Robust Estimation of Monthly Employment Growth Rates for Small Areas in the Current Employment Statistics Survey* October 2008

Julie Gershunskaya[1] and Partha Lahiri[2]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Suite 4985, Washington, DC, 20212
[2]University of Maryland, College Park

Abstract

Each month, the Bureau of Labor Statistics publishes estimates of employment for industrial supersectors at the metropolitan statistical area (MSA) level. The survey-weighted ratio estimator that is used to produce estimates for large domains is generally less reliable for MSA level estimation due to the unavailability of adequate sample from a given MSA. We also note that the effect of a few establishments, which are influential in terms of unusual employment numbers or sampling weights or both, could be prominent for the small area estimation. In this paper, we develop an empirical hierarchical Bayes method based on a unit level model. Our proposed method is found to be less variable and less sensitive to influential establishments when compared to the direct survey-weighted ratio estimator or estimators based on an area level model. Empirical evaluation of the estimators is performed using the population data from administrative file.

Key Words: small area estimation, robust estimation, influential observations

1. Introduction

Complex surveys are usually designed to collect enough sample units from a population of interest and to make estimates of population quantities based on this sample with a satisfactory precision. However, at a progressively finer level of detail, where the sample is sparse, direct sample based estimates could be highly imprecise. The problem of estimation at such detailed levels is known as small area estimation (SAE) problem. A model is formulated to "borrow strength" from additional sources (e.g., from the neighboring areas or historical data), to improve on the quality of the estimates for small areas. A thorough account of existing SAE methods is given in Rao (2003).

In this paper, we consider the situation when the population parameter of interest is given in *a pre-specified form*; the motivation for the form of the target does not necessarily come from a stochastic model. For example, in application considered in this paper, we are interested in estimation of the ratio of two population means: levels of employment in the current and previous months. A somewhat more complex set of examples give various forms of the price indexes, important economic indicators published from surveys conducted by national governments. The definition of the price indexes can be motivated using a range of requirements that do not assume any stochastic nature of the population measurements.

To estimate such pre-defined targets using a small sample, one may first derive estimates based on the sample and then stabilize them by applying an area-level SAE method, e.g., the Fay-Herriot model (Fay and Herriot 1979). In models of this type, assumptions are made on the area-level direct sample based estimates. In many situations, however, it is preferable to formulate model at a unit level. If the unit-level auxiliary information is available, a model that incorporates such information can be especially beneficial. However, there are reasons to consider a unit-level model even in the absence of such auxiliary data. In our application, we observe that the direct sample-based estimates can be heavily affected by the influential observations, which are common in the establishment surveys (see Hidiroglou 1994; Lee 1995; Rivest 1999; Section 3 of this paper).

One problem with the area-level model is that the sampling variance of the direct sample based estimate, considered to be known, does not always properly reflect the existence of influential observations in a particular realized sample. If this is the case, the adverse effect from such units carries over onto the resulting model estimate. On the other hand, the

* Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

effect from influential observations can be harnessed by fine-tuning the modeling assumptions on measurements at the unit level.

When the form of the finite population target is pre-specified, a unit level model is naturally obtained by considering the effect of measurements from individual units on the estimator of the given form. In infinite population settings, Hampel (1974) introduced the notion of the influence function. The influence function measures the effect of small changes in the underlying distribution on the estimator. In other words, the approach allows assessment of the effect of individual observations on the estimated target quantity by linearizing the target quantity as the increments of the underlying distribution. The approach to robust estimation based on the influence functions finds its way to estimation for finite population. Our approach is similar to that of Zaslavsky *et al* (2001) who adapt the influence function approach to linearize the target population quantity and construct estimator that is robust to appearances of influential observations.

To summarize, there are two related purposes for linearizing the estimator in the present paper. First, it provides a way to formulate a unit level small area model. Secondly, the form of the target finite population quantity, by means of its influence function, dictates which observations are to be considered as influential. Thus, the structure of the unit-level data is determined by the form of the target population parameter of interest, and the role of the model is to provide a useful and robust description of this structure.

In a typical small area estimation project, there are a few areas with relatively large sample sizes. We would like our model-based estimates to be design-consistent, that is, we would like the model-based estimates to be close to the direct estimates for these relatively large areas since they are quite reliable. This requirement of design-consistency offers some protection against possible model failure. The use of design-consistent estimators in SAE has been discussed by several researchers, including Sarndal (1984), Kott (1989), Prasad and Rao (1999), You and Rao (2003), Arora and Lahiri (1997), Folsom, Shah and Vaish (1999), Pfeffermann and Sverchkov (2007). In this paper, we employ survey weights in estimation of the target quantity. We derive the form of the estimator by exploiting the relationship between the sample and population distributions using the general approach of Sverchkov and Pfeffermann (2004). Each unit's measurement enters the estimator in combination with the unit's survey weight; thus, the impact of a unit on the estimator is determined by the combination of these two factors (see also the discussion in Zaslavsky *et al.* 2001). For this reason, the unit-level model is assumed on the weighted measurements as a whole.

