Robust Small Area Estimation Using a Mixture Model

Gershunskaya, Julie
U.S. Bureau of Labor Statistics
2 Massachusetts Ave NE, Suite 4985
Washington, DC, 20212, USA
E-mail: Gershunskaya.Julie@bls.gov

Lahiri, Parthasarathi
JPSM
1208 Lefrak Hall, University of Maryland
College Park, MD 20742, USA
E-mail: plahiri@survey.umd.edu

1. Introduction

Small area estimation (SAE) generally relies on either implicit or explicit modeling assumptions. It may happen that a relatively few observations do not fit the model that adequately explains bulk of the data. Such observations may adversely affect estimation of the model parameters. In the context of the well-celebrated Fay-Herriot area level normality-based model (see Fay and Herriot, 1979), Datta and Lahiri (1995) noted that even in the presence of one unusually large direct estimate, empirical Bayes estimates for all small areas collapse to the corresponding direct estimates and thereby loose the benefit of borrowing strength from relevant sources. To deal with such influential small areas, they proposed a hierarchical Bayes method based on a scale mixture of normal prior distribution, which has a heavier tail than the traditional normal prior. One problem with the area level model is that it assumes the sampling variances to be known although they are estimated by certain variance smoothing techniques such as the Generalized Variance Function (GVF) method (see Hawala and Lahiri 2010 and the references therein) and the model does not incorporate the variability due to the estimation of these sampling variances. To get around this problem, one may use unit level model such as the one proposed by Battese, Harter and Fuller (1988). Since a unit level model typically requires modeling of a large dataset, it is common to encounter some unusual observations. Several methods that are resistant to influential observations have been proposed in the SAE literature in recent years (Chambers and Tzavidis 2006**,** Sinha and Rao 2008, Gershunskaya, 2010).

Influential observations may suggest a real finite population structure that is not described by the assumed base model. Such influential observations or representative outliers (using Chambers' 1986 terminology) carry important information and it would be unwise to ignore it and rely only on the base model. In a non-SAE setting, Chambers (1986) proposed a bias correction to the initial estimator, where the initial estimator is based firmly on the assumed working model while the bias correction is an estimated mean of residuals after relaxing the modeling assumptions. In a SAE application, one may add area specific bias correction term to the initial predictor for small area parameter, a method explored by Chambers *et al.* (2009). The drawback of such adaptation of the non-SAE methodology is that inevitably the estimation of the bias correction terms for small areas would be based on small samples, potentially leading to inefficient estimates.

The approach proposed in the present paper is a slight modification of a classical linear mixed model application to SAE. The underlying distribution is a scale mixture of two normal distributions. This model explicitly describes the behaviour of the influential observations relative

to the other units; thus, it automatically produces estimates (e.g., using MLE) that account for influential observations.

A simple formulation of the mixture model used in this paper may still be too strong in certain assumptions about the distribution of influential observations. First, the outliers are assumed to appear randomly across areas. However, the outliers may be clustered in certain areas. This may lead to bias in the prediction of the area-level random effects. Gershunskaya (2010) proposed an area-level bias correction method that is different from the one of Chambers *et al.* (2009) and attempted to preserve the efficiency of the initial model by introducing the corrections only to select areas, after these areas have been tested on possible outlyingness. One could alternatively explore the possibility of using a heavy tailed distribution for the random effects such as the ones suggested by Datta and Lahiri (1995). Another potentially incorrect assumption is that the influential observations are distributed symmetrically around a common mean. Failure of this assumption may lead to an overall bias across areas. The overall bias correction (OBC) can be based on the data combined from all areas, thus the initial modeling assumptions can be more safely relaxed to estimate the correction at this higher level.

In Section 2, we briefly review several existing approaches to outlier resistant SAE. The proposed approach is detailed in Section 3. Section 4 contains results of a simulation study that compares several methods of robust SAE. We consider two applications of the proposed methodology in Section 5. Our first application concerns estimation of monthly employment changes in the metropolitan statistical areas (MSA) in the Current Employment Statistics (CES) Survey conducted by the U.S. Bureau of Labor Statistics (BLS). The second application relates to county level estimation of crop yield, which is of interest to the National Agricultural Statistical Service (NASS) of the United States Department of Agriculture (USDA).

