Additive Random Coefficient (ARC) Models for Robust Small Area Estimation

Ralph E. Folsom[1], Akhil K. Vaish[1], Avinash Singh[2]

[1]RTI International, 3040 Cornwallis Road, RTP, 27709
[2]NORC at the University of Chicago, Chicago, IL 60603

Abstract

Unit or person-level ARC models with linear, logistic, and log-linear marginal mean functions are developed for small area estimation. ARC models take the form of a first order Taylor series approximation to the associated general linear mixed model. The area-level random coefficient vectors specify effects for demographic groups. Protection against nonignorable sample designs is provided by a hybrid solution that combines the marginal [probability (P) sampling plus ARC model (ξ)] distribution of the fixed regression coefficients with the MCMC simulated Bayes posterior distributions for the small area specific random coefficient vectors. Survey weighted estimating equations are employed in the solution for the fixed and random coefficients along with sample design consistent covariance matrix estimators. A generalized design effect matrix is used to stabilize the area-level covariance matrices for the random coefficients. A simulation study for the logistic ARC model contrasts the new method's performance with a nonlinear version of You and Rao's (2003) pseudo hierarchical Bayes solution that discounts the effect of nonignorable samples on the mean squared errors of small area estimates.

Key Words: small area estimation; generalized design effects; nonignorable survey sample design; general linear mixed model; additive random coefficient; survey weighted estimating equations

1. Introduction

The Additive Random Coefficient **(**ARC) models utilized here for small area estimation (SAE) are versions of the generalized liner mixed model (GLMM) with additive random coefficients. ARC models can be derived as a first order Taylor series approximation to a GLMM. Our unit-level ARC model is an extension of Singh and Verret's (2006) aggregate level GLMARC model. Small area estimators based on unit level models have the potential to be more precise than aggregate model estimators. This potential derives from the unit model's use of person and neighbourhood level fixed predictors and main effect type random coefficients for demographic groups. A major impediment to using unit level models for SAE has been the difficulty in fully accounting for complex nonignorable sample designs. We provide such a solution here based on the unit level ARC model where

$$E(y_{dk} \mid \underset{\sim}{\eta}_d) = f(\underset{\sim}{X}_{dk}\underset{\sim}{\beta}) + [\partial f(\underset{\sim}{X}_{dk}\underset{\sim}{\beta})]\underset{\sim}{Z}_{dk}\underset{\sim}{\eta}_d \tag{1}$$

for $d = 1, \cdots, m$ small areas and $k = 1, \cdots, n_d$ area-d respondents. In this model $E(y \mid \underset{\sim}{\eta})$ equals a nonlinear marginal mean function $f(X\beta)$ and an additive random effect $\underset{\sim}{\eta}$ contribution with a derivative multiplier ∂f .

For the logistic ARC model

$$\text{Prob}(y_{dk} = 1 \mid \underset{\sim}{\eta}_d) = f_{dk} + \partial f_{dk} \underset{\sim}{Z}_{dk} \underset{\sim}{\eta}_d \tag{2}$$

where $f_{dk} = [1 + \exp(-\underset{\sim}{X}_{dk} \underset{\sim}{\beta})]^{-1}$ and $\partial f_{dk} = f_{dk}(1 - f_{dk})$. A typical Z vector takes the form $\underset{\sim}{Z}_{dk} = (1, g_{dk}, \underset{\sim}{a}_{dk}, \underset{\sim}{r}_{dk})$ with g denoting an indicator for male gender, $\underset{\sim}{a}$ specifying a vector of age group indictors, and $\underset{\sim}{r}$ containing indicator variables for race/ethnicity groups. We assume that the random effect vectors $\underset{\sim}{\eta}_d$ are q variate i.i.d. normal with zero mean and general covariance matrix Σ_η .

The reason we favor the ARC model over GLMM is the significant computational advantage it has for achieving our SAE goals. Our SAE goal is to produce point estimates and mean squared errors (MSEs) that account fully for complex nonignorable sample designs. Existing solutions that account for complex sample features as regards the fixed model parameters $\underset{\sim}{\beta}$ and Σ_η tend to discount the effect of nonignorable design on the MSEs of the random effects. The ARC model's two key computational advantages are that:

- The marginal means of the y_{dk} have the <u>fixed</u> model form f_{dk} . This makes it easier to estimate the fixed β coefficients.
- Solutions for the random effect vectors $\underset{\sim}{\eta}_d$ do not require Newton-Raphson iterations.