Finally, flexibility in modeling assumptions and the ability to adjust the model parameters based on the data at hand give the "influential" points appropriate place in the assumed distribution. Our proposed mixture model is quite insensitive to extreme values and thus yields estimator that is robust to appearances of the influential observations. For the present paper, we assume a relatively simple model based on the mixture of two normal distributions. Modifications of this approach are, of course, possible.

The paper is organized as follows. In Section 2 we describe the general idea of the use of the influence function in constructing an estimator for a small area problem. We apply this idea to the small area estimation in the Current Employment Statistics survey conducted by the U.S. Bureau of Labor Statistics. Section 3 contains a brief description of the survey and details of the application, including the formulation of the model and results of empirical evaluation. We conclude with a brief summary.

2. Approach Using the Influence Function

Suppose $\mathbf{y} = \left(\mathbf{y}_1, ..., \mathbf{y}_{N_a}\right)$, a vector of population measurements, is a realization from the probability distribution F_a (superpopulation distribution) in area *a* (in general, each $\mathbf{y}_j$ is a vector of measurements on unit *j*); P_a is the set of population units in *a* and S_a is the set of units sampled from *a*; N_a and n_a are the number of units in P_a and S_a, respectively; $P = \bigcup_{a=1}^{m} P_a, S = \bigcup_{a=1}^{m} S_a$, and m is the number of areas.

Let F_{N_a} denote the empirical distribution function (edf) of the finite population in area *a*. Suppose we are interested in estimating the finite population quantity $T\left(F_{N_a}\right)$, a smooth function of the finite population means. We assume that $T\left(F_{N_a}\right)$ is sufficiently regular and can be linearized near F_a using the following Taylor expansion:

$$T\left(F_{N_a}\right)=T\left(F_a\right)+N_a^{-1}\sum_{j=1}^{N_a}U_{F_a,T}\left(\mathbf{y}_j\right)+O_p\left(N_a^{-1}\right), \qquad (2.1)$$

where $T\left(F_a\right)$ is a superpopulation parameter and $U_{F_a,T}\left(\mathbf{y}_j\right)$ is the *influence function* of the functional *T*; the influence function $U_{F,T}\left(x\right)$ is defined as the Gateaux derivative of T at F and is viewed as a function of x:

$$\int U_{F,T}\left(x\right)dH\left(x\right)=\lim_{\varepsilon\to 0}\frac{T\left(\left(1-\varepsilon\right)F+\varepsilon H\right)-T\left(F\right)}{\varepsilon}$$

(Hampel *et al*, 1986).

We drop the remainder term in (2.1) and *redefine* the population quantity that we target in our estimation to be

$$\breve{T}\left(F_{N_a}\right)=T\left(F_a\right)+N_a^{-1}\sum_{j=1}^{N_a}U_{F_a,T}\left(\mathbf{y}_j\right). \qquad (2.2)$$

Given the *population* size in area *a* is large, this quantity is different from the ideal target, $T\left(F_{N_a}\right)$, by a small amount, $O_p\left(N_a^{-1}\right)$.

Of course, $T\left(F_a\right)$ in (2.2) is not known and has to be estimated from the data. If sample is large enough, one could simply use some design-consistent estimator to estimate $T\left(F_a\right)$. On the other hand, the second term on the right hand side of (2.2), namely, the values $U_{F_a,T}\left(\mathbf{y}_j\right)$, can provide an insight into the sample influential observations.

For small area estimation, the direct estimator is not reliable and it is common practice to make a suitable assumption about the proximity of the area levels to the aggregation of areas. Let F denote the distribution function of the population measurements in $P=\bigcup_{a=1}^{m}P_a$, the aggregation of all areas. Suppose $T\left(F_{N_a}\right)$ can be expanded in the neighborhood of F as

$$T\left(F_{N_a}\right)=T\left(F\right)\ +c_a^{-1}N_a^{-1}\sum_{j=1}^{N_a}U_{F,T}\left(\mathbf{y}_j\right)+R_{Na} \qquad (2.3)$$

for some c_a, where $\sum_{a=1}^{m}p_a c_a=1, p_a=N_a/N, N=\sum_{a=1}^{m}N_a$, and R_{Na} is the remainder term.

Example. Consider $T\left(F_{N_a}\right)\equiv\frac{\bar{Y}_{1a}}{\bar{Y}_{2a}}$, ratio of two population means $\bar{Y}_{1a}=N_a^{-1}\sum_{j=1}^{N_a}y_{1j}$ and $\bar{Y}_{2a}=N_a^{-1}\sum_{j=1}^{N_a}y_{2j}$. Let the superpopulation parameter in the aggregation of areas be a function of superpopulation means θ_1 and θ_2,

$T(F)=\dfrac{\theta_1}{\theta_2}$. We can define $c_a=\bar{Y}_{2a}/\theta_2$. In this particular case, the influence function is given by $U_{F,T}(\mathbf{y}_j)=\dfrac{1}{\theta_2}\left(y_{1j}-\dfrac{\theta_1}{\theta_2}y_{2j}\right)$, $\mathbf{y}_j=(y_{1j},y_{2j})'$ and the remainder term vanishes, $R_{Na}=0$.