2. Review of existing approaches

Under the prediction approach to surveys, an estimator of $\bar{Y}_i$, the mean of the *i*th small area, is given by:

$$\hat{\bar{Y}}_i = f_i \bar{y}_i + (1 - f_i)\hat{\bar{Y}}_{ir}, \tag{1}$$

where $\bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ is the sample mean, index ij denotes observation j from area i, $f_i = n_i / N_i$, N_i and n_i are the number of area i population and sample units, $\hat{\bar{Y}}_{ir}$ is a model-dependent predictor of the mean of the non-sampled part of area i $(i = 1, \cdots, m)$. Let $n = \sum_{i=1}^{m} n_i$ and $N = \sum_{i=1}^{m} N_i$.

For example, the predictor $\hat{\bar{Y}}_{ir}$ can be obtained using a linear mixed model. A comprehensive account on the application of the linear mixed model theory to SAE is given by Rao (2003) and Jiang and Lahiri (2006). To facilitate the subsequent discussion, we refer to the following special case of the linear mixed model, known as the nested-error regression model (Battese et al. 1988):

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \tag{2}$$

$$v_i \overset{iid}{\sim} N(0,\tau^2) \text{ and } \varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2), \quad (3)$$

where $\mathbf{x}_{ij}$ is a vector of known auxiliary variables for an observation ij, $\boldsymbol{\beta}$ is the corresponding vector of parameters; v_i are random effects. The distribution of the random effects describes deviations of the area means from values $\mathbf{x}_{ij}^T\boldsymbol{\beta}$; ε_{ij} are errors in individual observations $(j=1,...,N_i;\ i=1,...,m)$. The random variables v_i and ε_{ij} are assumed to be mutually independent. We assume that sampling is non-informative for the distribution of measurements y_{ij}, given the auxiliary information $\mathbf{x}_{ij}$ $(j=1,...,N_i;\ i=1,...,m)$. The best linear unbiased predictor (BLUP) of $\bar{Y}_{ir}$ has the form $\hat{\bar{Y}}_{ir} = \bar{\mathbf{x}}_{ir}^T\hat{\boldsymbol{\beta}} + \hat{v}_i$, where $\bar{\mathbf{x}}_{ir}^T = (N_i - n_i)^{-1}\sum_{j=n_i+1}^{N_i}\mathbf{x}_{ij}^T$, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, $\hat{v}_i$ is BLUP of v_i and it has the form: $\hat{v}_i = \tau^2(\sigma^2/n_i + \tau^2)^{-1}(\bar{y}_i - \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}})$. We obtain an empirical best linear unbiased predictor (EBLUP) of $\bar{Y}_{ir}$ after plugging in estimates of σ^2 and τ^2.

The linear mixed model assumptions about the distribution of the random terms, v_i and ε_{ij}, may hold for most of the observations. However, there may be areas that do not fit the assumption on the random effects v_i; there may also be individual observations that are not well described by the model assumption on the error terms ε_{ij}. The influence of the outlying areas or individual observations on estimation of the model parameters can be reduced by using bounded influence functions for the corresponding residual terms when fitting the model estimating equations. For the general case of the linear mixed models, this approach was considered by Fellner (1986). Modification of Fellner's approach, also involving the bounded influence functions, was proposed by Sinha and Rao (2008). The predictor for $\bar{Y}_{ir}$ based on such a robustified fitting of the linear mixed model is called the Robust Empirical Best Linear Unbiased Predictor (REBLUP):

$\hat{\bar{Y}}_{ir}^{REBLUP} = \bar{\mathbf{x}}_{ir}^T\hat{\boldsymbol{\beta}}^{REBLUP} + \hat{v}_i^{REBLUP}$. An alternative to the mixed model approach to robust SAE is based on M-quantile regression, which is a generalization of the quantile regression technique. This approach was proposed by Chambers and Tzavidis (2006).

