Robust Small Area Estimation Using a Mixture Model[1] October 2010

Julie Gershunskaya
U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Suite 4985, Washington, DC, 20212

Abstract:

Different methods have been proposed in the small area estimation literature to deal with outliers in individual observations and in the area-level random effects. In this paper, we propose a new method based on a scale mixture of two normal distributions. Using a simulation study, we compare the performance of a few recently proposed robust small area estimators and our proposed estimator based on a mixture distribution. We then compare the proposed method with the existing methods to estimate monthly employment changes in the metropolitan statistical areas using data from the Current Employment Statistics Survey conducted by the U.S. Bureau of Labor Statistics (BLS).

Key Words: robust estimation, small area, mixture model

1. Introduction

Small area estimation (SAE) generally relies on some, implicit or explicit, modeling assumptions. It may happen that a relatively few observations do not fit into the model that explains well bulk of the data. Such observations may adversely affect estimation of the model parameters. This calls for development of methods of estimation that are robust to the appearance of outliers, and several outlier resistant methods have been proposed in the SAE literature in recent years (Fellner 1986, Chambers and Tzavidis 2006, Sinha and Rao 2008).

On the other hand, outliers may suggest a real finite population structure that is not described by the assumed base model. Such 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 the non-SAE settings, Chambers (1986) proposed to apply 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. The bias correction idea in application to SAE is to add separate bias correction terms to the initial predictors for each area, 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, where outliers are assumed to have a larger variance than the "regular" observations. This model explicitly describes the behaviour of the outlying observations relative to the other units; thus, it automatically produces estimates (e.g., using MLE) that account for outliers.

A simple formulation of the mixture model used in this paper may still be too strong in certain assumptions about the distribution of outliers. First, the outliers are assumed to appear randomly across areas. In fact, however, the outliers may be clustered in certain areas. This may lead to bias in the prediction of the area-level random effects. We propose an area-level bias correction

[1] Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

method that is different from the one of Chambers *et al.* (2009): the proposed method attempts 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. Another potentially incorrect assumption is that the outliers 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. Application using administrative Quarterly Census of Employment and Wages (QCEW) data for estimation of monthly employment changes in the metropolitan statistical areas (MSA) using sample from the Current Employment Statistics (CES) Survey conducted by the U.S. Bureau of Labor Statistics (BLS) is described in Section 5.

2. Review of existing approaches

Under the prediction approach to surveys, an estimator of $\bar{Y}_m$, the small area m mean, is given by:

$$\hat{\bar{Y}}_m = f_m \bar{y}_m + (1 - f_m)\hat{\bar{Y}}_{mr}, \tag{1}$$

where $m = 1,...,M$; $\bar{y}_m = n_m^{-1}\sum_{j=1}^{n_m} y_{mj}$ is the sample mean, index mj denotes observation j from area m, $f_m = N_m^{-1} n_m$, N_m and n_m are the number of area m population and sample units, $\sum_{m=1}^{M} N_m = N$; $\sum_{m=1}^{M} n_m = n$; $\hat{\bar{Y}}_{mr}$ is a model-dependent predictor of the mean of the non-sampled part of area m.

In particular, the predictor $\hat{\bar{Y}}_{mr}$ can be obtained using a linear mixed model. A comprehensive account about application of the linear mixed model theory to SAE is given by Rao (2003). 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, Harter, Fuller 1998):

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

$$u_m \overset{iid}{\sim} N(0, \tau^2) \text{ and } \varepsilon_{mj} \overset{iid}{\sim} N(0, \sigma^2) \tag{3}$$

$$j = 1,...,n_m, m = 1,...,M,$$

where $\mathbf{x}_{mj}$ is a vector of auxiliary variables for an observation mj, $\boldsymbol{\beta}$ is the corresponding vector of parameters; u_m are random effects. The distribution of the random effects describes deviations of the area means from values $\mathbf{x}_{mj}^T \boldsymbol{\beta}$; ε_{mj} are errors in individual observations. The random variables u_m and ε_{mj} are assumed to be mutually independent. (We assume that sampling is non-informative for the distribution of measurements y_{mj}, given the auxiliary information $\mathbf{x}_{mj}$.)

The best linear unbiased predictor (BLUP) of $\bar{Y}_{mr}$ has the form

$$\hat{\bar{Y}}_{mr} = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}} + \hat{u}_m, \tag{4}$$

where $\bar{\mathbf{x}}_{mr}^T = (N_m - n_m)^{-1} \sum_{j=n_m+1}^{N_m} \mathbf{x}_{mj}^T$, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, $\hat{u}_m$ is BLUP of u_m and it has the form

$$\hat{u}_m = \frac{\tau^2}{\sigma^2 / n_m + \tau^2} (\bar{y}_m - \bar{\mathbf{x}}_m^T \hat{\boldsymbol{\beta}}). \tag{5}$$

When the variances σ^2 and τ^2 are not known, they are estimated from the data. This will yield the empirical best linear unbiased predictor (EBLUP) for $\bar{Y}_{mr}$.