In general, the closeness of F_a to F is implied by assumption that the remainder term in (2.3) is small. Then, similar to (2.2), we can redefine the target population parameter by dropping the remainder term:

$$\tilde{T}(F_{N_a})=T(F)\ +c_a^{-1}N_a^{-1}\sum_{j=1}^{N_a}U_{F,T}(\mathbf{y}_{aj}). \tag{2.4}$$

Remark 1: Note that if $R_{Na}=O_p(N_a^{-1})$, the weighted sum of the area levels is close to *T* on the aggregation of areas, which is a desirable property:

$$\sum_{a=1}^{m}p_a c_a T(F_{N_a})-T(F_N)=o_p(1). \tag{2.5}$$

Remark 2: In what follows, we consider a particular case when $c_a=1$. Even though this may lead to a larger discrepancy between the original and modified targets, on the other hand, since c_a in general depends on the area-level quantities, the necessity to estimate these quantities from small samples, in practice, may lead to a larger error. We can assess the approximation by considering the direct sample based estimate of the difference $R_{Na}=\tilde{T}(F_{N_a})-T(F_{N_a})$. For example, using an area-level SAE model, we can derive an estimator of R_{Na} and then use it to adjust the estimate of $\tilde{T}(F_{N_a})$. (This approach was not pursued in the present paper. Instead, after examining the scatter plot of the direct sample based estimates of R_{Na}, we make an assumption that R_{Na} is small and can be neglected)

Next, we discuss the estimation of $\tilde{T}(F_{N_a})$. First, let us re-write (2.4) as:

$$\tilde{T}(F_{N_a})=T(F)+f_a\bar{U}_{S_a}+(1-f_a)\bar{U}_{S_a^C}, \tag{2.6}$$

where

$f_a=n_a/N_a$ is the sampling fraction in area *a*,

$$\bar{U}_{S_a}=\frac{1}{n_a}\sum_{j\in S_a}U_{F,T}(\mathbf{y}_j) \text{ and} \tag{2.7}$$

$$\bar{U}_{S_a^C}=\frac{1}{N_a-n_a}\sum_{j\notin S_a}U_{F,T}(\mathbf{y}_j) \tag{2.8}$$

are means of the influence function for observations that are included, (2.7), and not included, (2.8), in the sample.

Let us suppose for a moment that we know the value of $T(F)$. Under the prediction approach to inferences in sampling from finite populations, the goal is to predict values in the non-sampled part of the population, S_a^C (also referred to as the *sample-complement*), using the sample measurements. In our formulation, the problem is to predict the value of $\bar{U}_{S_a^C}$.

The distribution of the sample measurements may differ from the distribution of the population measurements. If this is the case, it is important to account for the difference in order to avoid the estimation bias. Sverchkov and Pfeffermann (2004) established the relationship between the distributions of values in the sample and sample-complement parts of the population. The formula that links the two distributions accounts for the probabilities of units to be included in the sample. One of the corollaries of a more general expression is the formula relating the expectations under the sample and sample-complement distributions using the sampling weights w_j. The sampling weight w_j is defined as the inverse of π_j, the inclusion probability. We assume that π_j is a realization of a random variable, a function of variables used to design the survey. For a general pairs of a random vector variables $\left(\mathbf{u}_j, \mathbf{v}_j\right)$,

$$E_C\left[\mathbf{u}_j \mid \mathbf{v}_j\right] = \frac{E_S\left[\left(w_j - 1\right)\mathbf{u}_j \mid \mathbf{v}_j\right]}{E_S\left[w_j - 1 \mid \mathbf{v}_j\right]} \tag{2.9}$$

where E_C and E_S are expectations under the sample and sample-complement distributions, respectively. In our case, to predict $\bar{U}_{S_a^C}$, we need to estimate the sample-complement expectation $E_C\left[U_{F,T}\left(\mathbf{y}_j\right) \mid j \in S_a^C\right]$ from the sample.

Using (2.9), the population quantity can be expressed as

$$\tilde{T}\left(F_{N_a}\right) = T\left(F\right) + f_a \bar{U}_{S_a} + \left(1 - f_a\right) E_S\left[\frac{\left(w_j - 1\right)}{E_S\left[w_j - 1\right]} U_{F,T}\left(\mathbf{y}_j\right) \mid j \in S_a\right]. \tag{2.10}$$

Since $T\left(F\right)$ is defined on the aggregation of areas, it can be estimated from the sample with a satisfactory precision and plugged into the formula (2.10). This is the approach considered in the present paper. The estimator of $\tilde{T}\left(F_{N_a}\right)$ takes the form

$$\hat{\tilde{T}}\left(F_{N_a}\right) = \hat{T}\left(F_N\right) + f_a \frac{1}{n_a}\sum_{j=1}^{n} \hat{u}_j + \left(1 - f_a\right)\frac{E_S\left[\left(w_j - 1\right)\hat{u}_j \mid j \in S_a\right]}{E_S\left[w_j - 1 \mid j \in S_a\right]}, \tag{2.11}$$

where $\hat{u}_j$ is an estimate of $U_{F,T}\left(y_i\right)$ and depends on the choice of $\hat{T}\left(F_N\right)$.