2. Survey Weighted Estimating Functions

To estimate the β parameters we first considered the survey weighted pseudo-optimum estimating functions

$$S_w(\underset{\sim}{\beta} \mid \Sigma_\eta) = \sum_{d=1}^{m} \sum_{k=1}^{n_d} w_{dk} (\partial f_{dk} \div v_{dk}) \underset{\sim}{X}'_{dk} [y_{dk} - f_{dk}(\underset{\sim}{\beta}) - \partial f_{dk}(\underset{\sim}{\beta}) \underset{\sim}{Z}_{dk} \hat{\underset{\sim}{\eta}}_{wd}(\underset{\sim}{\beta}, \Sigma_\eta)] \tag{3}$$

where $v_{dk} = E_{\underset{\sim}{\eta}_d} \text{var}(y_{dk} \mid \underset{\sim}{\eta}_d)$ and $(\partial f_{dk} \div v_{dk}) = [1 - \partial f_{dk}(\underset{\sim}{Z}_{dk} \Sigma_\eta \underset{\sim}{Z}'_{dk})]^{-1}$ for the logistic ARC model. We used an optimum GEE type solution to derive the pure ARC model version of these equations and then inserted survey weights w_{dk} to protect against bias from nonignorable sample designs. Equation (3) includes ARC model weighting factors involving ratios of the ∂f_{dk} derivatives and the conditional variances v_{dk} of

$y_{dk} \mid \underset{\sim}{\eta}_d$. For continuous data linear ARC models and count data Poisson/Exponential ARC models these ratios are constants. While the $(\partial f_{dk} \div v_{dk})$ ratios are not constant for the logistic model, we chose to discard them anyway in favor of the survey weighted data fitting equations. To further reduce the computational burden, we have chosen to use the consistent but less efficient GEE or SUDAAN type estimating equations that do not include the random effect $\hat{\underset{\sim}{\eta}}_{wd}(\underset{\sim}{\beta}, \Sigma_\eta)$ residual corrections. In the second generation of our software design, we plan to provide an option that implements the more efficient version of equation (3) with the $\hat{\underset{\sim}{\eta}}_{wd}(\underset{\sim}{\beta}, \Sigma_\eta)$ residual corrections.

Turning to the survey weighted estimates of the random effect vector, we first define the column vectors $\underset{\sim}{\xi}_{dk} \equiv \underset{\sim}{Z}'_{dk}[y_{dk} - f_{dk}(\underset{\sim}{\beta})]$ for a given $\underset{\sim}{\beta}$. We then compute area-*d* level survey weighted total vectors $\underset{\sim}{\xi}_{wd} = \sum_{k=1}^{n_d} w_{dk} \underset{\sim}{\xi}_{dk}$. Taking the dual expectation of $\underset{\sim}{\xi}_{wd}$ first over the probability sample-*s* given the data-*y* and then over the superpopulation ARC model for *y* given $\underset{\sim}{\eta}_d$, the resulting expected value is $\Delta_{\Omega d} \underset{\sim}{\eta}_d$ where $\Delta_{\Omega d}$ is the universe level matrix

$$\Delta_{\Omega d} \equiv \sum_{k=1}^{N_d} \partial f_{dk}(\beta) \underset{\sim}{Z}'_{dk} \underset{\sim}{Z}_{dk}. \tag{4}$$

Computing $\Delta_{\Omega d}$ from the universe file for a given $\underset{\sim}{\beta}$ we then use $\Delta_{\Omega d}^{-1}$ to form $\underset{\sim}{\tau}_{wd} \equiv \Delta_{\Omega d}^{-1} \underset{\sim}{\xi}_{wd}$ which is an unbiased estimate for the $\underset{\sim}{\eta}_d$ realization.

