol{D}_{do}$ for each area-d based on an initial estimator for $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_o$, as follows:

1. Expand area-d sample respondent indices-k to allow for sampling strata $h = 1, ..., H_d$, and stratum-h primary sample clusters $c = 1, ..., r_{dh}$.

2. Form the Taylor-linearized variate column vectors $\boldsymbol{\alpha}_{dhck}$ with *a*-th element $\alpha_{dhcka} = w_{dhck} Z_{dhcka} (\xi_{dk} - \bar{\xi}_{sda})$, where $\xi_{dk} \equiv (y_{dk} - f_{dk})$ and $\bar{\xi}_{sda}$ denotes the w_{dk} weighted residual mean for domain-a in area-d.

3. Compute the primary cluster-level vector totals, $\boldsymbol{\alpha}_{dhc} = \sum_{k \in s_{dhc}} \boldsymbol{\alpha}_{dhck}$, and their simple averages, $\bar{\boldsymbol{\alpha}}_{dh} = \left(\sum_{c=1}^{r_{dh}} \boldsymbol{\alpha}_{dhc} / r_{dh} \right)$.

4. Compute the Taylor series linearized covariance matrix for the poststratified vector of the residual totals as follows:

$$
\boldsymbol{C}\xi_{sdo} = \sum_{h=1}^{H_d} r_{dh} \left[\sum_{c=1}^{r_{dh}} (\boldsymbol{\alpha}_{dhc} - \bar{\boldsymbol{\alpha}}_{dh})(\boldsymbol{\alpha}_{dhc} - \bar{\boldsymbol{\alpha}}_{dh})^T / (r_{dh} - 1) \right], \tag{2.6}
$$

which is the stratified probability proportional to size and with replacement (PPSWR) primary cluster sampling covariance estimator.

5. Ignoring the stratum-h and cluster-c indices, compute the unclustered domain-stratified PPSWR variance estimators S_{sda} for the poststratified residual totals as follows:

$$S_{dao} = n_{da}\left[\sum_{k\in S_{da}} (\alpha_{dka} - \bar{\alpha}_{da})^2 / (n_{da} - 1)\right].$$

$$= \left[n_{da} / (n_{da} - 1)\right]\left[\sum_{k\in S_{da}} w_{dk}^2 (\xi_{dk} - \bar{\xi}_{sda})^2\right]$$

.

6. Now, with $\boldsymbol{S}_{do}^{-1/2} = \underset{1\le a\le q}{Diag}\,(S_{dao}^{-1/2})$ we form

$$\boldsymbol{D}_{do} = \boldsymbol{S}_{do}^{-1/2}\ \boldsymbol{C\xi}_{sdo}\ \boldsymbol{S}_{do}^{-1/2}. \tag{2.7}$$

Averaging over our $d = 1,...,m$ small areas yields the generalized mean design effect matrix $\bar{\boldsymbol{D}}_o$ and the corresponding stabilized version of $\boldsymbol{C\xi}_{sdo}$, namely

$$\boldsymbol{C\xi}_{*do} = \boldsymbol{S}_{do}^{1/2}\ \bar{\boldsymbol{D}}_o\ \boldsymbol{S}_{do}^{1/2}. \tag{2.8}$$

This prescription for forming a stabilized covariance matrix for the small area-level vectors of domain-specific residual totals views the areas as sample strata. In practice, the small areas will more generally represent geographic domains. We presume that our *m* small areas are typically represented by two or more primary sample clusters with all *q* of the subpopulation domains of interest represented in each area-d sample. A few exceptions to this ground truth requirement will be accommodated by providing estimates for empty sample domains conditional on the estimable domain small area estimates. Although we acknowledge the existence of between-small area sampling covariances because target areas are typically not sample design strata, we treat these between-small area sampling covariances as relatively negligible.

Based on our stabilized covariance matrix $\boldsymbol{C\xi}_{*do}$, we form the associated $\boldsymbol{C\tau}_{*do}$ matrix $\boldsymbol{\Delta}_{\Omega do}^{-1}\boldsymbol{C\xi}_{*do}\boldsymbol{\Delta}_{\Omega do}^{-1}$ and assume that given $\boldsymbol{\beta}_o$, the vectors $\boldsymbol{\tau}_{do} = \boldsymbol{\Delta}_{\Omega do}^{-1}\xi_{sdo}$ are approximately *q*-variate normal with mean vectors $\boldsymbol{\eta}_d$, the finite population realizations of our small area random effect vectors, with covariance matrices represented by the stabilized estimates $\boldsymbol{C\tau}_{*do}$. With the $\boldsymbol{\eta}_d$ vectors assumed to have been generated by *m* independent draws from the same *q*-variate $N(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta)$ distribution, the conditional empirical Bayes posterior mean estimator for the $\boldsymbol{\eta}_d$ given $\boldsymbol{\beta}_o, \boldsymbol{\tau}_{do}$, and an estimator for $\boldsymbol{\Sigma}_{\eta o}$ is

$$\hat{\boldsymbol{\eta}}_{do} = \boldsymbol{\gamma}_{do}\ \boldsymbol{\tau}_{do}, \tag{2.9}$$

where $\boldsymbol{\gamma}_{do} = \boldsymbol{\Sigma}_{\eta o}\left[\boldsymbol{\Sigma}_{\eta o} + \boldsymbol{C\tau}_{*do}\right]^{-1}$. The associated conditional posterior covariance matrix for $\boldsymbol{\eta}_d$ is $\boldsymbol{\gamma}_{do}\boldsymbol{C\tau}_{*do}$. In the next section, we employ these empirical Bayes estimators for the area-d random effect vectors to form the corresponding empirical Bayes estimators for our *q*-small area estimator domain totals.

3. Small Area Domain Total Estimates

Given our empirical Bayes estimators for the domain-specific random effect vectors $\hat{\boldsymbol{\eta}}_{do}$ and the assumption that all of the domain-a finite population corrections (FPCs) $\left(n_{da} / N_{da}\right)$ are negligible, the associated ARC model SAE for the area-d vector of y_{dk} domain totals is

$$\boldsymbol{T}_{do}^{ARC} = \sum_{k \in \Omega_d} \boldsymbol{Z}^T{}_{dk} \left(f_{dko} + \partial f_{dko} \boldsymbol{Z}_{dk} \hat{\boldsymbol{\eta}}_{do} \right) = \boldsymbol{\mathcal{T}}_{\Omega do} + \Delta_{\Omega do} \hat{\boldsymbol{\eta}}_{do}. \tag{3.1}$$

With $\boldsymbol{G}_{do} = \left(\boldsymbol{\Delta}_{\Omega do} \boldsymbol{\Sigma}_{\eta o} \boldsymbol{\Delta}_{\Omega do}\right)\left[\left(\boldsymbol{\Delta}_{\Omega do} \boldsymbol{\Sigma}_{\eta o} \boldsymbol{\Delta}_{\Omega do}\right) + \boldsymbol{C}\boldsymbol{\xi}_{*do}\right]^{-1}$, we can recast $\boldsymbol{T}_{do}^{ARC}$ as a convex linear combination of the synthetic marginal mean predictor totals $\boldsymbol{\mathcal{T}}_{\Omega do}$ and the sample design-consistent generalized regression estimator $\boldsymbol{T}_{sdo}^{GREG} = \left(\boldsymbol{\mathcal{T}}_{\Omega do} + \boldsymbol{\xi}_{sdo}\right) = \left[\boldsymbol{Y}_{sd} - \left(\boldsymbol{\mathcal{T}}_{sdo} - \boldsymbol{\mathcal{T}}_{\Omega do}\right)\right]$. Given these definitions, the recast ARC model SAE vector has the following form:

$$\begin{aligned} \boldsymbol{T}_{do}^{ARC} &= \left(I - \boldsymbol{G}_{do}\right) \boldsymbol{\mathcal{T}}_{\Omega do} + \boldsymbol{G}_{do} \left(\boldsymbol{\mathcal{T}}_{\Omega do} + \boldsymbol{\xi}_{sdo}\right) \\ &= \left(I - \boldsymbol{G}_{do}\right) \boldsymbol{\mathcal{T}}_{\Omega do} + \boldsymbol{G}_{do} \boldsymbol{T}_{sdo}^{GREG}. \end{aligned} \tag{3.2}$$

Although this estimator was originally conceived from empirical Bayes considerations, it also has a strong frequentist motivation. One can show that given an estimator $\boldsymbol{t}_d$ of the form in equation (3.2), the matrix $\boldsymbol{G}_d$, which minimizes the ARC model expectation of the sampling MSE for any linear contrast $\left(\boldsymbol{l}^T \boldsymbol{t}_d\right)$, is the same $\boldsymbol{G}_d$ matrix we obtained from empirical Bayes considerations.

With our $\boldsymbol{\xi}_{sdo}$ vectors of sample residual totals assumed to be approximately *q*-variable normal, the conditional posterior covariance matrix for $\boldsymbol{T}_{do}^{ARC}$ given $\boldsymbol{\tau}_{sdo}$, $\boldsymbol{\beta}_o$, and $\boldsymbol{\Sigma}_{\eta o}$ is $\boldsymbol{CT}_{do} = \boldsymbol{G}_{do}\, \boldsymbol{C}\boldsymbol{\xi}_{*do}$. The matrix $\boldsymbol{CT}_{do}$ is also $E_m MSE_{s|m}\left(\boldsymbol{T}_{do}^{ARC}\right)$, where E_m denotes the expectation over the ARC superpopulation model and $MSE_{s|m}\left(\boldsymbol{T}_{do}^{ARC}\right)$ denotes the MSE matrix for $\boldsymbol{T}_{do}^{ARC}$ over repeated samples. In the following section, we propose a benchmarked alternative to the empirical Bayes estimator in equation (3.2).