The linear mixed model assumptions about the distribution of the random terms, u_m and ε_{mj}, may hold for most of the observations; however, there may be areas that do not fit the assumption on the random effects u_m; there may also be individual observations that are not well described by the model assumption on the error terms ε_{mj}. 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 taken by Fellner (1986). Modification of Fellner's approach, also involving the bounded influence functions, was proposed by Sinha and Rao (2008). The predictor for $\bar{Y}_{mr}$ based on such a robustified fitting of the linear mixed model is called the Robust Empirical Best Linear Unbiased Predictor (REBLUP):

$$\hat{\bar{Y}}_{mr}^{REBLUP} = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}}^{REBLUP} + \hat{u}_m^{REBLUP}. \tag{6}$$

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

$$\sum_{j=1}^{n} \psi_q(r_{jq}) \mathbf{x}_j = \mathbf{0}, \tag{7}$$

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 m as $\bar{q}_m = n_m^{-1} \sum_{j=1}^{n_m} q_{mj}$. Therefore, each area's slope $\boldsymbol{\beta}_{\bar{q}_m}$ is determined by the value of the area's average quantile $\bar{q}_m$. The M-quantile estimator of $\bar{Y}_{mr}$ is given by

$$\hat{\bar{Y}}_{mr}^{MQ} = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}}_{\bar{q}_m}^{MQ}, \tag{8}$$

where $\hat{\boldsymbol{\beta}}_{\bar{q}_m}^{MQ}$ is the estimate of the area's m slope.

We next describe the bias correction approach proposed by Chambers *et al.* (2009). The estimation consists of two steps. First, find robust estimates using any outlier robust estimation method, for example, one of the approaches described above. Second, estimate the bias of the initial robust estimate using, again, an outlier robust approach but 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 step one. 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 both $\phi(\cdot)$ and $\psi(\cdot)$ be some bounded functions, where $\phi(\cdot)$ is not as restrictive as $\psi(\cdot)$

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

$$\hat{\bar{Y}}_{mr}^{REBLUP+BC} = \hat{\bar{Y}}_{mr}^{REBLUP} + n_m^{-1}\sum_{j=1}^{n_m} s_m^{REBLUP}\phi\left(\frac{y_{mj} - \mathbf{x}_{mj}^T\hat{\boldsymbol{\beta}}^{REBLUP} - \hat{u}_m^{REBLUP}}{s_m^{REBLUP}}\right). \tag{9}$$

The bias-corrected version of Chambers and Tzavidis' approach is

$$\hat{\bar{Y}}_{mr}^{MQ+BC} = \hat{\bar{Y}}_{mr}^{MQ} + n_m^{-1}\sum_{j=1}^{n_m} s_m^{MQ}\phi\left(\frac{y_{mj} - \mathbf{x}_{mj}^T\hat{\boldsymbol{\beta}}_{\bar{q}_m}^{MQ}}{s_m^{MQ}}\right). \tag{10}$$

Here s_m^{REBLUP} and s_m^{MQ} are some robust estimates of scale for the respective sets of residuals in area m.

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 (11)-(13):

$$y_{mj} = \mathbf{x}_{mj}^T\boldsymbol{\beta} + u_m + \varepsilon_{mj}, \tag{11}$$

$$u_m \overset{iid}{\sim} N(0,\tau^2) \text{ and } \varepsilon_{mj} \mid z \overset{iid}{\sim} (1-z)N(0,\sigma_1^2) + zN(0,\sigma_2^2), \tag{12}$$

$$j = 1,...,n_m, m = 1,...,M,$$

and the mixture part indicator is a random binomial variable

$$z \mid \pi \sim Bin(1;\pi), \tag{13}$$

where

π is the probability of belonging to mixture part 2 (the outlier part, $\sigma_2 \geq \sigma_1$).

Note that, conditional on the value of the mixture part indicator z, the model is the usual mixed effects model as given by (2) and (3).

We used the EM algorithm for estimation of the model parameters (see Appendix). The predictor is given by

$$\hat{\bar{Y}}_{mr}^{N2} = \bar{\mathbf{x}}_{mr}^T\hat{\boldsymbol{\beta}}^{N2} + \hat{u}_m^{N2}. \tag{14}$$

Let $\boldsymbol{\theta} = (\sigma_1, \sigma_2, \tau, \pi, \boldsymbol{\beta})$ denote the set of model parameters.