To account further for complex nonignorable sample designs we employ a stratified pps with replacement cluster sample estimator for the variance-covariance matrix of $\underset{\sim}{\xi}_{wd}$, say $C_{\xi d}$. The preferred sample designs for SAE are these where the target small areas are design strata. We assume here that the small areas are either design strata or geographic domains comprising parts of one or more design strata. For the latter case, we chose to ignore any cross area sampling covariances between $\underset{\sim}{\xi}_{wd}$ and $\underset{\sim}{\xi}_{wd'}$. Given the $C_{\xi d}$ matrices we then use a design effect matrix averaged over the *m* small areas to stabilize $C_{\xi d}$ which will often be based on relatively few clusters. With $\bar{C}_{\xi d}$ denoting the stabilized covariance matrix for $\underset{\sim}{\xi}_{wd}$, the sampling covariance matrix for $\underset{\sim}{\tau}_{wd}$ is

$$\mathrm{cov}_{s|y}(\underset{\sim}{\tau}_{wd} \mid \underset{\sim}{\eta}_d) = \Delta_{\Omega d}^{-1} \bar{C}_{\xi d} \; \Delta_{\Omega d}^{-1} \equiv \bar{C}_{\tau d}. \tag{5}$$

Given this stabilized version of $C_{\tau d}$, the correct sample design based shrinkage matrix for predicting $\underset{\sim}{\eta}_d$ is $\gamma_{wd} = \Sigma_\eta (\Sigma_\eta + \bar{C}_{\tau d})^{-1}$. This leads to the $\underset{\sim}{\eta}_d$ predictor $\hat{\underset{\sim}{\eta}}_{wd}(\underset{\sim}{\beta}, \Sigma_\eta) = \gamma_{wd}\, \underset{\sim}{\tau}_{wd}$ with the mean squared prediction error

$mspe(\hat{\underset{\sim}{\eta}}_{wd} \mid \underset{\sim}{\beta}, \Sigma_\eta) = \gamma_{wd}\bar{C}_{\tau d}$. Assuming that $\underset{\sim}{\tau}_{wd}$ has a joint distribution over the sample and the ARC model that is *q*-variate normal with mean vector $\underset{\sim}{\eta}_d$ and covariance matrix consistently estimated by $\bar{C}_{\tau d}$, then the conditional posterior distribution of $\underset{\sim}{\eta}_d \mid \underset{\sim}{\tau}_{wd}$ is normal with mean vector $\gamma_{wd}\ \underset{\sim}{\tau}_{wd}$ and covariance matrix $\gamma_{wd}\bar{C}_{\tau d}$.

3. The ARC Model Small Area Estimates

Given $\underset{\sim}{\beta}, \Sigma_\eta$ and our predictor for $\underset{\sim}{\eta}_d$ we can form the ARC model small area estimates for the area-*d* vector $\hat{\underset{\sim}{T}}_{\Omega d}$ of demographic domain totals. Assuming that the area level sampling fractions $(n_d \div N_d)$ are all negligible, the area-*d* demographic domain totals are predicted by

$$\begin{aligned}\hat{\underset{\sim}{T}}_{\Omega d} &= \sum_{k=1}^{N_d} \underset{\sim}{Z}'_{dk}[\, f_{dk}(\underset{\sim}{\beta}) + \partial f_{dk}\underset{\sim}{Z}_{dk}\hat{\underset{\sim}{\eta}}_{wd}(\underset{\sim}{\beta}, \Sigma_\eta)] \\ &= \Im_{\Omega d} + \Delta_{\Omega d}\hat{\underset{\sim}{\eta}}_{wd}.\end{aligned} \tag{6}$$

The vector $\Im_{\Omega d}$ contains the universe level domain totals of the f_{dk} fixed marginal means. With $\Im_{wd}$ denoting the w_{dk} weighted sample total of $\underset{\sim}{Z}'_{dk} f_{dk}$ and Υ_{wd} depicting the corresponding weighted sample total of $\underset{\sim}{Z}'_{dk}\, y_{dk}$ the vector of area-*d* small area estimates has the following composite form

$$\hat{\underset{\sim}{T}}_{\Omega d} = (I - G_d)\Im_{\Omega d} + G_d[\Upsilon_{wd} - (\Im_{wd} - \Im_{\Omega d})] \tag{7}$$

with $G_d \equiv (\Delta_{\Omega d}\Sigma_\eta\Delta_{\Omega d})[(\Delta_{\Omega d}\Sigma_\eta\Delta_{\Omega d}) + \bar{C}_{\xi d}]^{-1}$. Note that the compositing matrix G_d has a shrinkage form incorporating the $\Delta_{\Omega d}$ matrices. Note also that $[\Upsilon_{wd} - (\Im_{wd} - \Im_{\Omega d})]$ is a vector valued nonlinear survey regression estimator for the domain totals. The conditional posterior covariance matrix for $\hat{\underset{\sim}{T}}_{\Omega d}$ is $G_d\bar{C}_{\xi d}$.