In M-quantile regression, a separate set of linear regression parameters is considered for quantiles q of the conditional distribution of y given x. The M-estimator of the vector $\boldsymbol{\beta}_q$ of the q th quantile regression coefficients is a solution to estimating equations of the form $\sum_{j=1}^{n}\psi_q(r_{jq})\mathbf{x}_j = \mathbf{0}$, where $r_{jq} = y_j - \mathbf{x}_j^T\boldsymbol{\beta}_q$ are residuals, $\psi_q(r_{jq}) = 2\psi(s^{-1}r_{jq})\{qI(r_{jq}>0) + (1-q)I(r_{jq}\le 0)\}$, ψ is a bounded influence function, *s* is a robust estimate of scale. Suppose an observation j falls into quantile q_j. The second step consists of finding the average quantile of the observations in each area *i* as $\bar{q}_i = n_i^{-1}\sum_{j=1}^{n_i}q_{ij}$. Therefore, each area's slope $\boldsymbol{\beta}_{\bar{q}_i}$ is determined by the value of the area's average quantile $\bar{q}_i$. The M-quantile estimator of $\bar{Y}_{ir}$ is given by $\hat{\bar{Y}}_{ir}^{MQ} = \bar{\mathbf{x}}_{ir}^T\hat{\boldsymbol{\beta}}_{\bar{q}_i}^{MQ}$, where $\hat{\boldsymbol{\beta}}_{\bar{q}_i}^{MQ}$ is the estimate of the area's *i* slope.

We next describe the bias correction approach proposed by Chambers *et al.* (2009). The estimation consists of two steps. First, robust estimates are obtained using any outlier robust estimation method, for example, one of the approaches described above. Second, the bias of the initial robust estimate is obtained using an outlier robust approach with different tuning parameters in the corresponding bounded influence functions. The second step tuning parameters should be less restrictive than the ones used at the initial step; that is, there is more reliance on the data rather than on the model assumptions so that the purpose of the second step is to "undo" the effect of a possible model misspecification imposed at the first step. The final estimate is the sum of the robust estimate computed at the first step and the bias correction term computed at the second step.

Let $\phi(.)$ be some bounded function that is not as restrictive as $\psi(.)$.

The bias-corrected version of REBLUP (either Fellner's or Sinha and Rao's approach) is

$\hat{\bar{Y}}_{ir}^{REBLUP+BC} = \hat{\bar{Y}}_{ir}^{REBLUP} + n_i^{-1}\sum_{j=1}^{ni} s_i^{REBLUP}\phi\left((y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{REBLUP} - \hat{v}_i^{REBLUP})/s_i^{REBLUP}\right)$. The bias-corrected version of Chambers and Tzavidis' approach is

$\hat{\bar{Y}}_{ir}^{MQ+BC} = \hat{\bar{Y}}_{ir}^{MQ} + n_i^{-1}\sum_{j=1}^{n_i} s_i^{MQ}\phi\left((y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\bar{q}_i}^{MQ})/s_i^{MQ}\right)$, where s_i^{REBLUP} and s_i^{MQ} are some robust estimates of scale for the respective sets of residuals in area *i*.

3. Proposed approach

The proposed approach uses the same general form (1). The predictor for the sample-complement part is derived from a model (denoted N2) that is based on mixture of two normal distributions with common mean and different variances. The model is given by (4)-(6):

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \hat{v}_i + \varepsilon_{ij}, \tag{4}$$

$$v_i \overset{iid}{\sim} N(0,\tau^2) \text{ and } \varepsilon_{ij} \mid z_{ij} \overset{iid}{\sim} (1-z_{ij})N(0,\sigma_1^2) + z_{ij}N(0,\sigma_2^2), \tag{5}$$

and the mixture part indicator is a random binomial variable

$$z_{ij} \mid \pi \overset{iid}{\sim} Bin(1;\pi), \tag{6}$$

where π is the probability of belonging to mixture part 2 ($j = 1,...,N_i; i = 1,...,m$). Note that, conditional on the value of the mixture part indicator z_{ij}, the model is the usual linear mixed effects model as given by (2) and (3).