Each observation has its own conditional probability $P\{z = 1 \mid y_{mj}, \mathbf{x}_{mj}, \boldsymbol{\theta}\} = E\left[z \mid y_{mj}, \mathbf{x}_{mj}, \boldsymbol{\theta} \right]$ 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 "downweighted" according to the formula $\hat{\boldsymbol{\beta}}^{N2} = \sum_{m=1}^{M} \sum_{j=1}^{n_m} w_{mj} \mathbf{x}_{mj}^T (y_{mj} - \hat{u}_m) \Big/ \sum_{m=1}^{M} \sum_{j=1}^{n_m} w_{mj} \mathbf{x}_{mj}^T \mathbf{x}_{mj}$,

where the weights are given by

$$w_{mj} = \hat{\sigma}_1^{-2} (1 - \hat{z}_{mj}) + \hat{\sigma}_2^{-2} \hat{z}_{mj} \tag{15}$$

with $\hat{z}_{mj} = E\left[z \mid y_{mj}, \mathbf{x}_{mj}, \hat{\boldsymbol{\theta}} \right]$.

The predictor for the random effect $\hat{u}_m^{N2}$ has the form

$$\hat{u}_m^{N2} = \frac{\tau^2}{D_m^{N2} + \tau^2} (\hat{\bar{y}}_m^{N2} - \hat{\bar{\mathbf{x}}}_m^{N2} \hat{\boldsymbol{\beta}}^{N2}), \tag{16}$$

where $D_m^{N2} = \left(\sum_{j=1}^{n_m} w_{mj} \right)^{-1}$, $\hat{\bar{y}}_m^{N2} = \left(\sum_{j=1}^{n_m} w_{mj} \right)^{-1} \sum_{j=1}^{n_m} w_{mj} y_{mj}$, and $\hat{\bar{\mathbf{x}}}_m^{N2} = \left(\sum_{j=1}^{n_m} w_{mj} \right)^{-1} \sum_{j=1}^{n_m} w_{mj} \mathbf{x}_{mj}^T$.

Note that the "direct" estimate $\hat{\bar{y}}_m^{N2}$ in (16) 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. The sample average $\bar{y}_m$ may or may not be affected by outliers. It depends on the contents of the area: for example, if an area contains several units that have a high probability of belonging to the "outlier" part of the mixture, it is possible that the whole area would tend to be an outlier. Note that if outliers tend to be clustered in some areas, this would mean that the distribution of the mixture indicators depends on the area label, which would contradict the model assumption (13). The failure of the *independence assumption* may lead to significant biases in the areas with a larger portion of the outlying observations. We propose a test to determine that an area is not an outlying area and a simple method for the area-level bias correction in areas where the test fails, as described below.

First, consider the following "bias corrected" variation of $\hat{\bar{Y}}_{mr}^{N2}$:

Bias Correction 1 (BC1). Denote residuals $e_{mj}^{N2} = y_{mj} - \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}}^{N2} - \hat{u}_m^{N2}$.

For each area, find the estimate of the mean residual using a mixture of two normal distributions model and by treating areas *as fixed effects*:

$$e_{mj}^{N2} = \mu_m + \varepsilon_{mj}, \tag{17}$$

$$\varepsilon_{mj} \mid z \overset{iid}{\sim} (1 - z) N(0, \sigma_1^2) + z N(0, \sigma_2^2), \tag{18}$$

$j = 1, ..., n_m, m = 1, ..., M,$ and

$$z \mid \pi \sim Bin(1; \pi). \tag{19}$$

The BC1 estimator is

$$\hat{\bar{Y}}_{mr}^{N2+BC1} = \bar{\mathbf{x}}_{mr}^{T}\hat{\boldsymbol{\beta}}^{N2} + \hat{u}_{m}^{N2} + \hat{\mu}_{m}^{BC1}, \tag{20}$$

where $\hat{\mu}_{m}^{BC1}$ is the estimate of μ_m from the above model.

Bias Correction 2 (BC2). As a general rule, BC1 may be inefficient in areas where the estimates of μ_m are based on a small number of observations. Therefore, we propose to use $\hat{\bar{Y}}_{mr}^{N2+BC1}$ only when we can demonstrate that an area *m* is an outlying area. Consider the following statistic:

$$\hat{\pi}_m = n_m^{-1}\sum_{j=1}^{n_m}\hat{z}_{mj}. \tag{21}$$

The distribution of the statistic $\hat{\pi}_m$ under the independence assumption can be simulated using the estimated model parameters. These simulated values can be used to obtain a threshold. If the actual estimated $\hat{\pi}_m$ is greater than the threshold, the whole area is considered an outlier. The detailed procedure for an area *m* can be described by the following steps:

1. Generate $\gamma \sim Bin(1;\hat{\pi})$ and $\eta \sim \begin{cases} N(0,\hat{\sigma}_1^2+\hat{\tau}^2) & \text{if } \gamma=0 \\ N(0,\hat{\sigma}_2^2+\hat{\tau}^2) & \text{if } \gamma=1 \end{cases}$.