4. MCMC Posterior Variance Steps

To estimate the random effects covariance matrix Σ_η we use a hierarchical Bayes solution based on an inverse Wishart prior with (*q*+2) degrees of freedom and prior mean matrix $\Sigma_{\eta 0}$. The details are sketched in the MCMC steps outlined here:

1. Use SUDAAN to obtain β_0 and its covariance matrix $C_{\beta 0}$.
2. Sample $t = 1, \cdots, K$ vectors $\underset{\sim}{\beta}_t \sim N(\underset{\sim}{\beta}_0, C_{\beta 0})$.
3. Access the universe file for area-*d* and compute $\Im_{\Omega d}(\underset{\sim}{\beta}_t)$ and $\Delta_{\Omega d}(\underset{\sim}{\beta}_t)$.

4. Given $\Sigma_{\eta(t-1)}$ form $\hat{\underset{\sim}{T}}_{\Omega dt}$ and its covariance matrix $G_d \overline{C}_{\xi d}$.
5. Sample $\underset{\sim}{\eta}_{dt}(iid) \sim N(\gamma_{wdt}\underset{\sim}{\tau}_{wdt}, \gamma_{wdt}\overline{C}_{\tau dt})$.
6. Form $A_t = (\sum_{d=1}^{m} \underset{\sim}{\eta}_{dt}\underset{\sim}{\eta}'_{dt})$ and draw $\Sigma_{\eta t} \sim W^{-1}[m+q+2,(A_t+\Sigma_{\eta 0})]$.
7. Use the Rao-Blackwell formula to compute the covariance matrix of $\overline{\underset{\sim}{T}}_{\Omega d} = \sum_{t=1}^{K} \hat{\underset{\sim}{T}}_{\Omega dt} \div K$.

5. Nonignorable Sample Simulation

For our nonignorable sample simulation we generated area-*d* (*d*=100 areas) populations of $N_d \sim 6,000$ binary observations using a $N(0,1)$ latent variable and unequal fractions of the conditional means μ_{dk} as the tail probabilities. Specifically, with $e_{dk} \sim N(0,1)$

$$
\begin{aligned}
y_{dk} &= 1 \ if \ e_{dk} \geq \Phi^{-1}[1-0.75\mu_{dk}] \ or \ e_{dk} \leq \Phi^{-1}[0.25\mu_{dk}] \\
&= 0 \ otherwise
\end{aligned}
$$

where $\mu_{dk} = \mathrm{Prob}(y_{dk} = 1 | \underset{\sim}{\zeta}_d)$ and $\Phi(\,)$ is the $N(0,1)$ cumulative distribution function. We specified the μ_{dk} using a logistic mixed model

$$
\log[\mu_{dk} \div (1-\mu_{dk})] = \underset{\sim}{X}_{dk}\underset{\sim}{\alpha} + \underset{\sim}{Z}_{dk}\underset{\sim}{\zeta}_d \tag{8}
$$

with $\underset{\sim}{X}_{dk} = (1, a_{dk}, r_{dk}, x_{dk})$ and $\underset{\sim}{Z}_{dk} = (1, a_{dk}, r_{dk})$ where a_{dk} and r_{dk} denote 1/0 indicators for binary age and race groups; the continuous covariate was generated using $x_{dk} = u_d + \varepsilon_{dk}$ with $u_d \sim N(0,1)$ and $\varepsilon_{dk} \sim N(0,1)$; and $\underset{\sim}{\zeta}_d \sim N(\underset{\sim}{0}_3, \Sigma_\zeta)$.

We set the fixed coefficients $\underset{\sim}{\alpha}$ and Σ_ζ to generate a wide range of small area domain proportions and then drew 160 sample records from each area-*d* with 120 drawn via srs from stratum Ω_{d+} where the latent variables $e_{dk} \geq 0$ and 40 from stratum Ω_{d-} where $e_{dk} < 0$. We generated such 200 populations with the same fixed predictors $\underset{\sim}{X}_{dk}$ and domain indicators $\underset{\sim}{Z}_{dk}$ but new random effects $\underset{\sim}{\zeta}_d$ and new latent variables e_{dk}.