4. Benchmarking the ARC Model Small Area Estimates

To provide some protection against model misspecification and to possibly add some cosmetic appeal, we benchmark our ARC model small area estimates (SAEs) to equal the sum over areas of the sample design-consistent generalized regression estimators, $\boldsymbol{T}_{sd}^{GREG} = \left(\boldsymbol{\mathcal{T}}_{\Omega d} + \boldsymbol{\xi}_{sd}\right)$. Considering the case where the full *q* element vector $\boldsymbol{\xi}_{sd}$ is available for all small areas-d and $\sum_{d=1}^{m} \Omega_{do} = \Omega_+$, the full population universe, then we require benchmarked ARC model SAEs $\boldsymbol{T}_d^{ARC-B}$ such that

$$\sum_{d=1}^{m} \boldsymbol{T}_d^{ARC-B} = \sum_{d=1}^{m} \left(\mathfrak{T}_{\Omega d} + \xi_{sd} \right) = \mathfrak{T}_{\Omega +} + \xi_{s+}. \tag{4.1}$$

We specify our benchmarked SAE total vectors $\boldsymbol{T}_d^{ARC-B}$ as the modal values of the joint conditional Bayes posterior distribution for the $\left(\boldsymbol{T}_d;\ d = 1,...,m\right)$ vectors subject to the constraint in equation (4.1). This yields the solution

$$\boldsymbol{T}_d^{ARC-B} = \mathfrak{T}_{\Omega d} + \boldsymbol{G}_d\ \boldsymbol{\xi}_{sd} + \boldsymbol{\psi}_d \left[\sum_{d'=1}^{m} \left(I - \boldsymbol{G}_{d'} \right) \boldsymbol{\xi}_{sd'} \right], \tag{4.2}$$

where $\boldsymbol{\psi}_d \equiv \boldsymbol{G}_d\ \boldsymbol{C}\boldsymbol{\xi}_d \left[\sum_{d'=1}^{m} \boldsymbol{G}_{d'} \boldsymbol{C}\boldsymbol{\xi}_{d'} \right]^{-1}$.

Although it is clear that this benchmarked ARC model SAE vector can yield infeasible negative mean estimates for binary and count data, our experience suggests these will be very rare occurrences that can be dealt with in an ad hoc fashion (i.e., set to a small positive value). For the simulation presented here in section 8, we produced 180,000 estimated percentages and observed only four negative values that were all very close to zero.

The ARC model-m expectation of the sampling MSE matrix for $\boldsymbol{T}_d^{ARC-B}$ is

$$E_m\ MSE_{s|m}\left(\boldsymbol{T}_d^{ARC-B}\right) = \boldsymbol{G}_d \boldsymbol{C}\boldsymbol{\xi}_d + \boldsymbol{\psi}_d \left[\sum_{d'=1}^{m} \left(I - \boldsymbol{G}_{d'} \right) \boldsymbol{C}\boldsymbol{\xi}_{d'} \right] \boldsymbol{\psi}_d^T \equiv \boldsymbol{M}_d. \tag{4.3}$$

To this point, our SAE development for the ARC model has been conditional on the fixed parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_\eta$ being known. In the next section, we detail our estimation approach for these parameters.

5. Fixed Parameter Estimation

To estimate the fixed regression coefficients $\boldsymbol{\beta}$, we have implemented a simple survey-weighted pseudo maximum likelihood estimation (MLE) solution available for fixed-effect generalized linear models (GLMs) in software packages such as SUDAAN®, STATA®, and SAS®. Although these are admittedly not the most efficient estimates, they are consistent under the ARC model. Because for the ARC model $f\left(\boldsymbol{X}_{dk}\boldsymbol{\beta}\right)$ is the marginal mean of the y_{dk} outcomes, the survey-weighted pseudo-MLE estimating equations,

$$\boldsymbol{S}_w(\boldsymbol{\beta}) = \sum_{d=1}^{m} \sum_{k \in S_d} w_{dk} \boldsymbol{X}_{dk}^T \left[y_{dk} - f_{dk}(\boldsymbol{\beta}) \right], \tag{5.1}$$

yield consistent $\boldsymbol{\beta}$ estimates. Note that these equations do not involve the $\boldsymbol{\Sigma}_\eta$ covariance matrix. With $\boldsymbol{\beta}_O$ depicting the solution to equation (5.1) and $\boldsymbol{C}\boldsymbol{\beta}_O$ representing an estimator for $Cov_S\left(\boldsymbol{\beta}_O\right)$, we assume that with an improper flat prior on $\boldsymbol{\beta}$ that the conditional posterior for $\boldsymbol{\beta}$ given $\boldsymbol{\Sigma}_\eta$ is p-variable normal with mean vector $\boldsymbol{\beta}_O$ and covariance matrix $\left(\boldsymbol{C}\boldsymbol{\beta}_O + \boldsymbol{C}\boldsymbol{\beta}_m\right)$, where $\boldsymbol{C}\boldsymbol{\beta}_m = Cov_m E_{s|m}\left(\beta_O\right)$, which depends on $\boldsymbol{\Sigma}_\eta$ and may not be negligible. Appendix A contains a short description of how we estimate $\boldsymbol{C}\boldsymbol{\beta}_m$ given $\boldsymbol{\Sigma}_\eta$.

We have considered a potentially more efficient solution for $\boldsymbol{\beta}$ inspired by the survey-weighted generalized estimating equations (GEEs) that are based on the ARC model covariance structure.

Given a set of $\hat{\eta}_d$ random effect estimates, these GEE-inspired estimating equations take the following form:

$$S_w(\beta) = \sum_{d=1}^{m} \sum_{k \in s_d} w_{dk} X_{dk}^T \left[y_{dk} - f_{dk}(\beta) - \partial f_{dk}(\beta) Z_{dk} \hat{\eta}_d \right], \tag{5.2}$$

which employs residuals based on the full conditional means $\mu_{dk}(\beta, \hat{\eta}_d)$. We plan to test this β solution in the next version of our ARC model software.

To estimate Σ_η, we assume that the conditional posterior distribution for η_d is q-variate normal with mean vector $\gamma_d \tau_d$ and covariance matrix $\gamma_d C\tau_{*d}$. Armed with a set of sample draws η_{dt} from these conditional posteriors, we form $\bar{A}_t = \left(\sum_{d=1}^{m} \eta_{dt} \eta_{dt}^T / m \right)$ and draw $\Sigma_{\eta t}$ from the inverse Wishart distribution with $(m+q+2)$ degrees of freedom and scale matrix $\left(m\bar{A}_t + \Sigma_{\eta o} \right)$. This assumes a proper inverse Wishart prior for Σ_η with $(q+2)$ degrees of freedom and scale matrix $\Sigma_{\eta o}$. This is a diffuse prior with mean $\Sigma_{\eta o}$ and infinite element variances.

In the next section, we outline a set of MCMC steps that we use to add the second order contributions to our expected MSE matrices. These contributions result from estimating the model parameters β and Σ_η that we conditioned on in section 3 to obtain the first order MSE matrix estimator. We also consider accounting for the fact that our small area-d by domain-a unstratified and unclustered PPSWR residual variance estimates S_{da} employed in the stabilized sampling covariance matrices $C\xi_{*d}$ are not strictly error free.

6. MCMC Second Order MSE Contribution

At the first step in our MCMC cycle, we obtain the survey-weighted pseudo-MLE solution β_o and its sampling covariance matrix $C\beta_o$, which does not depend on Σ_η. For $t = 1, ..., K$ MCMC cycles, we select vectors $\beta_t \sim N(\beta_o, C\beta_o + C\beta_{mt})$, recalling that $C\beta_{mt} = Cov_m E_{s|m}(\beta_o)$, which depends on $\Sigma_{\eta(t-1)}$ (see Appendix A). Given β_t, we form the universe-level vectors $\mathfrak{I}_{\Omega d}(\beta_t)$ of area-d domain-level synthetic predictor totals from equation (3.1) and the associated diagonal matrices $\Delta_{\Omega d}(\beta_t)$ of domain-level derivative totals defined in equation (2.3).

Now, given $\Sigma_{\eta(t-1)}$, we proceed to form T_{dt}^{ARC-B} and M_{dt} for our estimate for $EMSE_{dt}^{B}$, the ARC model-expected value of the sampling MSE matrix for T_{dt}^{ARC-B}. From equation (3.1), we note that the G_{dt} shrinkage matrix involves both $\Sigma_{\eta(t-1)}$ and $C\xi_{*dt} = S_{dt}^{1/2} \bar{D}_{do} S_{dt}^{1/2}$. Recognizing that the domain stratified-unclustered PPSWR residual variance estimates S_{dat} on the diagonal of S_{dt} are not strictly known without error, we have considered adopting the assumption applied by Wang and Fuller (2003) for small areas with estimated variances such as our S_{dat}, namely, that $(n_{da}-1)S_{dat} \sim V_{dat} \chi^2(n_{da}-1)$ where $V_{dat} \equiv E_S(S_{dat})$ and χ_{df}^2 is a chi-squared random variable with df degrees of freedom. Assuming further that V_{dat} have improper priors of

the form $P(V_{dat}) = V_{dat}^{-(\nu/2)-1}$ with $\nu = 2$, then the posterior for V_{dat} given S_{dat} is inverse Gamma with $\alpha = \left[(n_{da}+1)/2\right]$ and $\beta = \left[(n_{da}-1)S_{dat}/2\right]$. To account for this sampling error in the S_{dat} variance estimates, one could substitute draws V_{dat} from the inverse-Gamma (α,β) distribution with α and β as specified above. This would yield V_{dat} values with $E(V_{dat}) = S_{dat}$ and $Var(V_{dat}) = \left[2(S_{dat})^2/(n_{da}-3)\right]$. This would lead to

$$\boldsymbol{C\xi}_{*dt} = \boldsymbol{V}_{dt}^{1/2}\bar{\boldsymbol{D}}_{do}\boldsymbol{V}_{dt}^{1/2}. \tag{6.1}$$

We plan to implement this enhancement in the next update of our ARC model software.

Having formed the benchmarked total vectors $\boldsymbol{T}_{dt}^{ARC-B}$ and their estimated conditional $EMSE_{dt}^{B}$ matrices $\boldsymbol{M}_{dt}$, we then draw the $\boldsymbol{\eta}_{dt}$ random effect vectors from their conditional $N(\gamma_{dt}\tau_{sdt};\gamma_{dt}\boldsymbol{C\tau}_{dt})$ posteriors, recalling that $\boldsymbol{C\tau}_{dt} = \left(\boldsymbol{\Delta}_{\Omega dt}^{-1}\boldsymbol{C\xi}_{*dt}\boldsymbol{\Delta}_{\Omega dt}^{-1}\right)$ and $\gamma_{dt} = \boldsymbol{\Sigma}_{\eta(t-1)}\left[\boldsymbol{\Sigma}_{\eta(t-1)} + \boldsymbol{C\tau}_{dt}\right]^{-1}$. Now, with

$$\boldsymbol{A}_t \equiv \sum_{d=1}^{m}\boldsymbol{\eta}_{dt}\boldsymbol{\eta}_{dt}^{T}, \tag{6.2}$$

we draw $\boldsymbol{\Sigma}_{\eta t} \sim W^{-1}\left[(m+q+2),\left(\boldsymbol{A}_t + \boldsymbol{\Sigma}_{\eta o}\right)\right]$, where $E\left(\boldsymbol{\Sigma}_{\eta t}\right) = \left(\boldsymbol{A}_t + \boldsymbol{\Sigma}_{\eta o}\right)/(m+1)$.