Some modeling methods can be used to find an estimator for the last term in (2.11). In addition, auxiliary information, if available, can also be used in the modeling.

In the present paper, we consider two methods for treating the last term in (2.11). The first method uses the assumption of normality for $\hat{u}_j^w = \left(w_j - 1\right)\hat{u}_j$. Alternatively, we assume that $\hat{u}_j^w$ are the realizations from the mixture of two normal distributions with different variances but means varying only by areas and not by parts of the mixture. The second approach, or its possible extension to the case of the mixture of more than two distributions with different variances, in a sense, is a robust approach to estimation and can be useful in dealing with a nonstandard distribution of $\hat{u}_j^w$.

Note that we have treated weights as random variables and combined $\hat{u}_j$ and w_j into a single random variable $\hat{u}_j^w$ for our modeling purpose. By treating the survey weights as random variables we allow for the simultaneous treatment of the outlying survey weights, measurements, or of their combined effect.

3. Application to CES

Current Employment Statistics (CES) is a large-scale establishment survey conducted by the U.S. Bureau of Labor Statistics. The survey produces monthly estimates of employment and other important indicators of the U.S. economy. The estimates are published every month at various levels of industrial and geographical detail.

While estimation of the National and State level employment is of a central importance in CES, there is also a lot of interest in publication of estimates for the smaller domains defined at a finer industrial and geographical detail. At these levels, the sample is often sparse and a single influential observation, if left untreated, may affect the resulting estimates.

To facilitate the discussion, we describe briefly relevant details of the CES sample selection and estimation methods.

3.1 CES Frame and Sample Selection

The CES sample is selected once a year from a frame based on the Quarterly Census of Employment and Wages (QCEW) data file. This is the administrative dataset that contains records of employment and wages for nearly every U.S. establishment covered by the States' unemployment insurance (UI) laws. The QCEW data becomes available to the BLS on a lagged basis and serves not only for the sampling selection but also for the benchmarking purposes; the historical QCEW data is also a valuable source for researchers (see *BLS Handbook of Methods*, 2004, for more information about QCEW).

The frame is divided into strata defined by State, industrial supersector based on the North American Industrial Classification System (NAICS) and on the total employment size of establishments within a UI account. A stratified simple random sample of UI accounts is selected using optimal allocation to minimize, for a given cost per State, a State level variance of the monthly employment change estimate.

3.2 CES Estimator

The relative growth of employment from the previous to the current month is estimated using a matched sample S_t of establishments, that is, establishments reporting positive employment in both adjacent months:

$\hat{R}_t = \dfrac{\sum_{j \in S_t} w_j y_{j;t}}{\sum_{j \in S_t} w_j y_{j;t-1}}$, where the suffixes *j* and *t* denote the establishment and the current month, respectively.

The numerator of the ratio is the survey weighted sum of the current month reported employment; similarly, the denominator is the survey weighted sum of the previous month employment.
Once a year, an estimate is benchmarked to a census level figure (that becomes available on a lagged basis): $\hat{Y}_{t=1} = Y_0 \hat{R}_{t=1}$; monthly estimates of the employment level at subsequent months are derived by application of estimate $\hat{R}_t$ of employment trend to the previous month estimate of the employment level: $\hat{Y}_t = \hat{Y}_{t-1} \hat{R}_t$. See the *BLS Handbook of Methods* (2004, Chapter 2) for further details.

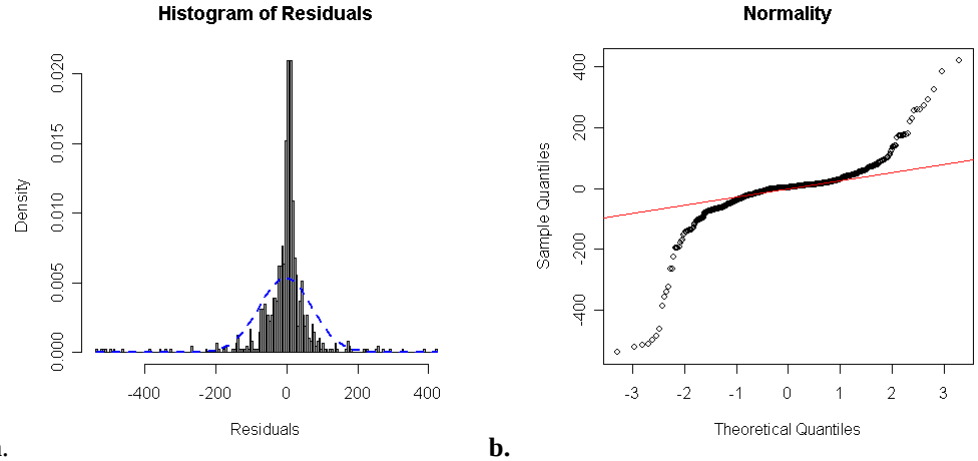
3.3 Influential Observations in CES

As is common for many establishment surveys, the CES sample often contains a small fraction of observations that may seriously affect the survey estimate. Establishments in the target population vary greatly by size. The population consists of a relatively small number of large companies, but most of the national employment is associated with small-size enterprises. Businesses are disproportionately selected into the sample, and the resulting survey weights are very heterogeneous; hence, the survey weighted estimators may become very unstable for some small domains.