As a competitor for the ARC model we implemented a nonlinear version of You and Rao's (2003) pseudo-hierarchical Bayes (PHB) solution. The PHB solution assumes the sample is ignorable for posterior variance estimation.

6. Simulation Results

To present the results we focus on one of the two age groups where the sample sizes per area were ~80. Results for the other demographic domains were similar. For comparison the small area totals for age group 1 were converted to percentages. For both the PHB and ARC solutions there were 100 area percentage estimates for each of 200 populations yielding a total of 20,000 estimates for each method. To examine the performance as a function of the true finite population percentages we formed 20 groups of 1000 area by population combinations using the true finite population percents to rank the 20,000 combinations. Figure 1 graphs the average bias calculations for the 20 groups. Both methods have fairly linear bias plots that exhibit some over shrinkage on the low and high ends. The ARC model fares a bit better in this regard.

Figure 1. Bias: ARC vs. PHB

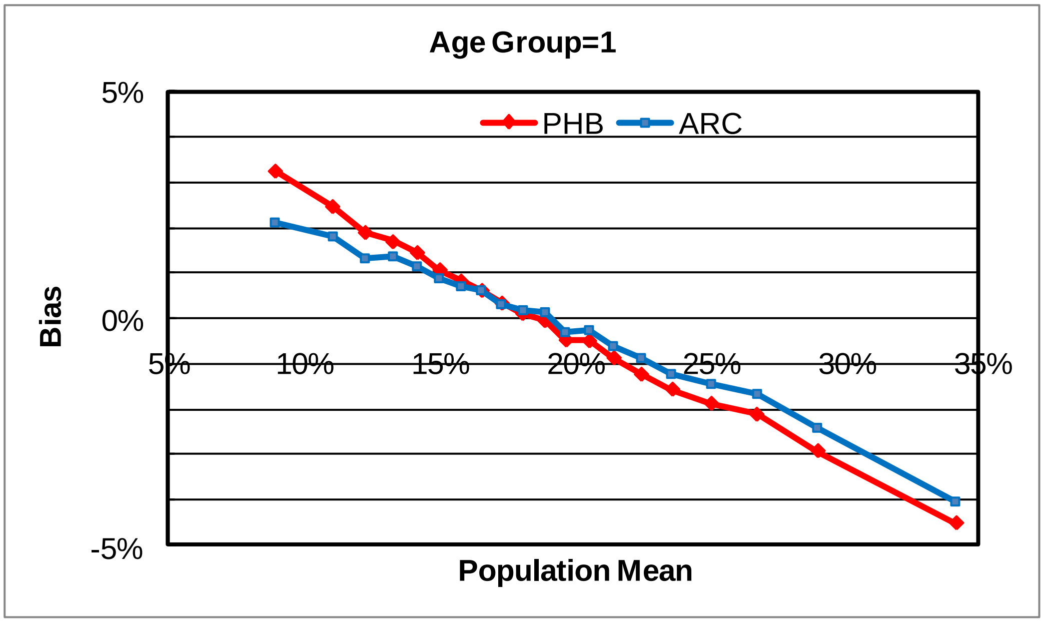
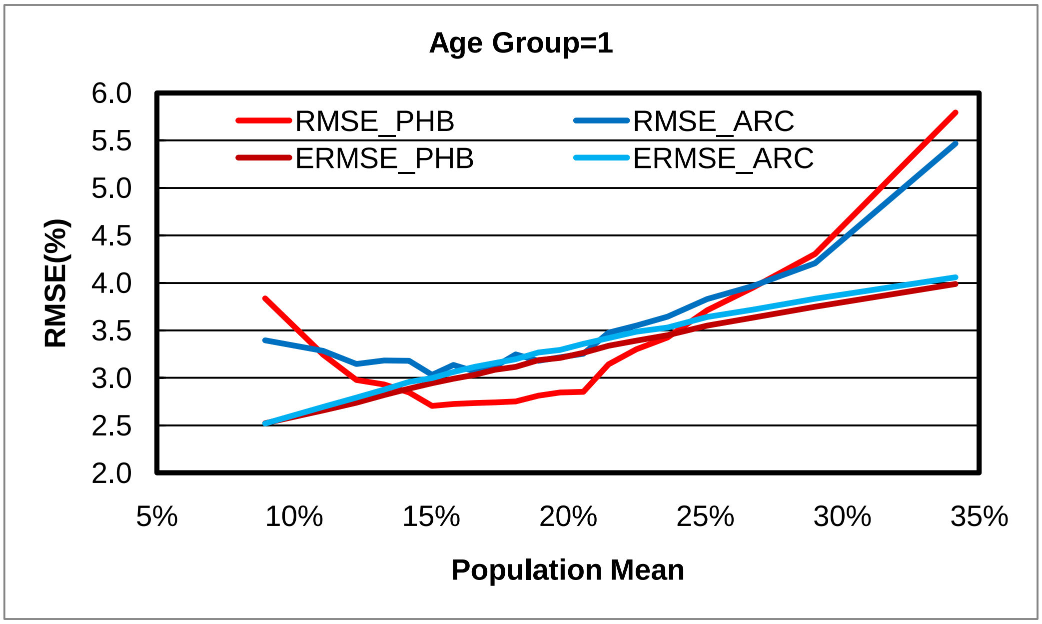