The predictor is given by $\hat{\bar{Y}}_{ir}^{N2} = \bar{\mathbf{x}}_{ir}^T\hat{\boldsymbol{\beta}}^{N2} + \hat{v}_i^{N2}$. Let $\phi = (\sigma_1, \sigma_2, \tau, \pi, \boldsymbol{\beta})$ denote the set of model parameters. We used the EM algorithm for estimation of the model parameters.

Each observation has its own conditional probability $P\{z_{ij} = 1 \mid y_{ij}, \mathbf{x}_{ij}, \phi\} = E\left[z_{ij} \mid y_{ij}, \mathbf{x}_{ij}, \phi\right]$ of belonging to part 2 of the mixture, so that the observations in the sample can be ranked according to these probabilities. The estimate of $\boldsymbol{\beta}$ (thus, the synthetic part of the estimator) is outlier robust because the outlying observations would be classified with a higher probability to the higher variance part of the mixture; hence, they would be "down-weighted" according to the formula

$\hat{\boldsymbol{\beta}}^{N2} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij} (y_{ij} - \hat{v}_i)$, where the weights are given by $w_{ij} = \hat{\sigma}_1^{-2}(1-\hat{z}_{ij}) + \hat{\sigma}_2^{-2}\hat{z}_{ij}$ with $\hat{z}_{ij} = E\left[z_{ij} \mid y_{ij}, \mathbf{x}_{ij}, \hat{\varphi} \right]$. The predictor for the random effect $\hat{v}_i^{N2}$ has the form

$$\hat{v}_i^{N2} = \frac{\tau^2}{D_i^{N2} + \tau^2} (\hat{\bar{y}}_i^{N2} - \hat{\bar{\mathbf{x}}}_i^{N2} \hat{\boldsymbol{\beta}}^{N2}), \tag{7}$$

where $D_i^{N2} = \left(\sum_{j=1}^{n_i} w_{ij} \right)^{-1}$, $\hat{\bar{y}}_i^{N2} = \left(\sum_{j=1}^{n_i} w_{ij} \right)^{-1} \sum_{j=1}^{n_i} w_{ij} y_{ij}$, and $\hat{\bar{\mathbf{x}}}_i^{N2} = \left(\sum_{j=1}^{n_i} w_{ij} \right)^{-1} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij}^T$.

Note that the "direct" estimate $\hat{\bar{y}}_i^{N2}$ accounts for outliers. In fact, this estimate is not exactly "direct" because it depends on units from other areas through the estimates of variances and the probabilities of belonging to part 2 of the mixture.

Let $e_{ij}^{N2} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^{N2} - \hat{v}_i^{N2}$, then the overall bias corrected EBP is given by $\hat{\bar{Y}}_{ir}^{N2+OBC} = \hat{\bar{Y}}_{ir}^{N2} + n^{-1} s^R \sum_{i=1}^{m} \sum_{j=1}^{n_i} \phi_b \left(e_{ij}^{N2} / s^R \right)$, where s^R is a robust measure of scale for the set of residuals $\left\{ e_{ij}^{N2}; j = 1,...,n_i, i = 1,...,m \right\}$, e.g., $s^R = med \left| e_{ij}^{N2} - med(e_{ij}^{N2}) \right| / 0.6745$ and ϕ_b is a bounded Huber's function with the tuning parameter $b = 5$.

4. Simulation Study

The purpose of the simulation study is to compare the performances of different methods under different scenarios. We use the same setup as in Chambers *et al.* (2009) and briefly describe it here. From each area, a sample is selected using simple random sampling without replacement. Each area has 100 population units and 5 sampled units. The auxiliary variable x_{ij} is generated from the lognormal distribution with mean 1.004077 and standard deviation of 0.5 and the population values y_{ij} are generated as $y_{ij} = 100 + 5x_{ij} + v_i + \varepsilon_{ij}$. There are several scenarios for distribution of v_i and ε_{ij}, as described below.

