On Weight Smoothing in the Current Employment Statistics Survey October 2014

Julie Gershunskaya, Michael Sverchkov

Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington, DC, 20212

Abstract

The sampling weight in the Current Employment Statistics Survey is determined at the time of sample selection. It depends on a unit's State, industry, and size class. However, the population of businesses is highly dynamic. Establishments constantly grow or contract; sometimes they also change their industrial classification or geographical location. Even the number of population units is not fixed but continuously changes over time. A unit may change its size class at the time of estimation or the content of the original stratum may change. Under such circumstances, application of the original survey weights may increase volatility of survey estimates. In this paper we investigate if the survey estimates can be improved by adjusting the original weights.

Key Words: sampling weights, extreme observation, business survey, stratum jumper

1. Introduction

Under the classical design-based approach to inferences from survey sampling, the sampling weights are defined at the design stage of a survey and viewed as *non-random* quantities at the estimation stage. In contrast, Pfeffermann and Sverchkov (1999) and Beaumont (2008) view the sampling weights as realizations of a *random* vector. This allows modeling the weights and applying a new "smoothed" set of weights estimated from the model. The approach has the potential for the improved efficiency compared to the estimator based on the original weights.

Consider a version of the expansion estimator for the population total, where a set of *smoothed* weights is used in place of the original sampling weights. The method has the potential to give good results when the new weights are better related to the response variable.

The usual layperson's interpretation of the sampling weights in the expansion estimator goes as follows: think of sample unit's weight as the number of corresponding units in the population having the same value of the sample unit's response variable. If this correspondence would hold exactly for all sample units, then the sample weighted estimator provides a perfect estimator of the population total. It "estimates" the total without an error.

Of course, in reality, sampling weights never exactly represent the number of such units in the population. Each weight may be considered as *an estimate* of the number of population units with like values. One can try to improve this estimate by exploiting the relationship between weights and sample responses and finding an "average" value of the weight for units with similar measurements. Thus, one is *smoothing* the weights.

The theoretical idea is promising; however, the method depends on finding an appropriate model for the weights. In practice, the choice of a good model may be challenging, and the model failure may lead to a

bias in estimation. With a good model also, the model parameters need to be estimated from the data, and this contributes to the variance of the resulting survey estimator. Keeping in mind these practical difficulties, application of the method needs to be thoroughly tested.

In this paper we consider a nonparametric approach to estimation of the smoothed weights based on the values of response variables. The nonparametric approach does not require explicitly formulating a model; a drawback is that the nonparametric estimation, generally, is less efficient than the parametric approach. We apply the method to estimation in the Current Employment Statistics (CES) survey and compare results with the currently used estimator.

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. Here, we consider estimation for the one month relative employment change for industrial divisions (supersectors) in metropolitan statistical areas (MSA).

In Section 2, we give a brief description of relevant details of the CES sample selection and estimation methods and provide motivation for considering the weights smoothing method. In Section 3 we adapt the theoretical concepts developed by Pfeffermann and Sverchkov (1999, 2003) and Beaumont (2008) to the case of the CES estimator of the relative over-the-month change. We describe the proposed estimators for the CES survey, the evaluation criteria, and provide the results in Section 4. The last section contains the summary.

2. Details of the CES survey

2.1 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 BLS on a lagged basis and serves for the sampling selection and for the benchmarking purposes; (see *BLS Handbook of Methods*, http://www.bls.gov/opub/hom/pdf/homch2.pdf, for more information about QCEW).

The QCEW based sampling 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.

2.2. 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:

$$\hat{R}_t = \frac{\sum_{j \in S_t} w_j y_{j;t}}{\sum_{j \in S_t} w_j y_{j;t-1}}, \qquad (1)$$

where *j* denotes establishment, *t* is the current month, $y_{j;t}$ and $y_{j;t-1}$ denote, respectively, a unit's current and previous months reported employment; w_j is the selection weight.

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 Y_0 (the QCEW-based level 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* (Chapter 2) for further details.

2.3 Motivation for weights adjustment and treatment of influential observations in CES

In CES, every month, we are essentially measuring employment change in the population. Thus, we should be looking at the relationship between the weights and employment changes. More precisely, because we are using the ratio estimator and estimating the relative change, we should consider the relationship between the weights and residuals $r_{t,j} = y_{t,j} - R_t y_{t-1,j}$.

Indeed, by the first order Taylor decomposition, (1) can be approximated as,

$$\hat{R}_t \doteq R_t + \frac{1}{Y_{t-1}} \sum_{j \in S_t} w_j (y_{j;t} - R_t y_{j;t-1}), \qquad (2)$$

where Y_{t-1} and R_t are total employment in month *t-1* and relative growth of employment from month *t-1* to month *t*.