Another aspect of a survey of businesses is the potential change in the establishment attributes that are used for sample selection. For example, the establishment employment level may change after the sample has been selected. As a result, a larger than expected at the time of sampling employment size becomes associated with a large survey weight creating a predisposition for the influential observation.

A definition of influential observation must be tied to the form of the estimator. In a given month, CES estimates relative employment growth, the ratio of the two survey weighted sums. An influential report would have either

relatively large survey weight or large change in the size of its employment, the combination of moderately large weight with employment change may also produce influential report. In any given month, there are generally a handful of observations that stand out from the rest of the sample. One reason is the form of the distribution of employment changes: a large number of the establishments do not add or reduce the number of employees; some have a very little change in their employment. However, there are always units that have a substantial change in employment and at times they also have a large survey weight. The histogram of establishments employment change cannot be described by a nice bell-shaped distribution: it has a spike around zero with very long tails (see Fig.1). The sample is prone to outliers in the sense that there is a high probability that a handful of observations from the tails of the distribution are present in the sample.



a. **b.**

Fig.1. a. Histogram of residuals overlayed by the density of the normal distribution with the same first two moments as the estimated moments of the residuals distribution. **b.** Normal QQ plot of residuals.

3.4 Small Area Estimation in CES

The goal is to estimate the employment trend R_t^a at a given month t in areas $a=1,\ldots,m$, where areas are defined by intersection of industry and geography at MSA level within a given State.

For area a, the target finite population quantity at month t is given by

$$R_t^a = \frac{\sum_{j \in P_t^a} y_{j,t}}{\sum_{j \in P_t^a} y_{j,t-1}}, \qquad (3.1)$$

where P_t^a is a set of area a population establishments having non-zero employment in both previous (t-1) and current (t) months.

Assume the set of finite population observations at month t $\left\{\left(y_{j,t-1}, y_{j,t}\right) \mid j \in P_t\right\}$ (here, $P_t = \bigcup_{a=1}^{m} P_t^a$) to be independent realizations of a random vector $\left(Y_{t-1}, Y_t\right)$ having probability distribution F. Denote $\left(\theta_{t-1}, \theta_t\right)$ a vector of means of $\left(Y_{t-1}, Y_t\right)$. Let the population measurements in area a, $\left\{\left(y_{j,t-1}, y_{j,t}\right) \mid j \in P_t^a\right\}$, be independent realizations of a random vector $\left(Y_{t-1}, Y_t\right)$ with the probability distribution F_a and let $\left(\theta_{t-1}^a, \theta_t^a\right)$ be a vector of corresponding means . The superpopulation parameter of interest is a function of superpopulation means $\left(\theta_{t-1}^a, \theta_t^a\right)$:

$$T\left(F_a\right) = T\left(\theta_{t-1}^a, \theta_t^a; F_a\right) = \frac{\theta_t^a}{\theta_{t-1}^a}.$$

The influence function for T defined at F is given by

$$U_{F,T}\left(y_{j,t-1}, y_{j,t}\right) = \frac{1}{\theta_{t-1}}\left(y_{j,t} - \frac{\theta_t}{\theta_{t-1}} y_{j,t-1}\right). \tag{3.2}$$

As for $\hat{T}\left(F_N\right)$ appeared in the formula (2.11), we propose to apply the following survey weighted estimator:

$$\hat{R}_t = \frac{\hat{\theta}_t}{\hat{\theta}_{t-1}} = \frac{\sum_{j\in S_t} w_j y_{j;t}}{\sum_{j\in S_t} w_j y_{j;t-1}}, \tag{3.3}$$

using sample units from all areas, $S_t = \bigcup_{a=1}^{m} S_t^a$. This is the same form of the estimator that is normally used for large areas in the CES. The estimates of means are given by $\hat{\theta}_{t-1} = \frac{\sum_{j\in S_t} w_j y_{j;t-1}}{\sum_{j\in S_t} w_j}$ and $\hat{\theta}_t = \frac{\sum_{j\in S_t} w_j y_{j;t}}{\sum_{j\in S_t} w_j}$.