2. Using the Bayes formula, find the probability of belonging to part 2 of the mixture, given the value of η:

$$z^{(a)} = P\{z=1 \mid \eta;\hat{\boldsymbol{\theta}}\}$$

$$= \frac{\hat{\pi}P\{\eta \mid z=1;\hat{\boldsymbol{\theta}}\}}{(1-\hat{\pi})P\{\eta \mid z=0;\hat{\boldsymbol{\theta}}\} + \hat{\pi}P\{\eta \mid z=1;\hat{\boldsymbol{\theta}}\}}$$

$$= \frac{\dfrac{\hat{\pi}}{\sqrt{\hat{\sigma}_2^2+\hat{\tau}^2}}\exp\left(-\dfrac{1}{2}\dfrac{\eta^2}{\hat{\sigma}_2^2+\hat{\tau}^2}\right)}{\dfrac{1-\hat{\pi}}{\sqrt{\hat{\sigma}_1^2+\hat{\tau}^2}}\exp\left(-\dfrac{1}{2}\dfrac{\eta^2}{\hat{\sigma}_1^2+\hat{\tau}^2}\right)+\dfrac{\hat{\pi}}{\sqrt{\hat{\sigma}_2^2+\hat{\tau}^2}}\exp\left(-\dfrac{1}{2}\dfrac{\eta^2}{\hat{\sigma}_2^2+\hat{\tau}^2}\right)}.$$

3. Repeat steps 1 and 2 n_m times: $a=1,...,n_m$.

4. Let $\pi_m^{(b)} = n_m^{-1}\sum_{a=1}^{n_m} z^{(a)}$ be the average of n_m simulated values of z.

5. Repeat steps 1-4 a large number of times: $b=1,...,B$ (say, $B=500$).

6. Using the simulated values $\pi_m^{(b)}$, $b=1,...,B$, estimate a "theoretical value" p_m^{α} such that $P\{\pi_m > p_m^{\alpha}\}$ is smaller than some predetermined level α. This value depends on the number of units in area *m*.

7. If the actual value, obtained as (21), is higher than p_m^{α}, then the area *m* has more outliers than would be in a "regular" area under the independence assumption; thus, it can be regarded as an outlying area, and the bias correction $\hat{\mu}_m^{BC1}$ from (20) is applied; otherwise, the bias correction is not applied. In our simulations, for application of the bias

adjustment, we required that an area had at least four sample units and $\hat{\pi}_m > p_m^{\alpha}$, where $\alpha = 0.05$:

$$\hat{\mu}_m^{BC2} = \begin{cases} \hat{\mu}_m^{BC1}, & \text{if } \hat{\pi}_m > p_m^{\alpha} \text{ and } n_m \geq 4 \\ 0, & otherwise \end{cases} \tag{22}$$

The BC2 estimator is

$$\hat{\bar{Y}}_{mr}^{N2+BC2} = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}}^{N2} + \hat{u}_m^{N2} + \hat{\mu}_m^{BC2}, \tag{23}$$

Remark 1. The data consists of the individual measurements y_{mj}, along with the corresponding area labels, while the area-level effects u_m are not observable. It is not obvious what is meant by "outlyingness" of an unobserved quantity in the REBLUP approaches. The mixture model formulation, on the other hand, allows the description of the outlying areas in terms of the observable quantities, i.e., as individual outliers clustered in certain areas.

Overall Bias Correction, (OBC). By using (22), we correct biases in specific outlying areas. Still, it is possible that the assumption that outliers are distributed symmetrically around a common mean may not hold. Failure of this assumption would result in an overall bias. In the simulation study reported in this paper, we correct the initial estimate by adding a robust estimate of the overall mean of residuals to each small area prediction $\hat{\bar{Y}}_{mr}^{N2+BC2}$. (Alternatively, the overall bias may be corrected by benchmarking the small area estimates to a more reliable aggregate level estimate. We did not pursue this approach in the current paper.) The data from all areas are involved in estimation of the overall bias. Thus, the OBC estimation is not a problem of small area estimation, and the assumptions may be considerably relaxed. Details follow.