As described in Section 2.1, CES stratifies based on the employment size (within industry and geography) and allocates optimally for a given cost. This strategy is intended to produce an efficient sample weighted estimator. Given how we sample, larger weights are usually associated with smaller businesses. The smaller businesses also tend to have smaller changes in employment and thus smaller unweighted residuals.

However, we often observe a relatively large weight associated with a large change in employment. This happens for various reasons. The general explanation is the dynamic nature of the population of businesses (businesses may jump from one size class to another; the number of population units may change; labels, such as industrial classification, also change during the estimation period). Even with the optimal sampling design, one cannot account for the future changes in the population at the time of planning and selecting the sample. Therefore, the design weights are hardly optimal for any given month of the estimation.

The Robust Estimation procedure is the estimation method currently used in CES. It is designed to reduce the effect of the influential observations on the estimate of the relative over-the-month change. The Robust estimator identifies a limited number of units having extreme values of weighted residuals. These units receive special treatment: their weights are reduced; in the most egregious cases, they are considered self-representing atypical units and are removed from the formula (1).

From (2), the influential reports are those having large positive or negative values of the weighted residuals, $w_j(y_{j;t} - R_t y_{j;t-1})$, compared to the other sample units. The extreme residuals are reduced to specific cut-off values. The cut-off values depend on the distribution of the weighted residuals in a given series and are determined independently for each month and industry series. Pushing the extreme residuals to the cut-off values is accomplished by using an appropriate weight adjustment factor.

The procedure used for the CES robust estimation is a particular variation of a general method of weight reduction known as Winsorization. See Kokic and Bell (1994), Gershunskaya (2011). The actual cut-off values are determined by examining the relative distances of units with extreme weighted residuals to the nearest but less extreme values in the same cell and month. See the *BLS Handbook of Methods* (Chapter 2) for further details.

This procedure helps to reduce volatility of the estimator. Still, especially at the lower estimation cell levels, the estimator remains unstable. At the lower levels, weights may be further modified using the proposed method of weights smoothing.

3. Sample weights smoothing

We consider the sampling process as a result of three stage procedure (see Pfeffermann and Sverchkov 2009). At the first stage, the finite population, $U = \{y_{j,\tau}: \tau = 0,...,t, j = 1,..., N, z\}$, is generated from some unknown distribution, $f(y_0,..., y_t, Z)$, which is usually called the super-population distribution or model, here z denotes the set of design variables (frame). The sampling weights $w_j, j = 1,..., N$ are defined on this realized final population. Then, the sample, $S = \{y_{j,\tau}, w_j: \tau = 0,..., t, j = 1,..., n\}$ is selected from the finite population. Finally, since some units do not respond, S can be decomposed into monthly sets S_t containing units that respond in month t and $t-1$. Under this model the outcome variables and the sampling weights are random and follow the sample distribution (for exact definitions see Pfeffermann and Sverchkov 2009). Therefore, (2) can be approximated as,

$$\hat{R}_t \doteq R_t + \frac{1}{Y_{t-1}} \sum_{j \in S_t} w_j (y_{j;t} - R_t y_{j;t-1}) \doteq R_t + \frac{1}{Y_{t-1}} \sum_{j \in S_t} E[w_j (y_{j;t} - R_t y_{j;t-1}) \mid j \in S_t]$$

$$= R_t + \frac{1}{Y_{t-1}} \sum_{j \in S_t} E\{E[w_j (y_{j;t} - R_t y_{j;t-1}) \mid y_{j;t} - R_t y_{j;t-1}, j \in S_t] \mid j \in S_t\}$$

$$= R_t + \frac{1}{Y_{t-1}} \sum_{j \in S_t} E[\underbrace{E(w_j \mid y_{j;t} - R_t y_{j;t-1}, j \in S_t)}_{v_j} (y_{j;t} - R_t y_{j;t-1}) \mid j \in S_t]$$

$$\doteq R_t + \frac{1}{Y_{t-1}} \sum_{j \in S_t} v_j (y_{j;t} - R_t y_{j;t-1}),$$

where $v_j = E(w_j \mid y_{j;t} - R_t y_{j;t-1}, j \in S_t)$ are the smoothed weights. Here, in the first line we approximate weighted residuals by their expectations and in the last line we do the opposite.