When we have completed K of these MCMC cycles and thinned them sufficiently to remove excess serial correlation, we use the Rao-Blackwell prescription to compute the second order adjusted version of our estimated $EMSE\left(\boldsymbol{T}_{d}^{ARC-B}\right)$ matrices. Specifically, we estimate $EMSE\left(\boldsymbol{T}_{d}^{ARC-B}\right)$ as follows:

$$EMSE\left(\boldsymbol{T}_{d}^{ARC-B}\right) \triangleq \bar{\boldsymbol{M}}_d + \sum_{t=1}^{K}\left(\boldsymbol{T}_{dt}^{ARC-B} - \bar{T}_{d}^{ARC-B}\right)\left(\boldsymbol{T}_{dt}^{ARC-B} - \bar{T}_{d}^{ARC-B}\right)^{T}/K, \tag{6.3}$$

where $\boldsymbol{T}_{dt}^{ARC-B}$ and $\boldsymbol{M}_{dt}$ are defined in equation (4.2) and equation (4.3) at MCMC cycle-t. The $\bar{T}_{d}^{ARC-B}$ and $\bar{\boldsymbol{M}}_d$ versions are simple averages over a thinned subset of the MCMC cycles.

Instead of using the thinned cycle average of the $\boldsymbol{T}_{dt}^{ARC-B}$ benchmarked SAE totals to specify our point estimates, we choose to use an empirical Bayes-type estimator based on the pseudo-MLE $\boldsymbol{\beta}_o$ vectors and shrinkage matrices $\boldsymbol{G}_{do}$ based on the MCMC cycle averaged $\boldsymbol{\Sigma}_{\eta t}$ matrices, say $\bar{\boldsymbol{\Sigma}}_{\eta}$, and our original $\boldsymbol{C\xi}_{*do}$ stabilized sampling covariance matrices. We favor these empirical Bayes point estimates because they are benchmarked in terms of the original $\boldsymbol{\beta}_o$ pseudo-MLE estimates and the associated aggregate residual total vector $\sum_{k=1}^{m}\boldsymbol{\xi}_{sdo} \equiv \boldsymbol{\xi}_{s+o} = \varnothing_q$, the q element null vector. This identity holds assuming that the q domain indicators in $\boldsymbol{Z}_{dk}$ are always included as fixed predictors in $\boldsymbol{X}_{dk}$. This leads to the benchmark constraint:

$$\sum_{d=1}^{m} \boldsymbol{T}_{do}^{ARC-B} = \sum_{d=1}^{m} \mathfrak{T}_{\Omega do}. \tag{6.4}$$

7. Cluster Sample Simulation Design

To evaluate the performance of our ARC model SAEs and Bayes-assisted frequentist solution, we generated for $m = 100$ small areas $N_d \approx 6,000$ binary observations y_{dk} using latent $N(0,1)$ random variables e_{dk}. To facilitate the selection of an informative sample where the logistic mixed model that holds in the population does not hold in the selected samples, we placed one fourth of $prob(y_{dk} = 1) = \mu_{dk}$ in the lower tail and $(3/4)\mu_{dk}$ in the upper tail and set the following:

$$y_{dk} = \begin{cases} 1 \text{ if } e_{dk} \geq \phi^{-1}\left[1-(0.75)\mu_{dk}\right] \text{ or } e_{dk} \leq \phi^{-1}\left[(0.25)\mu_{dk}\right] \\ 0 \text{ otherwise,} \end{cases} \tag{7.1}$$

where $\phi(\ )$ denotes the $N(0,1)$ CDF.

The logistic mixed model we use to specify μ_{dk} has the form $\ln\left[\mu_{dk}/(1-\mu_{dk})\right] = \boldsymbol{x}_{dk}\boldsymbol{\alpha} + \boldsymbol{z}_{dk}\boldsymbol{\eta}_d$, where $\boldsymbol{x}_{dk} = (1, g_{dk}, a_{dk}, \chi_{dk})$ with g_{dk} and a_{dk} denoting $1/0$ indicator variables for male gender and the first of two age groups, respectively. The continuous predictors χ_{dk} have the form $\chi_{dk} = u_d + \varepsilon_{dk}$ with $u_d \sim N(0,0.1)$ and $\varepsilon_{dk} \sim N(0,1)$. To generate a wide range of small area domain proportions, we set $\alpha_0 = -2.4$ for the intercept, $\alpha_1 = 0.6$ for males, $\alpha_2 = 0.1$ for age group-1, and $\alpha_3 = 1.0$ for the continuous predictors χ_{dk}.

Random effects η_{da} were estimated in each small area for four demographic domains defined by the four age by gender cross-class cells. Specifically, $\boldsymbol{Z}_{dk} = \langle g_{dk}a_{dk}, g_{dk}(1-a_{dk}), (1-g_{dk})a_{dk}, (1-g_{dk})(1-a_{dk})\rangle$. We patterned the 4×4 random effect covariance matrix $\boldsymbol{\Sigma}_\eta$ used for our simulations after similar covariance matrices observed in our National Survey on Drug Use and Health (NSDUH) substate region small area estimate modeling.

Specifically, we use the following correlation matrix:

$$\boldsymbol{R}_\eta = \begin{bmatrix} 1.00000 & 0.58019 & 0.60189 & 0.48326 \\ 0.58019 & 1.00000 & 0.43023 & 0.67175 \\ 0.60189 & 0.43023 & 1.00000 & 0.41131 \\ 0.48326 & 0.67175 & 0.41131 & 1.00000 \end{bmatrix}.$$

This matrix is used in combination with the diagonal cross-class domain variance components $\sigma_\eta(11) = 0.06968, \sigma_\eta(22) = 0.15859, \sigma_\eta(33) = 0.08780,$ and $\sigma_\eta(44) = 0.20159$. The male gender indicators g_{dk} were generated as independent Bernoulli random variables with $P = (0.5)$. The indicators for age group-1 were independent Bernoulli with $P = (0.6)$.

We group the $N_d \approx 6,000$ records in each area into around 70 clusters ranging in size from 60 to 140 records with varying numbers of records where the latent variable $e_{dk} \geq 0$. We denote this subpopulation of area-d by $[k \in \Omega_{d+}]$. For our informative sample, we selected eight clusters from each area-d using the probability proportional to size (PPS) without replacement method (Sampford

1967) while also using a composite size measure (Folsom, Potter, and Williams 1987) that oversamples records from the latent domain Ω_{d+} where $e_{dk} \geq 0$. This composite size measure has the form

$$S_{dc} = (120 / N_{d+}) N_{dc+} + (40 / N_{d-}) N_{dc-}, \tag{7.2}$$

where N_{d+} and N_{d-} are the area-d population counts for latent domains Ω_{d+} and Ω_{d-}, respectively. The associated domain counts for cluster-c are N_{dc+} and N_{dc-}. Having thus selected 8 sample clusters, we then drew 20 records from each cluster via stratified random sampling with randomly rounded versions of the stratum allocations $n_{dc+} = 20\left[(120 / N_{d+}) N_{dc+} / S_{dc}\right]$ and $n_{dc-} = \left[20 - n_{dc+}\right]$. This cluster sampling scheme yields equal probability samples from the two latent domains in each small area with inverse selection probability weights:

$$w_{dck} = \begin{cases} (N_{d+} / 120) & \text{for } k \in \Omega_{d+} \\ (N_{d-} / 40) & \text{for } k \in \Omega_{d-} \end{cases}. \tag{7.3}$$

Note that by placing $(3/4)$th of μ_{dk} in the upper tail critical region for the latent variable e_{dk}, $\text{Prob}(y_{dk} = 1 \mid e_{dk} \geq 0) = (1.5)\mu_{dk}$ and $\text{Prob}(y_{dk} = 1 \mid e_{dk} < 0) = (0.5)\mu_{dk}$. Therefore, selecting cases with $e_{dk} \geq 0$ at 3 times the rate of cases with $e_{dk} < 0$ yields an informative sample that does not conform to the population logistic mixed model.

To specify a complex cluster sample for each small area-d that is noninformative, we use the sign of our continuous model predictor χ_{dk} to specify the small area domain for oversampling. If M_{d+} denotes the area-d population count for cases with $\chi_{dk} \geq 0$ and M_{d-} the count with $\chi_{dk} < 0$, then our noninformative sample selects eight small area clusters with the size measure $S_{dc} = (120 / M_{d+}) M_{dc+} + (40 / M_{d-}) M_{dc-}$, where the M_{dc+} and M_{dc-} represent cluster-dc level counts for cases with positive and negative χ_{dck} continuous predictors. The corresponding sample allocations for the stratified random sampling within-cluster selections are then allocated as before; namely, $m_{dc+} = 20\left[(120 / M_{d+}) M_{dc+} / S_{d+}\right]$ and $m_{dc-} = (20 - m_{dc+})$. Although this second sample overrepresents cases with positive continuous predictors, it is nevertheless noninformative because the population model is conditional on the χ_{dck} predictors and therefore still holds in the sample.

We contrast the performance of our logistic ARC model with the Folsom, Shah, Vaish (1999) pseudo-hierarchical Bayes (PHB) solution that fits a version of the true logistic mixed model used to generate our population data. The Folsom et al. (1999) PHB method is patterned after the Zeger and Karim (1991) hierarchical Bayes (HB) solution for generalized linear models. Our PHB version employs a survey-weighted pseudo-likelihood in combination with Gibbs sampling and an acceptance/rejection algorithm. We specify a diffuse inverse Wishart prior for $\boldsymbol{\Sigma}_{\eta}$. We also provide for comparison pure model-based HB small area estimates by running our PHB software with all of the survey weights set to 1.