Denote residuals $e_{mj}^{N2+BC2} = y_{mj} - \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}}^{N2} - \hat{u}_m^{N2} - \hat{\mu}_m^{BC2}$. Under the model assumption, errors are distributed symmetrically around zero as $N(0, \sigma_1^2)$ with probability $1-\pi$ and as $N(0, \sigma_2^2)$ with probability π. We propose to use these assumptions only in deriving a tuning parameter to curb the very extreme outliers. We then find the robust estimate of mean of the residuals e_{mj}^{N2+BC2} and use this estimate to correct the bias. If the model parameters $(\sigma_1, \sigma_2, \pi)$ are assumed known, then we can construct the following robust estimator of the mean of residuals. Let $r_{mj}^{N2+BC2} = e_{mj}^{N2+BC2} / s$, where $s = \sqrt{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}$. Note that

$$P\{r_{mj}^{N2+BC2} \leq c_{\alpha}\} = (1-\pi)\Phi\left(c_{\alpha}\frac{s}{\sigma_1}\right) + \pi\Phi\left(c_{\alpha}\frac{s}{\sigma_2}\right),$$

where $\Phi(x)$ is the standard normal distribution function. Thus, we choose a desirable precision α and obtain the "tuning parameter" c_{α} for a set of estimated parameters $(\sigma_1, \sigma_2, \pi)$.

The overall bias corrected estimator is

$$\hat{\bar{Y}}_{mr}^{N2+OBC} = \hat{\bar{Y}}_{mr}^{N2+BC2} + n^{-1}\sum_{m=1}^{M}\sum_{j=1}^{n_m} e_{mj}^{*}, \tag{24}$$

where $e_{mj}^{*} = s \cdot \min(c_{\alpha}, \max(r_{mj}^{N2+BC2}, -c_{\alpha}))$.

(In case of the standard normal distribution, the tuning parameter that equals *5* would correspond to the choice of $\alpha \approx 5 \cdot 10^{-7}$. In our simulations, we choose $\alpha \approx 10^{-6}$ to obtain a roughly comparable level of "tuning" with the standard robust estimation method that uses *5* for the tuning parameter; i.e., almost no restrictions on the estimation of the mean of residuals.)

Remark 2. We could have slightly modified the initial mixture model assumption and allow the outlying part to have a different mean. This, in our view, would contradict the definition of outlier, which is an unusual observation for a given model: In the absence of additional information in the initial model, we opt for the assumption of symmetry.

The REBLUP and MQ estimators also can be corrected using the overall bias correction; however, the OBC alone would not correct the bias in particular outlying areas. For example, the following OBC for the REBLUP (SR or Fellner's versions) estimator can be considered.

Let $e_{mj}^{REBLUP} = y_{mj} - \hat{\bar{Y}}_{mr}^{REBLUP}$, then the overall bias corrected REBLUP is

$$\hat{\bar{Y}}_{mr}^{REBLUP+OBC} = \hat{\bar{Y}}_{mr}^{REBLUP} + n^{-1} s^{REBLUP} \sum_{m=1}^{M} \sum_{j=1}^{n_m} \phi_b \left(\frac{e_{mj}^{REBLUP}}{s^{REBLUP}} \right), \tag{25}$$

where s^{REBLUP} is a robust measure of scale for the set of residuals $e_{mj}^{REBLUP}; j = 1,...,n_m, m = 1,...,M$, e.g., $s^{REBLUP} = med\left| e_{mj}^{REBLUP} - med(e_{mj}^{REBLUP}) \right| / 0.6745$ 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. Two versions of the population data are generated for 40 areas. From each area, a sample is selected using simple random sampling without replacement. The two versions are (a) each area has 100 population units and 5 sampled units or (b) each area has 300 population units and 15 sampled units. The auxiliary variable x_{mj} is generated from the lognormal distribution with mean 1.004077 and standard deviation of 0.5 and the population values y_{mj} are generated as $y_{mj} = 100 + 5x_{mj} + u_m + \varepsilon_{mj}$. There are several scenarios for distribution of u_m and ε_{mj}, as described below.

1. No contamination scenario, [0,0]: $u_m \sim N(0,3)$, $\varepsilon_{mj} \sim N(0,6)$;
2. Outlying areas, [0,u]: for the first 36 areas, $u_m \sim N(0,3)$; for the last four areas, $u_m \sim N(9,20)$; $\varepsilon_{mj} \sim N(0,6)$ for all observations;
3. Individual outliers, [e,0]: $u_m \sim N(0,3)$ for all areas; $\varepsilon_{mj} \sim N(0,6)$ with probability 0.97 and $\varepsilon_{mj} \sim N(20,150)$ with probability 0.03;
4. Individual outliers and outlying areas, [e,u]: for the first 36 areas, $u_m \sim N(0,3)$; for the last four areas, $u_m \sim N(9,20)$; $\varepsilon_{mj} \sim N(0,6)$ with probability 0.97 and $\varepsilon_{mj} \sim N(20,150)$ with probability 0.03;
5. Individual outliers only, $\varepsilon_{mj} \sim N(0,6)$ with probability 0.75 and $\varepsilon_{mj} \sim N(20,3000)$ with probability 0.25; random effects are $u_m \sim N(0,3)$. We included this version because it

somewhat resembles the CES situation (a high-peaked center of the distribution and very long tails), it is not considered in Chambers *et al.* (2009)