1. No contamination scenario, [0,0]: $v_i \sim N(0,3)$, $\varepsilon_{ij} \sim N(0,6)$;
2. Outlying areas, [0,v]: for the first 36 areas, $v_i \sim N(0,3)$; for the last four areas, $v_i \sim N(9,20)$; $\varepsilon_{ij} \sim N(0,6)$ for all observations;
3. Individual outliers, [e,0]: $v_i \sim N(0,3)$ for all areas; $\varepsilon_{ij} \sim N(0,6)$ with probability 0.97 and $\varepsilon_{ij} \sim N(20,150)$ with probability 0.03;
4. Individual outliers and outlying areas, [e,v]: for the first 36 areas, $v_i \sim N(0,3)$; for the last four areas, $v_i \sim N(9,20)$; $\varepsilon_{ij} \sim N(0,6)$ with probability 0.97 and $\varepsilon_{ij} \sim N(20,150)$ with probability 0.03;

5. Individual outliers only, $\varepsilon_{ij} \sim N(0,6)$ with probability 0.75 and $\varepsilon_{ij} \sim N(20,3000)$ with probability 0.25; random effects are $v_i \sim N(0,3)$. (This version was considered in Gershunskaya 2010)

The tuning parameters in the bounded Huber's function for REBLUP are set to *b=1.345*; for the bias-correction of REBLUP (Fellner and SR) and MQ, the tuning parameters are set to *b=3*. The tuning parameter for the overall bias correction is *b=5*. We used 250 simulation runs for each of the above scenarios and compared the estimates with the corresponding population area means.

To assess the quality of the estimators, we used the median value of the relative bias, $RB = 100 \cdot med_i \{250^{-1} \sum_{s=1}^{250} (\hat{\bar{Y}}_{is} - \bar{Y}_{is}) \Big/ 250^{-1} \sum_{s=1}^{250} \bar{Y}_{is}\}$, and the median of the relative root mean squared error, $RRMSE = 100 \cdot med_i \left\{ \sqrt{250^{-1} \sum_{s=1}^{250} (\hat{\bar{Y}}_{is} - \bar{Y}_{is})^2} \Big/ 250^{-1} \sum_{s=1}^{250} \bar{Y}_{is} \right\}$, index $s = 1,...,250$ denotes the simulation run.

Consider scenarios 1-4 (see Tables 4.1). In the no-outliers situations, the estimator N2 works similar to the regular EBLUP. If there are only individual outliers or only area level outliers, REBLUP and N2 (not the bias-corrected versions) have similar RRMSE's. Both the original and the bias-corrected versions of MQ are less efficient than REBLUP for the four outlying areas. (Some discrepancy between our results for MQ and the ones reported in Chambers *et al.* (2009) could be due to the sensitivity of MQ to the choice of the number of quantiles.) N2 estimator has a large bias when both the individual and area outliers are present. This bias is corrected in the N2+BC versions, so that the RRMSE's of the N2+BC versions in the four outlying areas is comparable to the other estimators (see Gershunskaya 2010). In the OBC* version of N2, we only correct the overall bias. The OBC version for N2 corrects both area and overall bias (see Gershunskaya 2010). It appears that N2+OBC works uniformly well for all considered scenarios. Gershunskaya (2010) found that for scenario 5, N2+OBC version is better than the other estimators. If a similar situation happens in the CES data, then this version of N2 estimators may be preferred.

Table 4.1: Simulation Results Scenarios 1-4 (250 runs), $N_i = 100$, $n_i = 5$

	No outliers		Individual outliers only		Area outliers	Individual and area outliers
Estimator / Scenario	[0,0]	[0,u]/1-36	[e,0]	[e,u]/1-36	[0,u]/37-40	[e,u]/37-40
Median values of Relative Bias (expressed as a percentage)						
EBLUP	-0.001	0.067	-0.004	0.191	-0.579	-1.546
REBLUP (F)	0.003	0.075	-0.374	-0.298	-0.625	-0.977
REBLUP (SR)	0.005	0.090	-0.370	-0.275	-0.538	-0.902
MQ	0.020	0.097	-0.374	-0.286	-1.003	-0.468
F+BC	-0.007	-0.003	-0.265	-0.258	-0.043	-0.233
SR+BC	-0.009	-0.001	-0.266	-0.255	-0.034	-0.225
MQ+BC	-0.006	0.001	-0.262	-0.258	-0.243	-0.156
N2	-0.001	0.068	-0.448	-0.321	-0.594	-3.250
N2+OBC*	-0.001	0.068	-0.235	0.071	-0.594	-2.885
N2+OBC	-0.005	0.003	0.002	-0.153	-0.073	-0.842