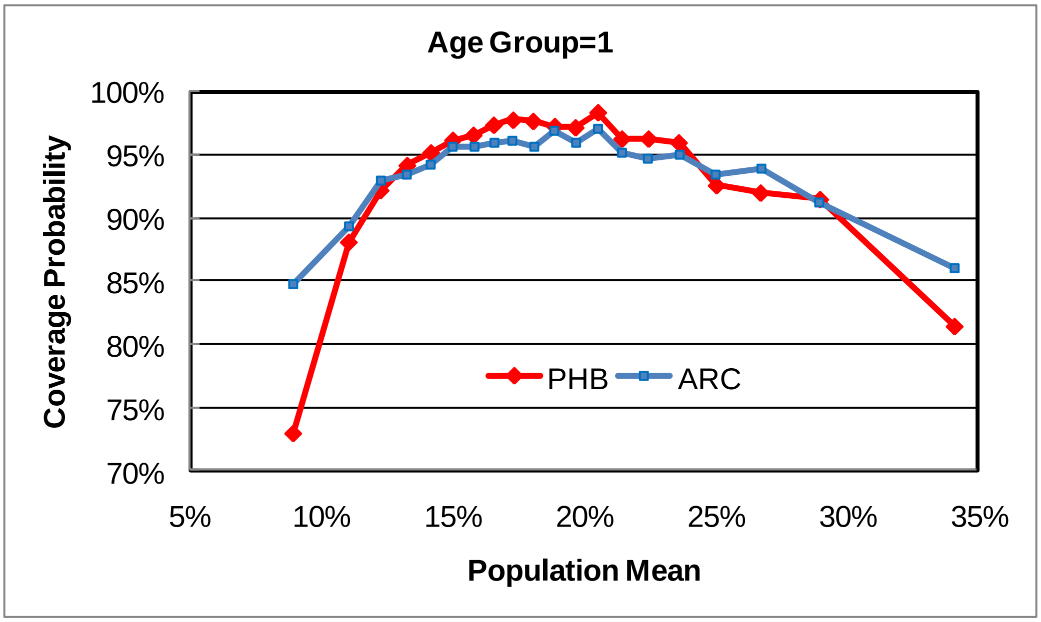
Figure 2 shows the true and the estimated root mean squared errors (RMSEs) by group. The two parabolic curves are the true RMSEs. The two linear curves are the estimated RMSEs. Both methods underestimates the true RMSE on the ends but the ARC estimate tracks the true values better in the mid-range.

Figure 2. RMSE (%): ARC vs. PHB



These MSE results are reflected in the interval coverage probabilities plotted in Figure 3. Both coverages drops off considerably in the outer groups with the PHB drops more pronounced. In the mid-range the ARC method stays closer to the desired coverage level.

Figure 3. 95% Coverage Probability: ARC vs. PHB



7. Future Developments

For a more challenging simulation we plan to select nonignorable cluster samples. We will also consider the more efficient estimating equations for β coefficients that are conditioned on the estimated random effect vectors.

References

Singh, A.C., and Verret. F. (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.

You, Y. and J. N. K. Rao (2003). Pseudo Hierarchical Bayes Small Area Estimation Combining Unit Level Models and Survey Weights, Journal of Statistical Planning and Inference, 111, 197-208.

Contact Information

Ralph Folsom: ref@rti.org
Akhil Vaish: avaish@rti.org
Avinash Singh: singh-avi@norc.org