The number of population units having non-zero employment in two consecutive months is not known and is estimated by $\hat{N}_a = \sum_{j\in S_t^a} w_j$, The sampling fraction is estimated by

$$\hat{f}_a = \frac{n_a}{\hat{N}_a}. \tag{3.4}$$

In this application, formula (2.11) reduces to:

$$\hat{\bar{R}}_t^a = \hat{R}_t + \hat{f}_a \frac{1}{n_a}\sum_{j=1}^{n_a} \hat{u}_{jt} + \left(1-\hat{f}_a\right)\frac{E_S\left[\hat{u}_{jt}^w \mid j \in S_t^a\right]}{\bar{w}_a - 1}, \tag{3.5}$$

where $\hat{u}_{jt} = \frac{1}{\hat{\theta}_{t-1}}\left(y_{j,t} - \hat{R}_t y_{j,t-1}\right); \hat{u}_{jt}^w = \left(w_j - 1\right)\hat{u}_{jt}; \bar{w}_a = E_S\left[w_j \mid j \in S_t^a\right]$.

As displayed on Fig.1, the distribution of $\hat{u}_{jt}^w$ is nearly symmetric; it does not appear to be normal and has a spike near 0 and long tails. A large number of establishments do not have changes in their employment level in a given two consecutive months; others change only by a few employees. This can explain the observed spike. However, there are also establishments that change by quite a large number of employees. The change can be magnified by application of the sampling weights. This economic phenomenon explains the long tails of the distribution.

From the above discussion, it is clear that the methods that assume normality or some other standard distribution may result in inefficient estimates. We attend to the problem by assuming the underlying distribution is a mixture of normal distributions and estimating unknown parameters of the distribution using the Bayesian normal mixture model. The MCMC utilizing Gibbs sampler provides a tool for this estimation. At present, we display results of the estimation under the assumption of the mixture of two normal distributions. We also considered the mixture of three and four distributions. The results turned out to be very similar to the two distributions version. On the other hand, the computational burden and monitoring for the convergence proved the approach with greater than two mixture parts to be infeasible in the situation of creating estimates for numerous domains in a short time period. Nevertheless, we formulate the model used in the estimation for the general case of an arbitrary number of mixtures.

Denote x_t^a an area-level auxiliary information (at present, we use historical QCEW data: $x_t^a = R_{t-12}^a$). We tested the following unit- and area-level models:

Model 1 (Mixture of Normal):

$$\hat{u}_{jt}^{w} \mid j \in S_t^a, C_{jt} = k, \mu_t^a, \sigma_k^2 \overset{ind}{\sim} N\left(\mu_t^a, \sigma_k^2\right), \text{where } \mu_t^a = \frac{\left(\bar{w}_a - 1\right)\left(R_t^a - R_t - f_a \bar{U}_{S_a}\right)}{\left(1 - f_a\right)}$$

$$R_t^a \overset{ind}{\sim} N\left(\beta x_t^a, \tau^2\right)$$

$$C_{jt} \sim Multi\left(1, \pi_1, ..., \pi_K\right)$$

$$\pi_k \sim Dirichlet\left(p, ..., p\right)$$

where $a = 1, ..., m;\ j = 1, ..., n$

K is the number of components of the normal mixture;

π_k is probability to belong to k^{th} component of the mixture, $k = 1, ..., K$;

C_{jt} is class indicator for observation i, which can take the values $1, ..., K$;

μ_t^a is common mean of the mixture distribution for observations from area a;

σ_k^2 is variance parameter of the k^{th} component of the mixture, common for all areas and

p is prior probability to belong to component k

Model 2 (Normal):

$$\hat{u}_{jt}^{w} \mid j \in S_t^a, \mu_t^a, \sigma^2 \overset{ind}{\sim} N\left(\mu_t^a, \sigma^2\right), \text{where } \mu_t^a = \frac{\left(\bar{w}_a - 1\right)\left(R_t^a - R_t - f_a \bar{U}_{S_a}\right)}{\left(1 - f_a\right)}$$

$$R_t^a \overset{ind}{\sim} N\left(\beta x_t^a, \tau^2\right)$$

where $a = 1, ..., m;\ j = 1, ..., n$

Model 3 (Fay-Herriot):

$$\hat{R}_t^a \mid R_t^a \overset{ind}{\sim} N\left(R_t^a, \sigma_a^2\right)$$

$$R_t^a \overset{ind}{\sim} N\left(\beta x_t^a, \tau^2\right)$$

where $a = 1, ..., m$

In models 1 and 2, for the assumed-to-be-known values of $\bar{w}_a$, R_t, f_a, we "plug in" their direct sample based estimates, so that $\mu_t^a = \dfrac{\left(\hat{\bar{w}}_a - 1\right)\left(R_t^a - \hat{R}_t - \hat{f}_a \dfrac{1}{n_a}\sum_{j=1}^{n_a} \hat{u}_{jt}\right)}{\left(1 - \hat{f}_a\right)}$, where $\hat{\bar{w}}_a = \dfrac{1}{\tilde{n}_a}\sum_{\substack{j \in S_t^a \\ w_j \neq 1}} w_j, \tilde{n}_a = \sum_{j \in S_t^a} I_{w_j \neq 1}$, and $\hat{R}_t$ and $\hat{f}_a$ are given by (3.3) and (3.4), respectively. The method of estimation can be labelled "the empirical hierarchical Bayes".