This implies that the relative growth of employment from the previous to the current month can be estimated also as,

$$\hat{R}_t^S = \frac{\sum_{j\in S_t} \hat{v}_j y_{j;t}}{\sum_{j\in S_t} \hat{v}_j y_{j;t-1}}, \qquad (3)$$

where $\hat{v}_j$ are estimates of $E(w_j \mid y_{j;t} - R_t y_{j;t-1}, j \in S_t)$.

Remark 1. (General justification for smoothing weights). Let s be a sample selected from a final population with inclusion probabilities $\pi_j = w_j^{-1}$. Pfeffermann and Sverchkov (1999) show that for any random variables ξ_j and x_j,

$$f(\xi_j \mid x_j) = \frac{E(w_j \mid \xi_j, x_j, j \in s)}{E(w_j \mid x_j, j \in s)} f(\xi_j \mid x_j, j \in s), \qquad (4)$$

where f is the probability density function when ξ_j is continuous and the probability function when ξ_j is discrete.

Therefore, for estimating relationships between variables ξ_j and x_j on the super-population from the observed sample data, the sampling weights can be replaced by their conditional expectations, $E(w_j \mid \xi_j, x_j, j \in s)$. For example, if one is interested in regression of ξ_j on x_j, then, by (4), $E(\xi_j \mid x_j) = E[\frac{E(w_j \mid \xi_j, x_j, j \in s)}{E(w_j \mid x_j, j \in s)} \xi_j \mid x_j, j \in s]$, the latter implies that one can use any weights w_j^* satisfying $E[\frac{w_j^*}{E(w_j^* \mid x_j, j \in s)} \xi_j \mid x_j, j \in s] = E[\frac{E(w_j \mid \xi_j, x_j, j \in s)}{E(w_j \mid x_j, j \in s)} \xi_j \mid x_j, j \in s]$ in this case.

Example. (Estimating an expectation over population). $E(\xi_j) = E[\frac{E(w_j \mid \xi_j, j \in s)}{E(w_j \mid j \in s)} \xi_j \mid j \in s]$, which suggests two estimators based on the smoothed weights: a) estimating the external expectation and expectation in the denominator by respective sample means $\hat{E}(\xi_j) = \frac{\sum_{j\in s} E(w_j \mid \xi_j, j \in s)\xi_j}{\sum_{j\in s} E(w_j \mid \xi_j, j \in s)}$ (analog of Hajek estimator); b) on the other hand, since for fixed size sampling schemes $E(w_j \mid j \in s) = N/n$, estimating the external expectation by the mean and substituting the later equality one can get $\hat{E}(\xi_j) = \frac{\sum_{j\in s} E(w_j \mid \xi_j, j \in s)\xi_j}{N}$ (analog of Horvitz-Thompson estimator).

The new weights are smoother than the original sampling weights and contain all necessary information on the relationship between the outcome variable, ξ_j, and the sampling weight. For example, if w_j is a deterministic function of the outcome, ξ_j, then the smoothed weight is the same as the original one, $v_j \overset{def}{=} E(w_j \mid \xi_j, j \in s) = w_j$, on the contrary, if the outcome and the sampling weights are unrelated, then the smoothed weight is constant. Therefore, the estimates based on the smoothed weights can be less variable (more efficient) than classical probability weighted estimators. The smoothed weights were used in Pfeffermann and Sverchkov (1999, 2003) in parametric estimation of linear and Generalized Linear Models.

One can find another theoretical justification for using smoothed weights in Beaumont (2008). Beaumont and Rivest (2009) suggested the use of smoothed weights to deal with influential observations.

Remark 2. The previous remark is correct for theoretical smoothed weight, $E(w_j \mid \xi_j, x_j, j \in s)$. In practice the latter expectation has to be estimated. If the estimate will be inaccurate then the final estimator can be biased and/or less efficient.

4. Proposed estimators and their evaluation

We consider the following set of estimators.

1) Unweighted Ratio estimator, $\hat{R}_t^{UNW} = \frac{\sum_{j \in S_t} y_{j;t}}{\sum_{j \in S_t} y_{j;t-1}}$. This estimator can be biased if sampling is informative.

2) Probability Weighted Ratio estimator, $\hat{R}_t^{PWR} = \frac{\sum_{j \in S_t} w_j y_{j;t}}{\sum_{j \in S_t} w_j y_{j;t-1}}$,. This estimator is unbiased over randomization distribution, can be not efficient if original sampling weights, w_j, and residuals, $y_{j;t} - R_t y_{j;t-1}$, are not strongly related, see Remark 1.