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, $RB = 100 \cdot med_m \{250^{-1} \sum_{s=1}^{250} (\hat{\bar{Y}}_{ms} - \bar{Y}_{ms}) \big/ 250^{-1} \sum_{s=1}^{250} \bar{Y}_{ms}\}$, and the median of the relative root mean squared error, $RRMSE = 100 \cdot med_m \left\{ \sqrt{250^{-1} \sum_{s=1}^{250} (\hat{\bar{Y}}_{ms} - \bar{Y}_{ms})^2} \Big/ 250^{-1} \sum_{s=1}^{250} \bar{Y}_{ms} \right\}$, index $s = 1,...,250$ denotes the simulation run.

First, consider scenarios 1-4 (see Tables 4.1 and 4.2). In the no-outliers situations, the estimator N2 works similar to the regular EBLUP. The bias corrected versions of N2 lost some efficiency compared to the uncorrected N2. 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. Finally, the OBC version for N2 seems to work uniformly well for all considered scenarios.

For scenario 5 (Table 4.3), 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.

5. Application to CES sample (based on Quarterly Census of Employment and Wages BLS data)

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 *m=1,...,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

$$R_{m,t} = \frac{\sum_{j \in P_{m,t}} y_{mj,t}}{\sum_{j \in P_{m,t}} y_{mj,t-1}}, \tag{26}$$

where $P_{m,t}$ is a set of the area *m* population establishments having non-zero employment in both previous and current months, i.e., $y_{mj,t-1} > 0$ and $y_{mj,t} > 0$. The direct sample estimate is

$$\hat{R}_{m,t} = \frac{\sum_{j \in S_{m,t}} w_{mj} y_{mj,t}}{\sum_{j \in S_{m,t}} w_{mj} y_{mj,t-1}}, \tag{27}$$

where $S_{m,t}$ is a set of the area m sample establishments having $y_{mj,t-1} > 0$ and $y_{mj,t} > 0$; w_{mj} is the sample weight for unit mj.

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:

$$y^*_{mj,t} = (1-\hat{f}_m)\frac{(w_{mj}-1)\hat{v}_{mj,t}}{\bar{\hat{w}}_m - 1} + \hat{R}_t + \hat{f}_m \bar{\hat{v}}_{m,t}, \tag{28}$$

where $\hat{R}_t$ is the estimated ratio of employment at a statewide level; $\hat{v}_{mj,t} = \hat{\bar{Y}}_{t-1}^{-1}(y_{mj,t} - \hat{R}_t y_{mj,t-1})$ is the estimated influence function for the ratio; $\hat{\bar{Y}}_{t-1}$ is an estimate of the previous month mean statewide employment; $\bar{w}_m = n_m^{-1}\sum_{j\in S_{m,t}} w_{mj}$ is area m average weight; $\bar{\hat{v}}_{m,t} = n_m^{-1}\sum_{j\in S_{m,t}} \hat{v}_{mj,t}$; $\hat{f}_m = \hat{N}_m^{-1} n_m$ is the estimated area sample fraction and $\hat{N}_m = \sum_{j\in S_{m,t}} w_{mj}$ 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 $R_{m,t-12}$ playing the role of an auxiliary variable (i.e., area-level auxiliary information for all observations in area m).

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 $E_{m,t} = 100\left|\hat{R}_{m,t} - R_{m,t}\right|$ and the empirical root mean squared error $ERMSE_t = \left[M^{-1}\sum_{m=1}^{M} E_{m,t}^2 \right]^{\frac{1}{2}}$.

Summaries of results for each State are reported in Tables 5.1 and 5.2. 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, 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.

Acknowledgements

I wish to thank Partha Lahiri for his useful comments and valuable discussions.

References

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

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/

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika* **93**,255-268.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2009). Outlier Robust Small Area Estimation. Invited Presentation, ISI 2009, South Africa

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

Fellner, W. H. (1986), Robust Estimation of Variance Components," T*echnometrics*, 28, 51-60.

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

Rao, J.N.K. (2003). *Small Area Estimation*, New-York, John Wiley & Sons, Inc.