We used the QCEW historical data to test the approach. We created estimates for September 2006 using the sample drawn from 2005 sampling frame, mimicking the production timeline. Relative growth rates (3.1) were estimated for four States (Alabama, California, Florida, and Pennsylvania). Small areas are defined as metropolitan statistical areas (MSA) at industrial supersector level. In the modeling, each industry was considered separately. The number of areas, *m*, varied by industries in the four States, ranging from *m*=12 to *m*=27. The number of units in the matched sample ranged from 2 to 969 across areas. The resulting estimates were compared to the corresponding true population quantities derived from QCEW.

In model 3, we used true design-based variances σ_a^2 of the direct sample estimator $\hat{R}_t^a$. For research purposes, when working with the historical data, we can derive the true variances using all population units. The variances σ_a^2 are assumed to be known in Fay-Herriot model. Of course, we do not have the true values of these variances in real time and would have used the smoothed values of the direct sample estimates of σ_a^2.

The empirical mean squared error (MSE) was calculated for each industry for the following methods:
(1) Model 1 (unit-level, the mixture of two normal distributions);
(2) Model 2 (unit-level, the normality assumption);
(3) Model 3 (area-level, Fay-Herriot model);
(4) the direct sample based estimates of the modified target, $\hat{\tilde{R}}_t^a$;
(5) the direct sample based estimates for the original form of the population target, $\hat{R}_t^a$.

$$MSE_{(k)t} = \frac{1}{m}\sum_{a=1}^{m}\left[100\left(\hat{R}_{(k)t}^a - R_t^a\right)\right]^2,$$

where R_t^a is the "true" value obtained from QCEW at month *t* in area *a* and industry *i* (index *i* is suppressed, for convenience); $\hat{R}_{(k)t}^a$ is the estimate of R_t^a obtained using method k, $k = 1,...,5$.

Below are the resulting tables for four states, by industry and overall (column '*n*' contains the total number of sample units in a given industry):

Alabama

Industry	n	Mixture	Normal	FH	Direct (Linear)	Direct
Construction	421	1.62	7.94	3.82	21.22	45.49
Manufacturing Durable Goods	237	0.72	1.13	0.83	1.57	1.18
Manufacturing Nondurable Goods	109	0.37	1.22	0.79	2.07	1.32
Wholesale Trade	231	1.55	1.45	2.83	11.39	12.74
Retail Trade	462	0.64	1.37	1.17	1.42	1.42
Transportation, Warehousing, Utilities	125	2.46	3.48	2.23	4.08	3.54
Information	70	2.00	1.05	1.36	7.80	5.19
Finance, Insurance, Real Estate and Rental and Leasing	238	1.51	4.85	4.13	4.12	4.86
Professional and Business Services	659	0.74	2.82	3.47	2.18	2.52
Education and Health Services	379	0.38	1.28	1.46	1.88	1.98
Leisure and Hospitality	420	0.88	4.09	1.93	12.21	18.60
Other Services	181	25.06	37.45	21.57	86.53	90.80
Overall		**3.16**	**5.68**	**3.80**	**13.04**	**15.80**

California

Industry	n	Mixture	Normal	FH	Direct (Linear)	Direct
Construction	2253	2.45	2.98	3.32	11.33	15.91
Manufacturing Durable Goods	906	2.28	2.65	3.24	6.10	9.53
Manufacturing Nondurable Goods	801	22.57	28.04	23.74	43.91	59.92
Wholesale Trade	1410	0.94	4.68	4.85	22.63	13.24
Retail Trade	1819	0.15	0.89	1.30	2.54	4.24

Transportation, Warehousing, Utilities	643	6.63	8.80	19.24	35.84	68.27
Information	410	0.97	1.66	2.93	1.85	7.01
Finance, Insurance, Real Estate and Rental and Leasing	1372	0.97	1.20	4.75	19.23	9.70
Professional and Business Services	3772	1.44	2.59	2.64	8.73	4.91
Education and Health Services	2147	0.53	1.68	3.46	4.97	5.83
Leisure and Hospitality	2466	1.77	2.30	2.34	4.91	5.10
Other Services	969	2.98	3.59	4.08	27.85	30.56
Overall		**3.64**	**5.09**	**6.32**	**15.82**	**19.52**

Florida

Industry	n	Mixture	Normal	FH	Direct (Linear)	Direct
Construction	769	0.58	4.34	2.97	8.23	7.90
Manufacturing Durable Goods	161	1.19	1.60	2.89	7.42	8.14
Manufacturing Nondurable Goods	86	4.93	2.99	4.52	138.79	15.30
Wholesale Trade	435	0.85	5.17	14.20	44.60	24.82
Retail Trade	652	0.08	0.30	0.34	0.36	0.51
Transportation, Warehousing, Utilities	188	1.63	2.12	4.23	14.50	18.22
Information	112	2.10	4.46	103.42	43.46	112.59
Finance, Insurance, Real Estate and Rental and Leasing	542	0.61	0.62	1.82	4.37	6.00
Professional and Business Services	1353	0.49	0.75	1.39	1.29	5.08
Education and Health Services	723	0.38	0.48	1.09	3.18	3.38
Leisure and Hospitality	689	1.80	2.73	3.53	17.17	13.33
Other Services	327	1.14	8.17	12.07	130.62	67.39
Overall		**1.32**	**2.81**	**12.71**	**34.50**	**23.55**