3) Robust estimator, $\hat{R}_t^R = \frac{\sum_{j \in S_t} w_j^R y_{j;t}}{\sum_{j \in S_t} w_j^R y_{j;t-1}}$, with weights, w_j^R, obtained by Robust Estimation Procedure described in Section 2. This estimator is protected against influential observations.

4) Robust smoothed estimator, $\hat{R}_t^{RS} = \frac{\sum_{j \in S_t} v_j^R y_{j;t}}{\sum_{j \in S_t} v_j^R y_{j;t-1}}$, where v_j^R are estimated by regressing w_j^R against estimated residuals $y_{j;t} - \hat{R}_t^{PWR} y_{j;t-1}$ by SAS Proc LOESS, http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#loess_toc.htm.

Remark 3. It is difficult to determine the functional form of the relationship between weights and residuals (see Figure 1). Therefore, we consider a nonparametric approach. This approach does not require explicitly formulating a model; the drawback is that the nonparametric estimation, generally, is less efficient than the parametric approach. We use the standard SAS LOESS procedure with default parameters. On the plot, stars represent the values of smoothed weights estimated using this procedure.

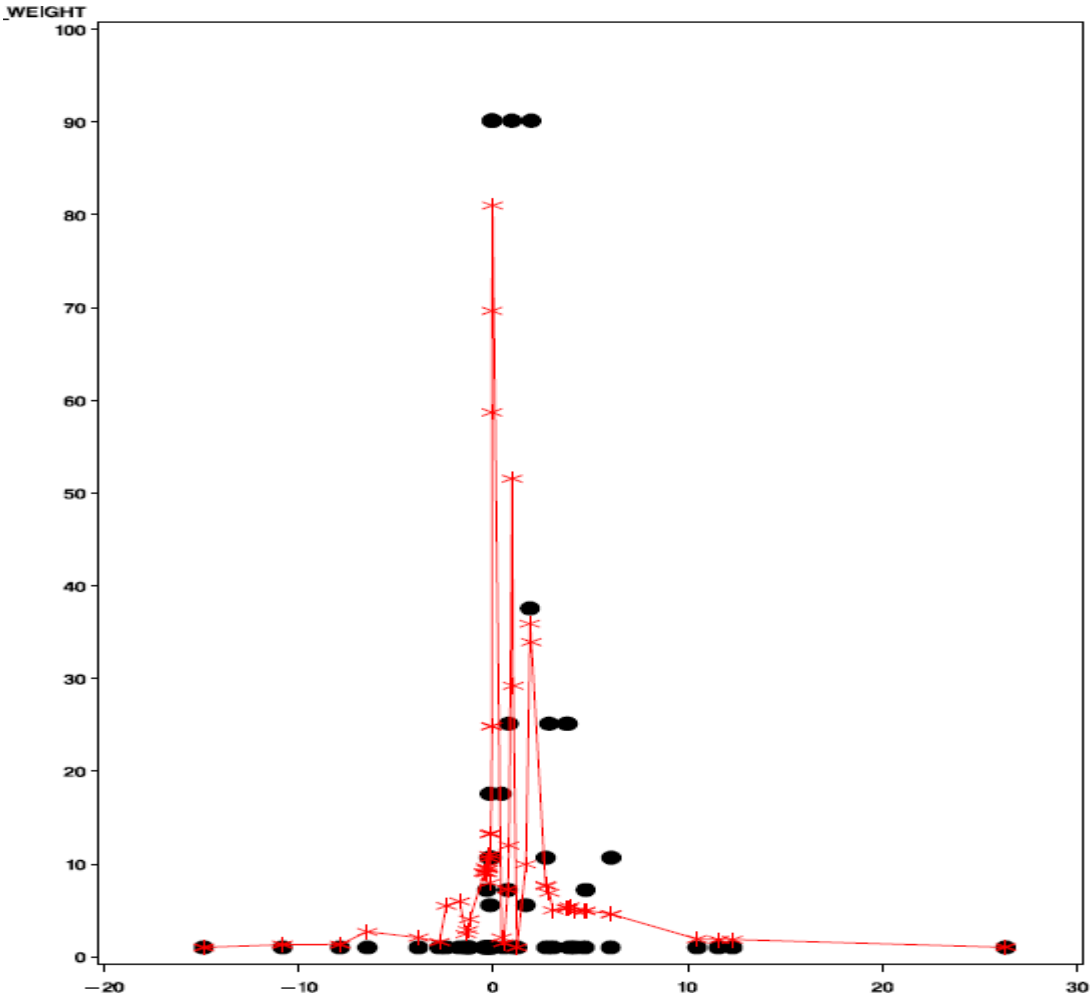

Figure 1: Robust sample weights w_j^R against estimated residuals $y_{j;t} - \hat{R}_t^{PWR} y_{j;t-1}$ (black) and SAS LOESS procedure estimates for regression $E(w_j^R \mid y_{j;t} - R_t y_{j;t-1}, j \in S_t)$ against estimated residuals for a particular MSA (red).