To improve the logistic ARC model approximation to our true logistic mixed model, we make use of the Zeger, Liang, and Albert (1998) approximation for the marginal mean of the logistic mixed model. Their approximation has for our true model the form $\text{logit}\left(f_{dk}^{m}\right) \approx \left(\boldsymbol{x}_{dk}\boldsymbol{\alpha}^{c}\right) / \left[1 + (0.3458)\sigma_{\eta a}^{2}\right]^{1/2}$ for persons-k belonging to the a-th gender by age cross-class domain, where $\boldsymbol{\alpha}^{c}$ denotes the vector of four fixed regression coefficients in the conditional on the $\boldsymbol{\eta}_{d}$ logistic mixed model and $\sigma_{\eta a}^{2}$ is the diagonal variance component in $\boldsymbol{\Sigma}_{\eta}$ corresponding to cross-class domain-a. This implies that our ARC model marginal means

$f_{dk}(\boldsymbol{x}_{dk}\boldsymbol{\beta})$ should have separate intercepts β_{oa} and continuous predictor χ_{dk} slopes β_{1a} for each of our four cross-class (gender by age group) domains. This leads to an expanded ARC model with eight fixed logistic coefficients, a separate intercept, and a separate χ-variable slope for each of our four gender by age group domains.

The goal of our simulation study is to evaluate the performance of our ARC model SAEs in terms of how well they estimate the small area finite population proportions derived from a single realization or draw from our superpopulation model. This led us to generate only one population using the $N(0,1)$ latent variable with imbalanced tail areas for setting the y_{dk} Bernoulli variables. Given this single generated population, we the selected 200 independent cluster samples from each of our 100 small areas using both the informative and noninformative sampling scheme outlined previously.

8. Simulation Results

In figures 1 and 2, we display simulation-based small area estimate biases for the pure model-based (HB) solution, the survey-weighted (PHB) results, and the ARC model estimates under the noninformative and informative sampling scenarios. Figure 1 is for females, the best performing domain for the ARC model with an average sample size of 80.2. The results are averages over 10 small areas that were grouped in rank order based on their population proportions for the female marginal domain. One would expect the small area estimates for females that are based on the true logistic mixed model, namely, the unweighted (HB) solution and the survey-weighted (PHB) result that accounts for survey weighting but otherwise ignores the complex cluster sample design features, to have the smallest average biases when the sample is noninformative. The ARC model's small area estimates, although derived from an ARC approximation to the logistic mixed model and further encumbered by unnecessary survey design weighting and clustered MSE estimation, actually exhibit the smallest biases for both the noninformative and informative samples. When the logistic mixed model does not hold in the sample, as illustrated here by our informative case, the pure model-based (HB) small area estimates have large biases across all 10 of the small area groups. Although the survey weighting in the ARC and PHB solutions tends to control for the informative sampling bias reasonably well in the interior area groupings, their average biases tend to increase in the outlying area groups. This result is reminiscent of the overshrinkage tendency that others have observed for small area estimates when their finite population targets rank on the low and high ends of their range. Figure 2 shows similar results for the males in age group-2 with an average area sample size of 31.9. This was the smallest cross-class domain (sample size wise) and was the domain where the ARC model had the poorest showing.

Figures 3 and 4 show the simulation-based true root-MSEs (RMSEs) for the two area-level domains (females and age group-2 males). Each figure contrasts the performance of the ARC , PHB, and HB solutions under the noninformative and informative samples. When the population logistic mixed model holds in the sample, the noninformative case, we see that as one would expect our ARC model's root-MSEs are somewhat inflated compared with the HB and PHB solutions that fit the correct model. On the other hand, when the logistic mixed model does not hold in the informative sample, we see that the pure model-based HB solution that makes no adjustments for the complex cluster sample design features exhibits substantial RMSE inflation.

Figures 5 and 6 display biases in the estimated RMSEs averaged over 10 adjacent population areas based on the domain population mean ranks. For the noninformative case, we see that the ARC model solution more than holds its own relative to the HB and PHB results that fit the correct model and are less encumbered with what should be superfluous adjustments for sample design features. For the informative sample cases, the HB solution consistently underestimates its RMSEs across all 10 area groups. Although the ARC and PHB model RMSE average biases are of similar magnitude for the eight interior area groupings, the ARC solution exhibits somewhat less bias in the two extreme groups where overshrinkage is the most pronounced.

In figures 7 and 8, we present confidence interval coverage rates for the HB, PHB, and ARC model small area estimates. We formed these intervals by matching the small area point estimates and their MSEs to the means and variances of the beta distribution and used the associated 2.5 and 97.5 distribution percentiles as the interval end points. For the noninformative case, the female domain small area estimates $(n \approx 80)$ under the HB, PHB, and ARC models all tend to approach the targeted 95 percent coverage rates for the eight interior small area groups and drop to 90 percent or below in the two extreme groups where overshrinkage comes into play. Under the informative sampling scenario, the ARC model and PHB solutions exhibit similar coverage rate patterns for the female domain while the HB version suffers a notable drop in coverage across all 10 small area groupings. The coverage results for the age group-2 males $(n \approx 32)$ in figure 8 tend to fall between 90 and 95 percent for all three solutions under the noninformative sample scenario except for the two outer groups where it drops substantially. It is clear that the ARC model coverage rates are deflated substantially less than the HB and PHB versions in these outer area groups, particularly at the low end. This result is also observed in figure 7. Under the informative scenario, the HB coverage rates continue to be seriously deflated. The ARC and PHB coverage rates for this small sample domain again tend to fall between 90 and 95 percent in the interior small area groups under the informative sampling scenario. The ARC model continues to exhibit some coverage advantage relative to the PHB results in the two extreme area groupings.

9. Conclusions

We have developed a Bayes-assisted methodology for fitting unit-level nonlinear small area estimation models. These models include multivariate additive random coefficients for targeted small area demographic and temporal domain statistics. An estimation methodology was presented that accounts for all of the potentially informative features of a complex survey design, including stratification, clustering, unequal selection probabilities, and weight calibration for nonresponse and frame undercoverage.

A simulation study was conducted to contrast the performance of our logistic ARC model solution with an unweighted hierarchical Bayes fit to the logistic mixed model used to generate the binary population data. A survey weighted pseudo-hierarchical Bayes solution (Folsom et al. 1999) was also evaluated as an ARC model competitor. Results were developed for 200 informative and 200 noninformative cluster samples drawn from each of 100 small areas. Small area estimates (percentages) were developed for four gender by age group domains, for the two gender and age group margins, and for the overall area mean percent.

Results were presented for the female margins $(n \approx 80)$ and the male by age group-2 cross-class domain $(n \approx 32)$. We averaged the results by 10 groupings of 10 small areas formed by rank ordering the areas by the respective target domains' population means. As expected, the small area estimate biases, estimated RMSE biases, and beta confidence interval (CI) coverage rates for the unweighted HB solution suffered notably for both domains under the informative sampling scheme. The ARC model and PHB results were comparable with the HB solution under the noninformative samples and were both reasonably robust to the informative scenario. All three methods tended to exhibit overshrinkage symptoms in the two outer (lowest and highest) population mean groupings. The ARC model solution appeared to be somewhat less susceptible to this overshrinkage-induced small area estimate bias and RMSE underestimation in the tail areas.

In conclusion, the nonlinear unit-level ARC model solution we propose for simultaneously estimating multiple small area domain statistics appears quite promising. The overshrinkage problem is, however, an issue that clearly requires further work.

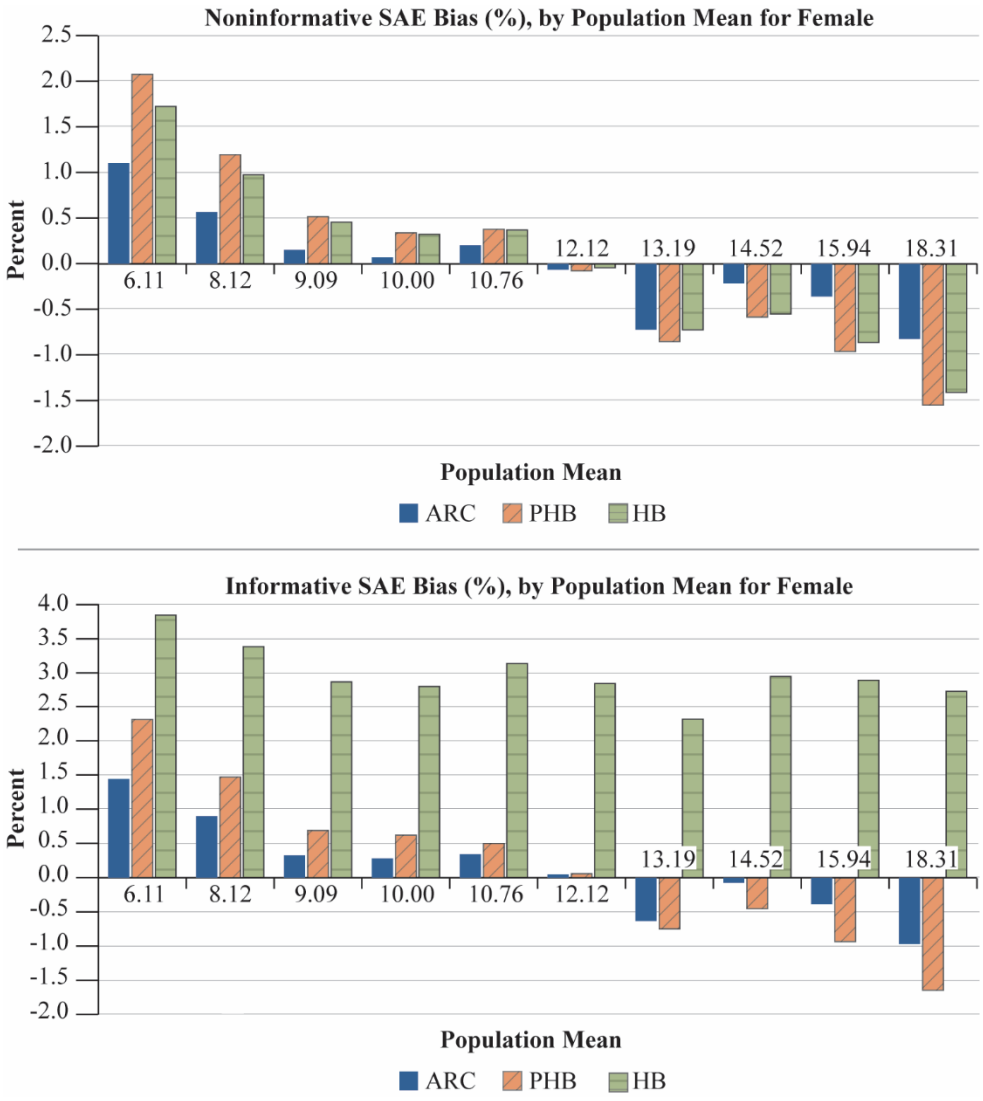

Figure 1. Noninformative and Informative Sample Small Area Estimate Biases (%), by Population Mean for the Female Margin Domain. ARC = additive random coefficient; PHB = pseudo-hierarchical Bayes; HB = hierarchical Bayes.