Median values of Relative Root MSE (expressed as a percentage)						
EBLUP	0.809	0.859	1.207	1.354	1.041	2.289
REBLUP (F)	0.821	0.823	0.989	0.972	1.076	1.396
REBLUP (SR)	0.825	0.827	0.991	0.966	1.035	1.342
MQ	0.844	0.846	0.996	0.975	1.650	1.468
F+BC	0.913	0.917	1.221	1.224	0.861	1.189
SR+BC	0.910	0.916	1.219	1.225	0.866	1.179
MQ+BC	0.914	0.920	1.223	1.226	0.994	1.421
N2	0.808	0.858	1.007	0.978	1.049	4.559
N2+OBC*	0.808	0.858	0.937	0.953	1.049	4.221
N2+OBC	0.859	0.878	0.921	0.944	0.879	1.308

5. Applications

5.1. U.S. Bureau of Labor Statistics Application

The purpose of this study is to provide a first glimpse into the prospect of using the alternative models for SAE in CES. In this simulation, historical administrative data from the Quarterly Census of Employment and Wages (QCEW) program of the U.S. Bureau of Labor Statistics played the role of "real" data. (In real time production, the estimates are based on the data collected by CES, which is somewhat different from the QCEW data; nevertheless, the use of the QCEW data is appropriate for preliminary research.)

In CES, the goal is to estimate the relative over-the-month change in employment at a given month t in areas $i=1,\ldots,m$, where the areas are formed by cross-classifying industries and metropolitan statistical areas (MSA). For area m, the target finite population quantity at month t is $R_{i,t} = \sum_{j\in P_{i,t}} y_{ij,t} \Big/ \sum_{j\in P_{i,t}} y_{ij,t-1}$, where $P_{i,t}$ is a set of the area m population establishments having non-zero employment in both previous and current months, i.e., $y_{ij,t-1} > 0$ and $y_{ij,t} > 0$. The direct sample estimate is $\hat{R}_{i,t} = \sum_{j\in S_{i,t}} w_{ij} y_{ij,t} \Big/ \sum_{j\in S_{i,t}} w_{ij} y_{ij,t-1}$, where $S_{i,t}$ is a set of the area m sample establishments having $y_{ij,t-1} > 0$ and $y_{ij,t} > 0$; w_{ij} is the sample weight for unit ij **.**

In order to work at a unit level, we expand $R_{m,t}$ around a hypothetical true superpopulation parameter (as in Gershunskaya and Lahiri 2008). Define the following variable: $y^*_{ij,t} = (1-\hat{f}_i)(\bar{\hat{w}}_i - 1)^{-1}(w_{ij}-1)\hat{v}_{ij,t} + \hat{R}_t + \hat{f}_i \bar{\hat{v}}_{i,t}$, where $\hat{R}_t$ is the estimated ratio of employment at a statewide level; $\hat{v}_{ij,t} = \hat{\bar{Y}}^{-1}_{t-1}(y_{ij,t} - \hat{R}_t y_{ij,t-1})$ is the estimated influence function for the ratio; $\hat{\bar{Y}}_{t-1}$ is an estimate of the previous month mean statewide employment; $\bar{w}_i = n_i^{-1}\sum_{j\in S_{i,t}} w_{ij}$ is area m average weight; $\bar{\hat{v}}_{i,t} = n_i^{-1}\sum_{j\in S_{i,t}} \hat{v}_{ij,t}$; $\hat{f}_i = \hat{N}_i^{-1} n_i$ is the estimated area sample fraction and $\hat{N}_i = \sum_{j\in S_{i,t}} w_{ij}$ is the estimated number of population units.