Pennsylvania

Industry	n	Mixture	Normal	FH	Direct (Linear)	Direct
Construction	657	3.80	2.00	4.69	19.97	24.49
Manufacturing Durable Goods	354	0.69	1.45	1.13	7.35	4.82
Manufacturing Nondurable Goods	215	1.80	1.63	1.10	21.15	5.33
Wholesale Trade	405	0.65	1.17	1.27	11.17	5.84
Retail Trade	691	0.71	0.40	0.83	1.76	2.99
Transportation, Warehousing, Utilities	282	20.41	4.18	2.21	25.30	27.51
Information	140	1.62	1.35	1.72	3.24	3.44
Finance, Insurance, Real Estate and Rental and Leasing	357	1.08	1.30	8.66	6.16	17.22
Professional and Business Services	1190	0.95	1.40	1.15	5.68	6.72
Education and Health Services	939	0.76	0.37	0.31	1.60	1.63
Leisure and Hospitality	951	7.13	2.88	2.23	11.79	10.85
Other Services	413	2.64	7.73	4.22	35.67	31.98
Overall		**3.52**	**2.16**	**2.46**	**12.57**	**11.90**

The values of the empirical mean squared errors of the direct sample based estimators for the original and modified targets are of a similar magnitude. All SAE methods considered help to reduce MSE of the direct estimators. Area-level Fay-Herriot estimator works satisfactory except for cases where it is affected by the outlying values (e.g., Florida, Information). Overall, either unit level model has lower MSE than the area level, and in three of the four states considered here, the performance of the mixture model is better. A smaller error in Normal and FH models (compared to the mixture model) for Transportation, Warehousing, Utilities industry of Pennsylvania can be explained by the fact that the historical values of the trend ($x_t^a = R_{t-12}^a$) seem to predict well the current trend, yet this predictor receives smaller weight in the mixture model. The model parameters can be adjusted to better account for the degree of reliability of the historical trends.

Summary

In this paper we proposed a unit level small area model for the case when the finite population quantity of interest has a nonlinear form. The proposed empirical hierarchical Bayesian method that uses a mixture of normal distributions to model the unit level observations has a potential for a useful tool in dealing with influential observations and non-standard distributions. However, more work is required to overcome practical difficulties associated with the efficient computations in the production environment. One example of the trade-off between efficiency of the procedure and practical considerations is the choice of the number of mixture groups. Larger number of mixture groups may be crucial for the processing time, yet have very little effect on the resulting estimates. Variance of the proposed estimator can be estimated using, for example, the parametric bootstrap method. Work will continue in this direction.

References

Arora, V., and Lahiri, P. (1997), "On the Superiority of the Bayesian Method Over the BLUP in Small Area Estimation Problems," *Statistica Sinica*, 7, 1053–1063

Bureau of Labor Statistics (2004), Chapter 2, "Employment, hours, and earnings from the Establishment survey," *BLS Handbook of Methods*. Washington, DC: U.S. Department of Labor.

Fay, R.E. and Herriot, (1979). Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data, *Journal of American Statistical Association*, 74, 269-277

Folsom, R., Shah, B., Vaish, A. (1999). Substance abuse in states: a methodological report on model based
estimates from the 1994–1996 National household survey on drug abuse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 371–375

Hampel, F. (1974). The influence curve and its role in robust estimation, *Journal of American Statistical Association*, 69, 383–393

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*. New-York, John Wiley & Sns, Inc.

Hidiroglou, M. (1994) Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 153-162

Kott, P. (1989). Robust small domain estimation using random effects modelling. *Survey Methodology*, 15, 1–12

Lee, H. (1995) Outliers in Business Surveys. *In Business Survey Methods* edited by B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. Colledge, and P. S. Kott, p. 503-526

Pfeffermann, D. and Sverchkov, M. (2007). Small area estimation under informative probability sampling of areas and within the selected areas, *Journal of American Statistical Association*, 102, 1427-1439

Prasad, N.G.N., Rao, J.N.K. (1999). On robust small area estimation using a simple random e.ects model. *Survey Methodology*, 25, 67–72

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

Rivest, L.-P. (1999). Stratum jumpers: Can we avoid them? *Proceedings of the Survey Research Methods of the American Statistical Association*, pp. 64-72.

Sarndal, C. E. (1984). Design-Consistent Versus Model-Dependent Estimation for Small Domains, *Journal of the American Statistical Association*, Vol. 79, No. 387, pp. 624- 631

Sverchkov, M., and Pfeffermann, D. (2004), "Prediction of Finite Population Totals Based on the Sample Distribution," *Survey Methodology*, 30, 79–92.

Zaslavsky, A.M., Schenker, N. and Belin, T.R. (2001). Downweighting influential clusters in surveys: application to the 1990 post enumeration survey. *Journal of the American Statistical Association*, 96, 858-869.

You, Y. and Rao, J. N. K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 111, 197-208.