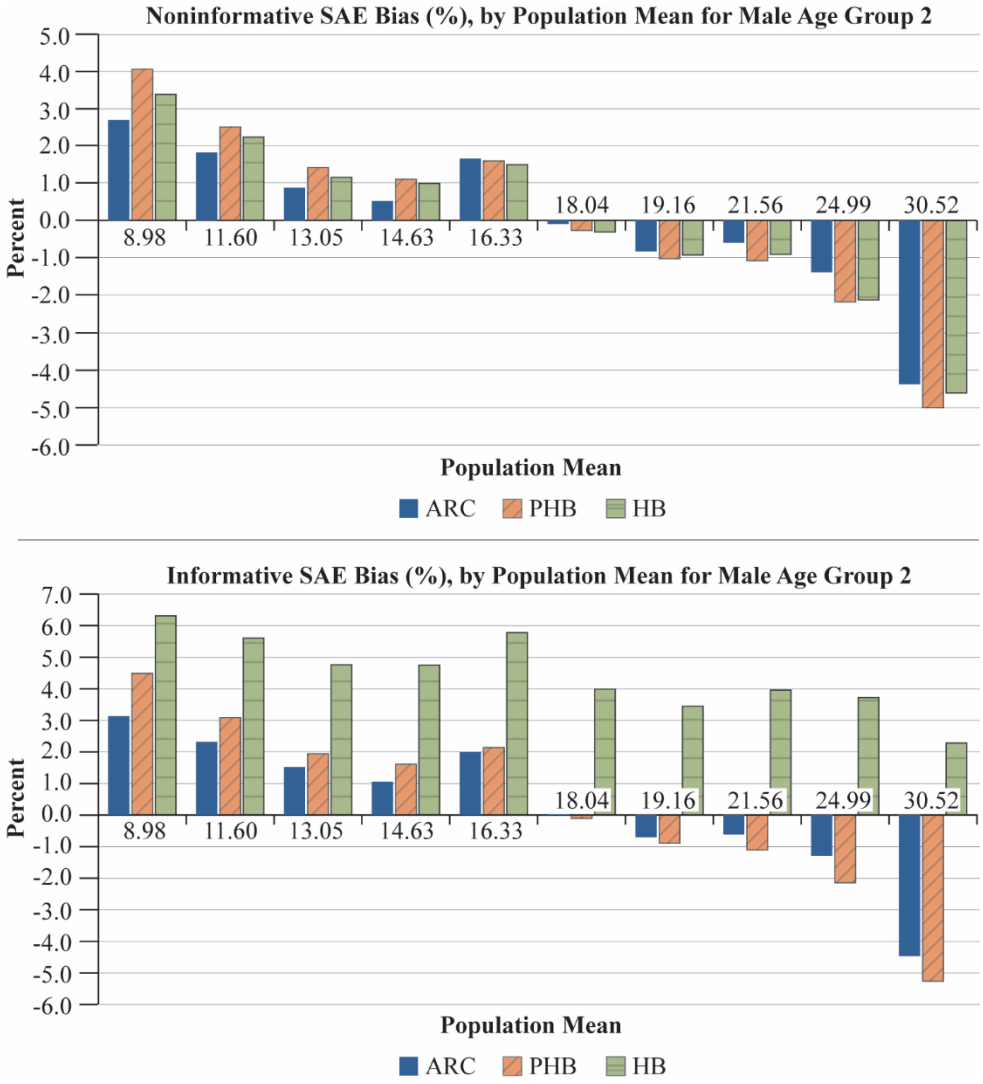

Figure 2. Noninformative and Informative Sample Small Area Estimate Biases (%), by Population Mean for the Male Age Group 2 Domain. ARC = additive random coefficient; PHB = pseudo-hierarchical Bayes; HB = hierarchical Bayes.

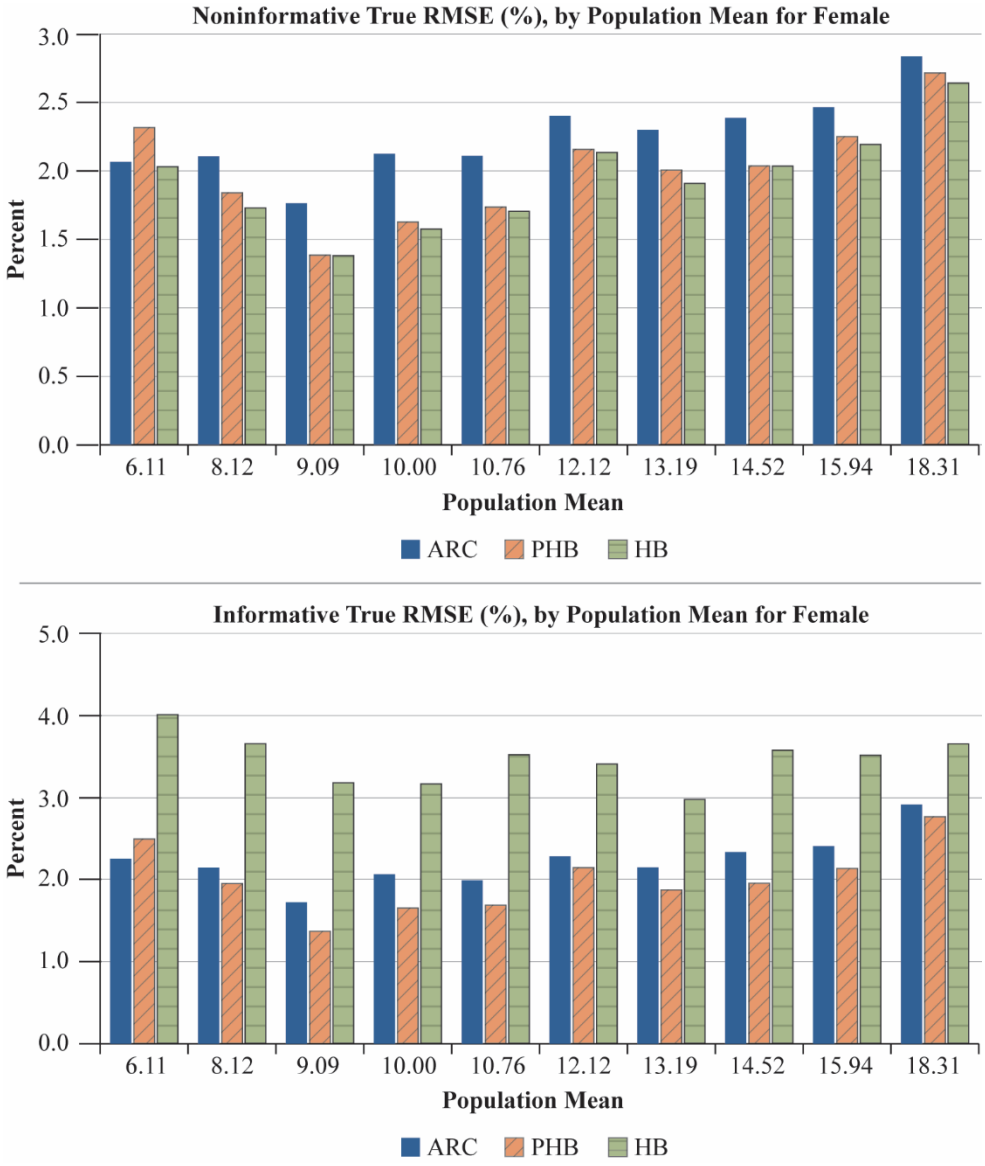

Figure 3. Noninformative and Informative Sample True Root-MSEs (RMSEs-%), by Population Mean for the Female Margin Domain. ARC = additive random coefficient; PHB = pseudo-hierarchical Bayes; HB = hierarchical Bayes.