We compared performances of several estimators: one estimator is based on the area-level Fay-Herriot model and the other estimators are based on different unit-level models. We used single

slope, without intercept linear models, with the past year's population trend $R_{i,t-12}$ playing the role of an auxiliary variable (i.e., area-level auxiliary information for all observations in area *i*).

We considered four States (Alabama, California, Florida, and Pennsylvania) and obtained estimates for September 2006 using the sample drawn from the 2005 sampling frame, mimicking the production timeline. We fit the models separately for each State's industrial supersector: a set of MSAs within States' industrial supersectors defined the set of small areas. The resulting estimates were compared to the corresponding true population values $R_{m,t}$ available from QCEW.

Performance of each estimator is measured using the 75th percentile of the absolute error $E_{i,t} = 100\left|\hat{R}_{i,t} - R_{i,t}\right|$ and the empirical root mean squared error $ERMSE_t = \left[m^{-1} \sum_{i=1}^{m} E_{i,t}^2 \right]^{\frac{1}{2}}$.

Summaries of results for each State are reported in Tables 5.1. Overall, the performance of N2 (and its bias-corrected versions) is quite satisfactory. In Alabama, the N2 estimator is slightly more efficient than REBLUP and better than the other estimators. In California, ERMSEs of REBLUP and MQ are smaller than of N2 but, in terms of the 75th percentile (reported in Gershunskaya 2010), these estimators are very close. In Florida, N2 is only slightly better than REBLUP for 75 percent of the areas but is much better in terms of the ERMSE, due to a significantly better performance in a few areas. In Pennsylvania, in several industries, N2 estimator had a large error due to the overall bias. The OBC version of N2 reduced the bias and made a good estimator.

Table 5.1: Empirical Root Mean Squared Error, %

State	FH	EBLUP	REBLUP (F)	MQ	F+BC	MQ+BC	N2	N2+OBC*	N2+OBC
AL	1.868	2.257	1.899	2.023	2.027	2.133	1.767	1.743	1.873
CA	2.502	2.339	2.099	2.040	2.388	2.378	2.165	2.158	2.307
FL	3.425	2.707	2.771	3.766	2.887	3.847	1.184	1.197	1.145
PA	1.418	1.318	1.754	1.664	2.092	2.129	1.627	1.547	1.264

5.2 NASS Application

The United States Agricultural Statistical Service (NASS) has been publishing county level crop and livestock estimates since 1917 (see Iwig 1993). County indications of crops such as harvested yield (i.e. production per unit harvested acreage) are needed to assist farmers, agribusinesses and government agencies in local agricultural decision making. Most NASS Field Offices conduct a separate County Estimates Survey every year. Data from multiple sample surveys (such as the County Acreage and Production Survey (CAPS) and Quarterly Agricultural Survey (QAS)) are used to estimate harvested yield for various crops (such as soybeans) at the county level.

We evaluated performances of EBLUP under normal model (2)-(3) and the EBP for mixture model (4)-(5) [N2] with direct (KB) and NASS official estimates for seven mid-western states in the year 2007, treating the census as gold standard. To save space, we report results for three states - Illinois, Iowa, and Minnesota. Following Bellow and Lahiri (2010), we report average absolute deviation (AAD), average squared deviation (ASD), average absolute relative deviation

(AARD), average squared relative deviation (ASRD) and percentage below census (PBPC) in Table 5.2. We use the following definition: (i) AAD: the mean of absolute deviations between county estimates and corresponding 2007 census (PC) values; (ii) ASD: the mean of squared deviations between estimates and PC values; (iii) AARD: the mean of ratios between absolute deviations and PC values; (iv) ASRD: the mean of squared ratios between absolute deviations and PC values; (v) PBC: the proportion of counties with estimate less than the corresponding PC value. Values of PBPC below (above) 0.5 indicate possible overestimation (underestimation) tendencies for an estimator. The EBLUP under normal model (2)-(3) and the EBP for mixture model (4)-(5) [N2] estimates are clearly superior to the Kott-Busselberg (KB) direct estimates for all the states considered. EBPs are also better than the official estimates in all but one state (Minnesota), where the official estimates have slight edge over EBPs. The OBC* correction to N2 provides similar results for most of the seven states. However, it provides slightly better results for Iowa, but slightly worse results for Minnesota. In the future, we plan to evaluate the other estimators considered in the paper.