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

Appendix. EM Algorithm for the scale mixture-mixed model (11)-(13)

Let $(\sigma_1^{(p)},\sigma_2^{(p)},\tau^{(p)},\pi^{(p)},\boldsymbol{\beta}^{(p)})$ be a set of parameter values after the pth iteration of EM algorithm. At the $(p+1)$th iteration, compute:

<u>*E-step*</u>

$$z_{mj}^{(p+1)}=\frac{\dfrac{\pi^{(p)}}{\sqrt{\sigma_2^{(p)2}+\tau^{(p)2}}}\exp\left(-\dfrac{1}{2}\dfrac{(y_{mj}-\mathbf{x}_{mj}^T\boldsymbol{\beta}^{(p)})^2}{\sigma_2^{(p)2}+\tau^{(p)2}}\right)}{\dfrac{1-\pi^{(p)}}{\sqrt{\sigma_1^{(p)2}+\tau^{(p)2}}}\exp\left(-\dfrac{1}{2}\dfrac{(y_{mj}-\mathbf{x}_{mj}^T\boldsymbol{\beta}^{(p)})^2}{\sigma_1^{(p)2}+\tau^{(p)2}}\right)+\dfrac{\pi^{(p)}}{\sqrt{\sigma_2^{(p)2}+\tau^{(p)2}}}\exp\left(-\dfrac{1}{2}\dfrac{(y_{mj}-\mathbf{x}_{mj}^T\boldsymbol{\beta}^{(p)})^2}{\sigma_2^{(p)2}+\tau^{(p)2}}\right)}$$

Denote

$$w_{mj}^{(p+1)}=\frac{1-z_{mj}^{(p+1)}}{\sigma_1^{(p)2}}+\frac{z_{mj}^{(p+1)}}{\sigma_2^{(p)2}};$$

$$\bar{y}_m^{(p+1)}=\left(\sum_{j=1}^{n_m}w_{mj}^{(p+1)}\right)^{-1}\sum_{j=1}^{n_m}w_{mj}^{(p+1)}y_{mj};$$

$$\bar{\mathbf{x}}_m^{(p+1)}=\left(\sum_{j=1}^{n_m}w_{mj}^{(p+1)}\right)^{-1}\sum_{j=1}^{n_m}w_{mj}^{(p+1)}\mathbf{x}_{mj};$$

$$D_m^{(p+1)}=\left(\sum_{j=1}^{n_m}w_{mj}^{(p+1)}\right)^{-1};$$

$$V_m^{(p+1)}=\left(\frac{1}{D_m^{(p+1)}}+\frac{1}{\tau^{(p)2}}\right)^{-1}.$$

Then

$$u_m^{(p+1)} = \frac{V_m^{(p+1)}}{D_m^{(p+1)}}(\bar{y}_m^{(p+1)} - \bar{\mathbf{x}}_m^{(p+1)T}\boldsymbol{\beta}^{(p)})$$

M-step

$$\pi^{(p+1)} = n^{-1}\sum_{m=1}^{M}\sum_{j=1}^{n_m} z_{mj}^{(p+1)}$$

$$\sigma_1^{(p+1)2} = \left(\sum_{m=1}^{M}\sum_{j=1}^{n_m}(1-z_{mj}^{(p+1)})\right)^{-1}\sum_{m=1}^{M}\sum_{j=1}^{n_m}(1-z_{mj}^{(p+1)})\{(y_{mj}-\mathbf{x}_{mj}^T\boldsymbol{\beta}^{(p)}-u_m^{(p+1)})^2+V_m^{(p+1)}\}$$

$$\sigma_2^{(p+1)2} = \left(\sum_{m=1}^{M}\sum_{j=1}^{n_m} z_{mj}^{(p+1)}\right)^{-1}\sum_{m=1}^{M}\sum_{j=1}^{n_m} z_{mj}^{(p+1)}\{(y_{mj}-\mathbf{x}_{mj}^T\boldsymbol{\beta}^{(p)}-u_m^{(p+1)})^2+V_m^{(p+1)}\}$$

$$\tau^{(p+1)2} = M^{-1}\sum_{m=1}^{M}(u_m^{(p+1)2}+V_m^{(p+1)})$$

$$\boldsymbol{\beta}^{(p+1)} = \left(\sum_{m=1}^{M}\sum_{j=1}^{n_m} w_{mj}^{(p+1)}\mathbf{x}_{mj}^T\mathbf{x}_{mj}\right)^{-1}\sum_{m=1}^{M}\sum_{j=1}^{n_m} w_{mj}^{(p+1)}\mathbf{x}_{mj}^T(y_{mj}-u_m^{(p+1)}).$$

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
N2	-0.001	0.068	-0.450	-0.329	-0.592	-3.528
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+BC1	-0.006	-0.006	-0.461	-0.466	0.007	-0.399
N2+BC2	-0.001	0.044	-0.451	-0.332	-0.035	-0.796
N2+OBC	-0.005	0.003	0.002	-0.153	-0.073	-0.842
SR+OBC	0.003	0.068	-0.224	-0.159	-0.558	-0.794
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