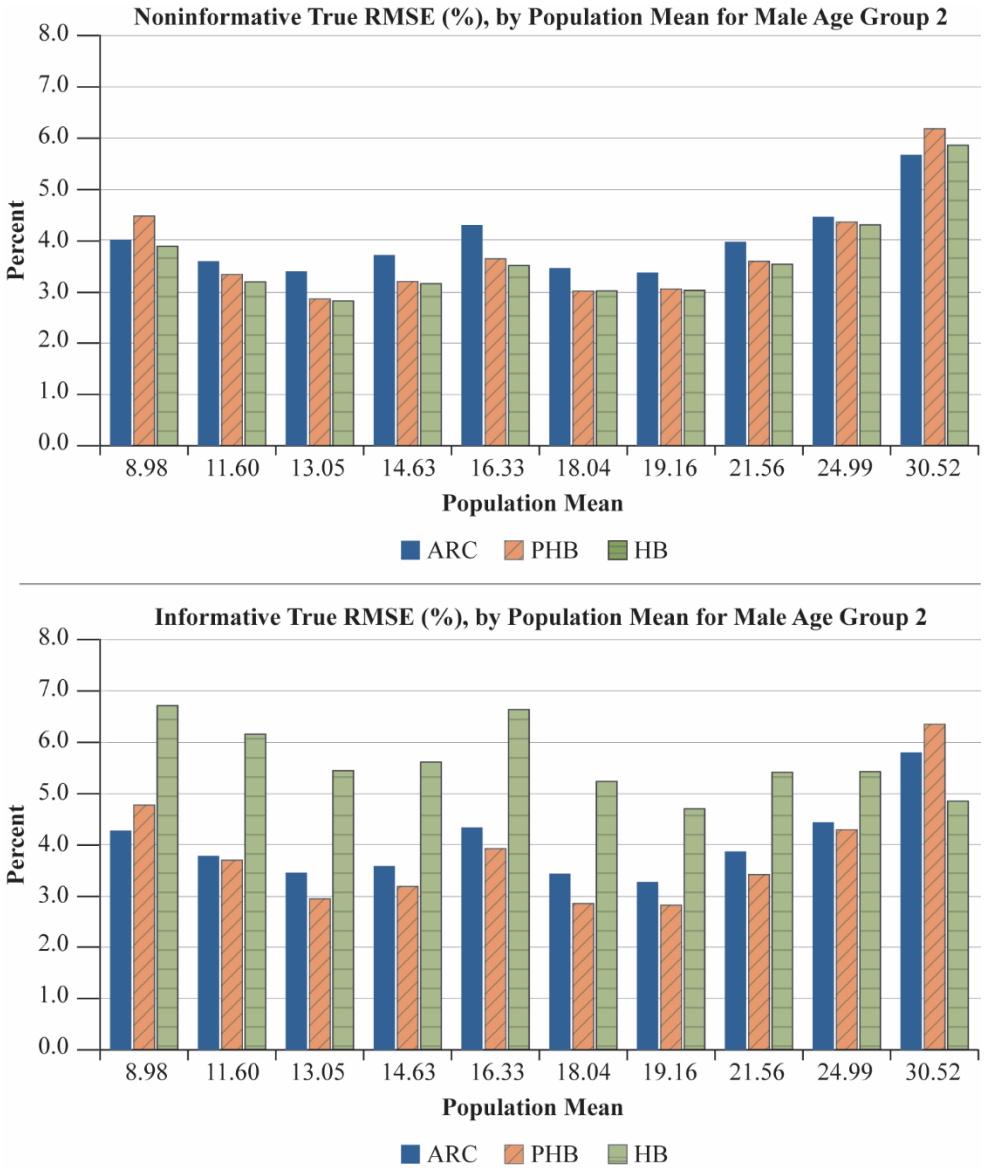

Figure 4. Noninformative and Informative Sample True Root-MSEs (RMSEs-%), by Population Mean for the Male Age Group 2 Domain. ARC = additive random coefficient; PHB = pseudo-hierarchical Bayes; HB = hierarchical Bayes.

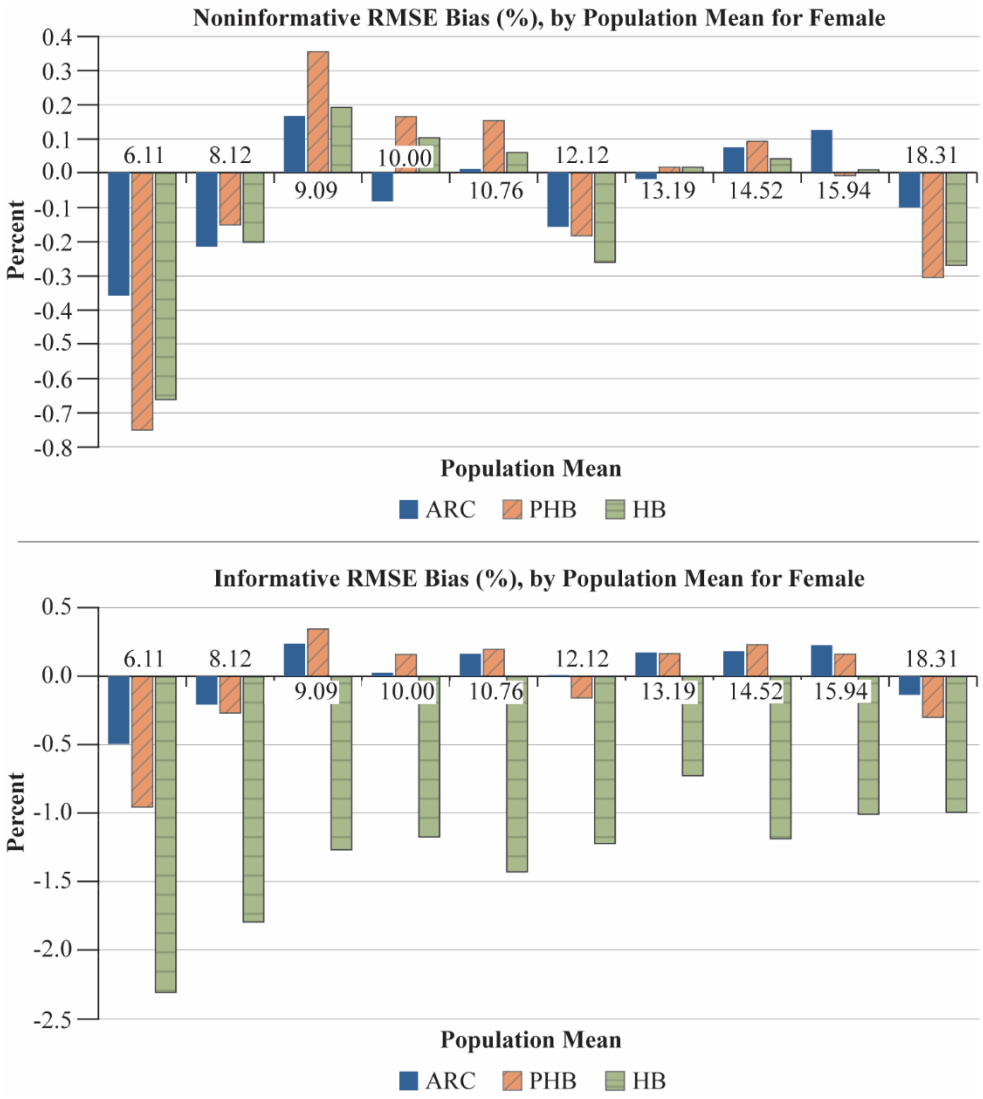

Figure 5. Noninformative and Informative Sample Estimated Root-MSE Biases (%), by Population Mean for the Female Margin Domain. ARC = additive random coefficient; PHB = pseudo-hierarchical Bayes; HB = hierarchical Bayes.

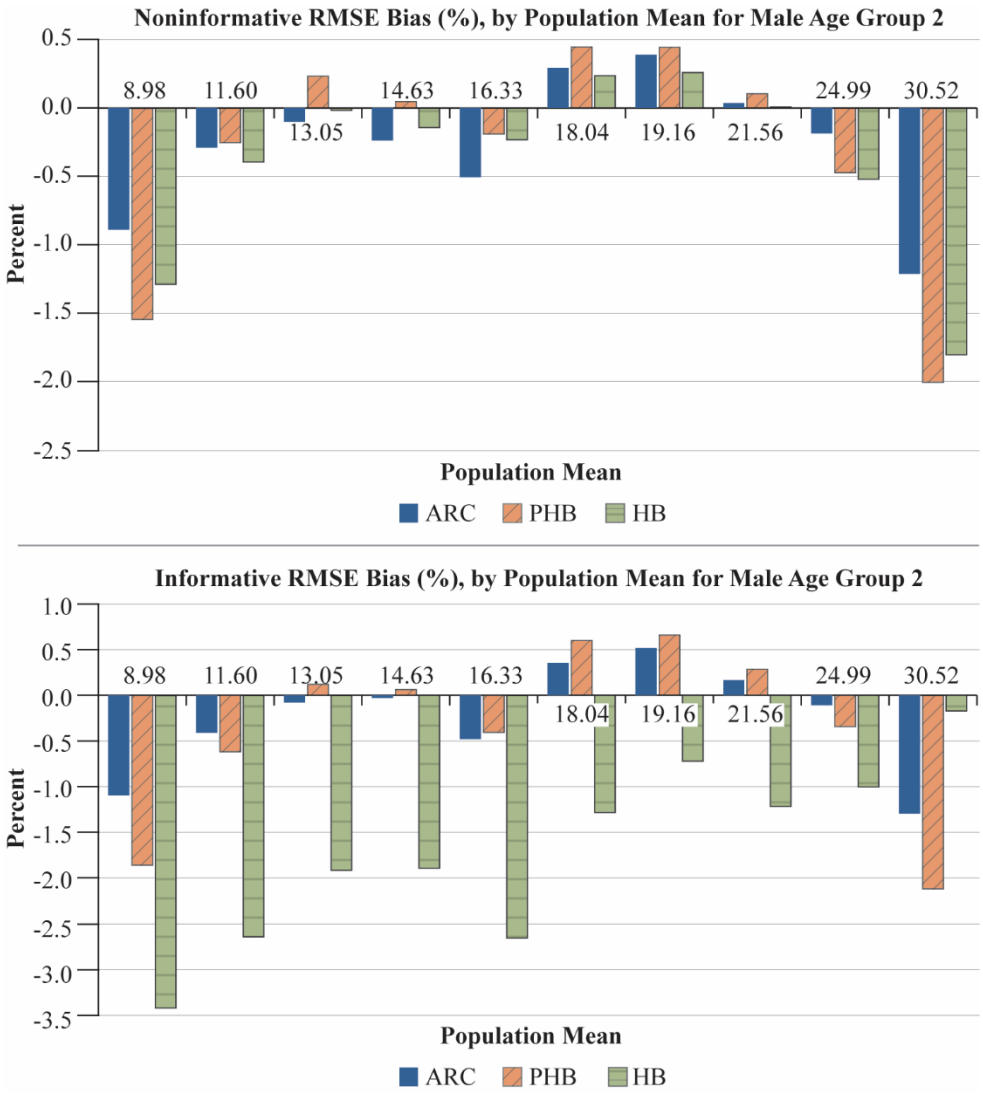

Figure 6. Noninformative and Informative Sample Estimated Root-MSE Biases (%), by Population Mean for the Male Age Group 2 Domain. ARC = additive random coefficient; PHB = pseudo-hierarchical Bayes; HB = hierarchical Bayes.