Table 5.2: Estimation Accuracy Measures for Harvested Yield*

State	Estimator	Metric				
		AAD	ASD	AARD	ASRD	PBC
Illinois	EBLUP	1.34	2.85	0.036	0.002	0.32
	KB	2.7	12.6	0.07	0.009	0.85
	N2	1.33	2.8	0.036	0.002	0.33
	N2+OBC*	1.33	2.8	0.036	0.002	0.32
	Official	1.82	5.18	0.048	0.004	0.42
Iowa	EBLUP	1.10	1.81	0.022	0.001	0.69
	KB	2.7	13.5	0.055	0.006	0.82
	N2	1.24	2.15	0.025	0.001	0.83
	N2+OBC*	0.95	1.48	0.019	0.001	0.72
	Official	2.12	5.94	0.043	0.002	0.08
Minnesota	EBLUP	1.32	3.92	0.037	0.004	0.31
	KB	3.46	26.0	0.095	0.022	0.85
	N2	1.23	4.04	0.036	0.004	0.36
	N2+OBC*	1.38	4.58	0.040	0.005	0.28
	Official	1.32	2.67	0.034	0.002	0.19

*The evaluation metric for KB and Official estimators are obtained from Bellow and Lahiri (2010).

6. Concluding Remarks

The EBP under mixture model with appropriate area level and overall bias correction appears to perform better than the rival estimators in most situations. We are currently investigating the possibility of resampling methods in estimating the mean squared error of the proposed estimators and the associated confidence interval problem. We shall report the findings in a separate paper.

Acknowledgements

The second author would like to thank his NASS colleagues for some helpful discussion.
The second author's research was supported in part by National Science Foundation grant SES-0851001 and a ASA/USDA-NASS Fellowship. Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics or NASS-USDA.

References

1. Battese, G. E., Harter, R. M. and Fuller, W. A. (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.
2. Bellow, M. and Lahiri, P. (2010), Empirical Bayes Methodology for the NASS County Estimation Program, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
3. Chambers, R. L. (2005) *What If... ? Robust Prediction Intervals for Unbalanced Samples.* Southampton, UK, Southampton Statistical Sciences Research Institute, 21pp. (S3RI Methodology Working Papers, M05/05) http://eprints.soton.ac.uk/14075/
4. Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93,255-268.
5. Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2009). Outlier Robust Small Area Estimation. Invited Presentation, ISI 2009, South Africa.
6. Datta, G.S. and Lahiri, P. (1995) Robust hierarchical Bayes estimation of small area characteristics in presence of covariates and outliers, *Journal of Multivariate Analysis*, Vol. 54, No. 2, 310-328.
7. Fay, R.E. and Herriot, R.A.(1979). Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data, *Journal of American Statistical Association*, 74, 269-277.
8. Fellner, W. H. (1986), Robust Estimation of Variance Components," *Technometrics*, 28, 51-60.
9. Gershunskaya, J. and Lahiri, P., (2008). Robust Estimation of Monthly Employment Growth Rates for Small Areas in the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
10. Gershunskaya, J. (2010), Robust Small Area Estimation Using a Mixture Model, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
11. Hawala, S. and Lahiri, P. (2010) Variance Modeling in the U.S. Small Area Income and Poverty Estimates Program for the American Community Survey, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
12. Iwig, W.C. (1993), The National Agricultural Statistics Service County Estimates Program, National Agricultural Statistics Service.
13. Rao, J.N.K. (2003). *Small Area Estimation*, New-York, John Wiley & Sons, Inc.
14. Sinha, S.K. and Rao, J.N.K. (2008). Robust methods for small area estimation. *Proceedings of the American Statistical Association*, Survey Research Methods Section, Alexandria, VA: American Statistical Association, 27-38.