N2	0.810	0.858	1.007	0.983	1.048	4.952
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+BC1	0.920	0.922	1.121	1.125	0.865	1.032
N2+BC2	0.862	0.880	1.004	0.980	0.876	1.375
N2+OBC	0.859	0.878	0.921	0.944	0.879	1.308
SR+OBC	0.826	0.825	0.950	0.942	1.045	1.274

Table 4.2: Simulation Results Scenarios 1-4 (250 runs), Ni =300, ni = 15

	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	0.002	0.030	0.003	0.094	-0.199	-0.662
REBLUP (F)	0.002	0.029	-0.384	-0.357	-0.220	-0.585
REBLUP (SR)	0.001	0.041	-0.379	-0.334	-0.111	-0.447
MQ	0.025	0.102	-0.392	-0.288	-1.007	-0.561
N2	0.002	0.030	-0.456	-0.380	-0.204	-1.569
F+BC	0.007	0.007	-0.309	-0.312	0.001	-0.285
SR+BC	0.007	0.008	-0.308	-0.310	0.002	-0.280
MQ+BC	0.007	0.011	-0.307	-0.302	-0.057	-0.292
N2+BC1	0.008	0.008	-0.466	-0.472	0.006	-0.414
N2+BC2	0.004	0.022	-0.456	-0.381	0.000	-0.536
N2+OBC	0.003	0.007	0.006	0.028	-0.014	-0.156
SR+OBC	0.001	0.019	-0.222	-0.207	-0.132	-0.329
Median values of Relative Root MSE (expressed as a percentage)						
EBLUP	0.506	0.517	0.864	0.940	0.526	1.102
REBLUP (F)	0.517	0.520	0.682	0.675	0.540	0.802
REBLUP (SR)	0.518	0.522	0.682	0.665	0.514	0.706
MQ	0.621	0.604	0.776	0.721	1.418	1.013
N2	0.506	0.516	0.710	0.676	0.528	2.129
F+BC	0.528	0.528	0.710	0.714	0.494	0.646
SR+BC	0.528	0.528	0.709	0.714	0.493	0.645
MQ+BC	0.528	0.527	0.709	0.710	0.525	0.698
N2+BC1	0.528	0.528	0.736	0.738	0.493	0.668
N2+BC2	0.518	0.518	0.713	0.677	0.496	0.789
N2+OBC	0.517	0.518	0.571	0.575	0.496	0.605

SR+OBC	0.516	0.519	0.612	0.615	0.517	0.637

Table 4.3: Simulation Results Scenario 5 (250 runs)

	Ni =100, ni = 5		**Ni =300, ni = 15**	
Estimator	***Med Rel Bias, %***	***Med Rel Root MSE,%***	***Med Rel Bias, %***	***Med Rel Root MSE,%***
EBLUP	0.109	3.440	0.097	2.209
REBLUP (F)	-3.305	4.276	-3.419	3.784
REBLUP (SR)	-3.344	4.276	-3.413	3.821
MQ	-3.250	4.563	-3.512	3.870
N2	-3.911	4.603	-3.896	4.161
F+BC	-2.300	6.584	-3.030	3.799
SR+BC	-2.282	6.535	-3.033	3.777
MQ+BC	-2.293	6.674	-3.046	3.793
N2+BC1	-3.907	4.696	-3.900	4.175
N2+BC2	-3.896	4.598	-3.894	4.164
N2+OBC	0.136	3.041	0.115	1.731
SR+OBC	-2.647	3.813	-2.725	3.235

Table 5.1: Empirical Root Mean Squared Error, %

State	**FH**	**EBLUP**	**REBLUP (F)**	**MQ**	**N2**	**F+BC**	**MQ+BC**	**N2+BC1**	**N2+BC2**	**N2+OBC**
AL	1.868	2.257	1.899	2.023	1.808	2.027	2.133	1.798	1.791	1.873
CA	2.502	2.339	2.099	2.040	2.179	2.388	2.378	2.344	2.316	2.307
FL	3.425	2.707	2.771	3.766	1.072	2.887	3.847	1.128	1.091	1.145
PA	1.418	1.318	1.754	1.664	1.642	2.092	2.129	2.036	1.733	1.264

Table 5.2: 75th Percentile Absolute Error, %

State	**FH**	**EBLUP**	**REBLUP (F)**	**MQ**	**N2**	**F+BC**	**MQ+BC**	**N2+BC1**	**N2+BC2**	**N2+OBC**
AL	1.843	1.696	1.382	1.457	1.350	1.667	1.686	1.425	1.350	1.494
CA	1.571	1.614	1.232	1.235	1.221	1.350	1.278	1.287	1.183	1.221
FL	1.586	1.316	1.069	1.135	1.018	1.295	1.339	1.046	1.022	1.018
PA	1.165	1.185	1.276	1.315	1.388	1.667	1.729	1.613	1.397	1.083