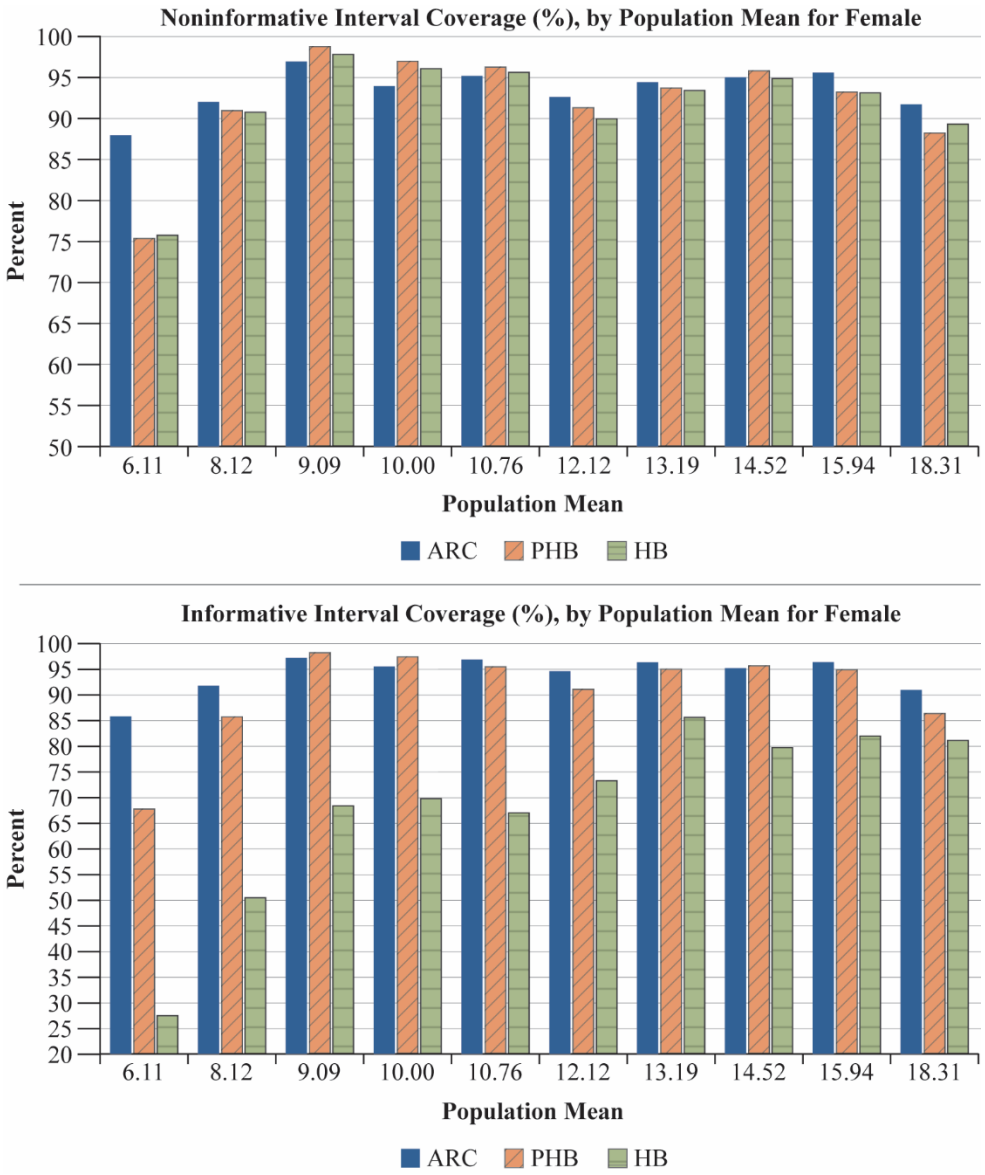

Figure 7. Noninformative and Informative Sample Interval Coverage Rates (%), by Population Mean for the Female Margin Domain. ARC = additive random coefficient; PHB = pseudo-hierarchical Bayes; HB = hierarchical Bayes.

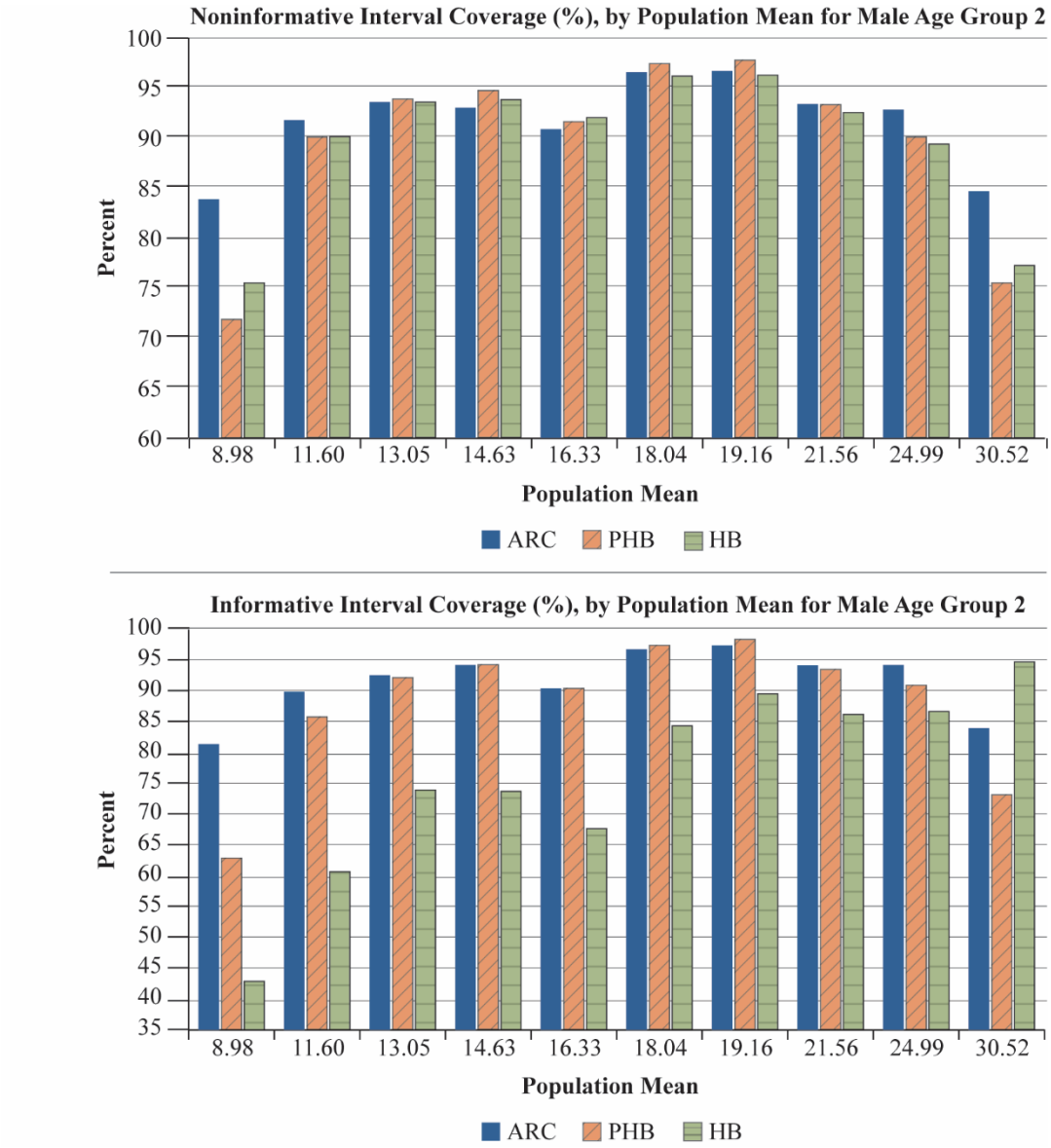

Figure 8. Noninformative and Informative Sample Interval Coverage Rates (%), by Population Mean for the Male Age Group 2 Domain. ARC = additive random coefficient; PHB = pseudo-hierarchical Bayes; HB = hierarchical Bayes.

Acknowledgments

This work was supported by RTI International (a registered trademark and a trade name of Research Triangle Institute).

Ralph Folsom Jr., former chief scientist at RTI International and ASA Fellow, passed away on December 14, 2022, in Raleigh, North Carolina. He started his professional career as a research associate and statistical consultant to the department of biostatistics, demography section, at the UNC, Chapel Hill (1966–1969), where he designed and analyzed sample surveys. He joined RTI in 1969 as a statistician and became a chief scientist in 1998. While working at RTI, he earned his PhD in biostatistics from UNC in 1984. He was a past member of the National Academy of Sciences' Panel to Evaluate the Survey of Income and Program Participation, the ASA working group to advise Census Bureau staff on the Survey of Income and Program Participation, the Board of Governors for the Panel Survey of Income Dynamics, and the Committee on National Statistics' Panel on Statistical Methods for Measuring the Group Quarter Population in the American Community Survey.

Ralph's 47-year career at RTI was filled with many innovative advancements in the field of survey data analyses. Ralph's early work on developing Taylor series standard errors for balanced effects also extended to Taylor series estimation of sampling errors for regression coefficients, which became the basis for the analysis of complex survey and other clustered data in SUDAAN® (statistical software to analyze clustered correlated data). Ralph was the first to propose using calibration weighting to adjust for unit nonresponse. He made his proposal at the 1991 ASA annual conference (ASA Proc. Soc. Statist. Sec., 197–202) before the concept of calibration weighting was formalized by Deville and Särndal (JASA, 1992). In a series of papers with Avinash Singh presented at the 2000 ASA annual conference, Ralph went on the generalize the class of calibration weighting schemes proposed by Deville and Särndal to cover their use for nonresponse and coverage-error adjustment in a more scientifically defensible manner.

In the mid-1990s, Ralph started working on developing small area estimation (SAE) methodologies to enable Substance Abuse and Mental Health Services Administration to produce reliable and cost-effective state and local area level estimates in a timely manner. He developed Survey Weighted Empirical Bayes (Folsom and Judkins, June 1997) SAE methodology for unit-level binary outcomes from complex survey data. The SWEB methodology used survey weights and worked on the same general principles as the well-known Multilevel Regression and Poststratification methodology (Gelman and Little, December 1997). Subsequently, Ralph developed the full hierarchical Bayes version of SWEB methodology and called it as the Survey Weighted Hierarchical Bayes (SWHB) methodology (Folsom, Shah, Vaish, 1999). Ralph's innovative work on SAE played a critical role in the expansion of the National Survey on Drug Use and Health (NSDUH) in 1999 from a national design to the currently implemented state stratified design. He collaborated with Babu Shah (developer of SUDAAN®) and developed a highly efficient state-of-the-art SWHB software. Since then, SWHB software is being used to produce annual state estimates and biennial substate estimates for several binary NSDUH outcome variables. The NSDUH state estimates on dependence and abuse provide the basis for calculations of treatment need by sub-state region, age, gender, and race presented in the Substance Abuse Prevention and Treatment Block Grant Application.

After developing SWHB methodology, Ralph started working on developing robust SAE methods applicable for informative samples (e.g., Singh, Folsom, and Vaish, 2002; Vaish, Folsom, and Singh, 2003; Singh, Folsom, and Vaish, 2006). He developed unit-level Additive Random Coefficient (ARC) SAE methodology for linear, logistic, and log-linear models applicable to data from informative samples (Folsom, Vaish, and Singh, 2011). Until his retirement in January 2017, he continued to make improvements to the ARC methodology, helped developed associated software, and co-authored this research manuscript.

References

Battese, G. E., R. M. Harter, and W. A. Fuller. (1988). "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28–36, doi:10.2307/2288915

Deville, J. C. and Särndal, C. E. (1992), Calibration Estimators in Survey Sampling, JASA, v. 87, No. 418, pp. 376-382.

Fay, R. E., and R. A. Herriot. (1979). "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association, 74,* 269–277, doi:10.2307/2286322

Folsom, R.E. Jr. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. ASA Proc. Soc. Statist. Sec., 197-202.

Folsom R. E., & Judkins D. R. (1997). Substance abuse in states and metropolitan areas: Model based estimates from the 1991-1993 NHSDA methodology report. Office of Applied Studies, Substance Abuse and Mental Health Services Administration, Methodological Series M-1 (DHHS Pub. No. SMA 97-3140). Rockville, MD.

Folsom, R. E., F. J. Potter, and S. R. Williams. (1987). "Notes on a Composite Size Measure for Self-Weighting Samples in Multiple Domains," Proceedings of the American Statistical Association Section on Survey Research Methods, pp. 792–796.

Folsom, R. E., B. Shah, and A. Vaish. (1999). "Substance Abuse in States: A Methodological Report on Model Based Estimates from the 1994-1996 National Household Surveys on Drug Abuse," Proceedings of the 1999 Joint Statistical Meetings, American Statistical Association, Survey Research Methods Section, Baltimore, MD, pp. 371–375.

Gelman, A., and T. C. Little. 1997. ''Postratification into Many Categories Using Hierarchical Logistic Regression.'' Survey Methodology 23:127–135.

Hidiroglou, M. A., and Y. You. (2016). "Comparison of Unit Level and Area Level Small Area Estimators," *Survey Methodology*, 42, 41–61.

Sampford, M. R. (1967). "On Sampling without Replacement with Unequal Probabilities of Selection," *Biometrika, 54*, 499-513, doi:10.2307/2335041

Singh, A. C. (2013). "A Bayesian-Frequentist Integrated Approach to Small Area Estimation," Proceedings of the 2013 Federal Committee on Statistical Methods (FCSM) Research Conference, 15 pp.

Singh, A. C., Folsom, R. E., & Vaish, A. K. (2002). A hierarchical Bayes generalization of the Fay-Herriot method to unit level nonlinear mixed models for small area estimation. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 3258–3263.

Singh, A. C., Folsom, R. E., & Vaish, A. K. (2006). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 3684–3688.

Singh, A. C., and F. Verret. (2006). "Mixed Linear Nonlinear Aggregate Level Models for Small Area Estimation from Surveys of Binary Counts," Proceedings of Statistics Canada Symposium on Methodological Issues in Measuring Population Health, Ottawa, ON, 14 pp.

Vaish, A. K., Folsom, R. E., & Singh, A. C. (2003). A simulation study to compare unit-level linear mixed model methods for small area estimation for survey data. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 4340–4344.

Folsom, R. E., Vaish, A. K., Singh A. (2011). Additive Random Coefficient (ARC) Models for Robust Small Area Estimation. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 4723–4729.

Wang, J. Y., and W. A. Fuller. (2003). "The Mean Squared Error of Small Area Predictors Constructed with Estimated Area Variances," *Journal of the American Statistical Association, 98*, 716–723, doi:10.1198/016214503000000620

Zeger, S. L., and M. R. Karim. (1991). "Generalized Linear Models with Random Effects; a Gibbs Sampling Approach," *Journal of the American Statistical Association, 86*, 79–86, doi:10.2307/2289717

Zeger, S. L., K. Y. Liang, and P. S. Albert. (1988). "Models for Longitudinal Data: A Generalized Estimating Equation Approach." *Biometrics*, *44*, 1049–1060, doi:10.2307/2531734

Appendix A

Note that $E_{S/m}(\beta_O) = \boldsymbol{\beta}_\Omega$ is the solution to

$$\boldsymbol{S}_\Omega(\boldsymbol{\beta}_\Omega) = \sum_{d=1}^{m} \sum_{k \in \Omega_d} \boldsymbol{X}_{dk}^T \left[y_{dk} - f_{dk}(\beta_\Omega) \right] = \boldsymbol{\phi}$$

and

$$Cov_m(\boldsymbol{\beta}_\Omega) = \boldsymbol{I}_\Omega(\boldsymbol{\beta})^{-1} E_m\left[\boldsymbol{S}_\Omega(\boldsymbol{\beta}) \boldsymbol{S}_\Omega(\beta)^T \right] \boldsymbol{I}_\Omega(\boldsymbol{\beta})^{-1} = C\beta_m,$$

where

$$\boldsymbol{I}_\Omega(\boldsymbol{\beta}) = -\left[\partial \boldsymbol{S}_\Omega(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} \right] = \left[\sum_{d=1}^{m} \sum_{k \in \Omega_d} \partial f_{dk}(\boldsymbol{\beta}) \boldsymbol{X}_{dk}^T \boldsymbol{X}_{dk} \right].$$

Under the additive random coefficient (ARC) model,

$$\boldsymbol{S}_\Omega(\boldsymbol{\beta}) = \sum_{d=1}^{m} \boldsymbol{M}_{\Omega d}(\boldsymbol{\beta}) \boldsymbol{\eta}_d + \sum_{d=1}^{m} \sum_{k \in \Omega_d} \boldsymbol{X}_{dk}^T (y_{dk} - \mu_{dk}),$$

with $\boldsymbol{M}_{\Omega d}(\boldsymbol{\beta}) = \sum_{k \in \Omega_d} \partial f_{dk}(\boldsymbol{\beta}) \boldsymbol{X}_{dk}^T \boldsymbol{Z}_{dk}$. If $\nu_{dk} \equiv E_m (y_{dk} - \mu_{dk})^2$, then we have

$$E_m\left[\boldsymbol{S}_\Omega(\boldsymbol{\beta}) \boldsymbol{S}_\Omega(\boldsymbol{\beta})^T \right] = \sum_{d=1}^{m} \boldsymbol{M}_{\Omega d}(\boldsymbol{\beta}) \boldsymbol{\Sigma}_\eta \boldsymbol{M}_{\Omega d}(\boldsymbol{\beta})^T + \sum_{d=1}^{m} \sum_{k \in \Omega_d} \boldsymbol{X}_{dk} \underset{\sim}{X}_{dk} \nu_{dk},$$

and therefore

$$Cov_m E_{S/m}(\boldsymbol{\beta}_O) = \boldsymbol{I}_\Omega(\boldsymbol{\beta})^{-1} \left[\sum_{d=1}^{m} \boldsymbol{M}_{\Omega d}(\boldsymbol{\beta}) \boldsymbol{\Sigma}_\eta \boldsymbol{M}_{\Omega d}(\boldsymbol{\beta})^T \right] \boldsymbol{I}_\Omega(\boldsymbol{\beta})^{-1}, \quad \text{(A.1)}$$

ignoring terms of large order $\left[N_+ = \sum_{d=1}^{m} N_d \right]^{-1}$. Now, we can recast

$$Cov_m E_{S/m}(\boldsymbol{\beta}_O) = \bar{\boldsymbol{I}}_\Omega^{-1} \left[\sum_{d=1}^{m} N_d^2 \left(\bar{\boldsymbol{M}}_{\Omega d} \boldsymbol{\Sigma}_\eta \bar{\boldsymbol{M}}_{\Omega d}^T \right) \right] \bar{\boldsymbol{I}}_\Omega^{-1} / N^2$$

$$= \frac{\left(1 + CVN^2\right)}{m} \left\{ \bar{\boldsymbol{I}}_\Omega^{-1} \left[\sum_{d=1}^{m} \zeta_d \left(\bar{\boldsymbol{M}}_{\Omega d} \boldsymbol{\Sigma}_\eta \bar{\boldsymbol{M}}_{\Omega d}^T \right) \right] \bar{\boldsymbol{I}}_\Omega^{-1} \right\},$$

where $\zeta_d = \left[N_d^2 / \left(\sum_{d'-1}^{m} N_{d'}^2 \right) \right]$. Therefore, $\boldsymbol{C\beta}_m$ is of order $(1/m)$. We estimate $\boldsymbol{C\beta}_m$ given $\boldsymbol{\beta}_o$ and $\hat{\boldsymbol{\Sigma}}_\eta$ as

$$\boldsymbol{C}\hat{\boldsymbol{\beta}}_m = \left[\boldsymbol{I}_s(\boldsymbol{\beta}_o) \right]^{-1} \left[\sum_{d=1}^{m} \boldsymbol{M}_{\Omega d}(\boldsymbol{\beta}_o)\, \hat{\boldsymbol{\Sigma}}_\eta \boldsymbol{M}_{\Omega d}^T \right] \left[\boldsymbol{I}_s(\boldsymbol{\beta}_o) \right]^{-1},$$

where $\left[\boldsymbol{I}_s(\boldsymbol{\beta}_o) \right]^{-1} = \left[\sum_{d=1}^{m} \sum_{k \in s} \partial f_{dk}(\boldsymbol{\beta}_o) \boldsymbol{X}_{dk}^T \boldsymbol{X}_{dk} \right]^{-1}$ is also used to form $\boldsymbol{C\beta}_o$, the sampling covariance matrix.