We made estimates of employment at the MSA supersector level. The estimates were constructed using data reported monthly over three years. For each year, we started the estimation cycle from the corresponding September QCEW employment level as the benchmark. Each year's estimation went on for 12 consecutive months using the estimation sequence described in Section 2.2.

The primary goal of this research is to find the estimator that improves the monthly volatility as compared to the currently used estimator. At the same time, revisions after 12 months of estimation (the annual revisions) should be, on average, at least as good as with the current estimator.

We make conclusions about the relative quality of the competing estimators based on the summary of the distances of corresponding estimates from the QCEW figures that serve as the truth. The summaries are derived over the set of MSAs in each supersector level and over the set of MSAs at the Total Private level.

For cell m at month t, the difference between the estimated, $\hat{Y}_{m,t}$, and the true, $Y_{m,t}$, employment levels is

$$d_{m,t} = \hat{Y}_{m,t} - Y_{m,t}$$

The difference, relative to the level (times 100), is

$$rel\_d_{m,t} = 100\left(\hat{Y}_{m,t} - Y_{m,t}\right)/Y_{m,t}.$$

The difference in the monthly changes is

$$c_{m,t} = \left(\hat{Y}_{m,t} - \hat{Y}_{m,t-1}\right) - \left(Y_{m,t} - Y_{m,t-1}\right).$$

The difference in the monthly changes, relative to the level (times 100), is

$$rel\_c_{m,t} = 100 c_{m,t} / Y_{m,t-1}.$$

In this paper, we publish results at the MSA Total Private level. The following summary statistics are presented in Table 1.

Mean revision: $\bar{d}_t = \frac{1}{M}\sum_{m=1}^{M} d_{m,t}$ and $\overline{rel\_d_t} = \frac{1}{M}\sum_{m=1}^{M} rel\_d_{m,t}$

Mean absolute revision: $\bar{d}_t^{\,a} = \frac{1}{M}\sum_{m=1}^{M}\left|d_{m,t}\right|$ and $\overline{rel\_d}_t^{\,a} = \frac{1}{M}\sum_{m=1}^{M}\left|rel\_d_{m,t}\right|$

75th percentile of the absolute revisions ($\left|d_{m,t}\right|$ or $\left|rel\_d_{m,t}\right|$) over the set of M domains: d_t^{75} and $rel\_d_t^{75}$.

The following summary statistics for the monthly changes are presented in Tables 2.

Mean revision: $\bar{c} = \frac{1}{12M}\sum_{t=1}^{12}\sum_{m=1}^{M} c_{m,t}$ and $\overline{rel\_c} = \frac{1}{12M}\sum_{t=1}^{12}\sum_{m=1}^{M} rel\_c_{m,t}$.

Mean absolute revision: $\bar{c}^{\,a} = \frac{1}{12M}\sum_{t=1}^{12}\sum_{m=1}^{M}\left|c_{m,t}\right|$ and $\overline{rel\_c}^{\,a} = \frac{1}{12M}\sum_{t=1}^{12}\sum_{m=1}^{M}\left|rel\_c_{m,t}\right|$

75th percentile of the absolute revisions ($\left|c_{m,t}\right|$ or $\left|rel\_c_{m,t}\right|$) over the set of M domains and 12 months: c^{75} and rel_c^{75}.

We obtained encouraging results: the summary statistics look consistently better for the new estimator (Tables 1 and 2). However, there are examples where the new estimator does not work as expected. Of course, it is not reasonable to expect that one estimator would work better than another in every case. However, we would like to be able to identify and correct certain cases where the new estimator is egregiously wrong (as in Figure 3).

Table 1: Differences from QCEW after 12 months of estimation. Summary over all MSAs at the Total Private Level

Estimator	**N**	$\bar{d}_{t=12}$	$\overline{rel\_d}_{t=12}$	$\bar{d}^{a}_{t=12}$	$d^{75}_{t=12}$	$rel\_d^{a}_{t=12}$	$rel\_d^{75}_{t=12}$
Based on September 2009 benchmark							
LOESS	390	-631	-0.09	2,703	2,787	1.67	2.19
Robust	390	-934	-0.36	3,529	3,602	2.53	3.35
UnwRatio	390	-2,731	-0.94	4,425	4,703	2.55	3.37
WRatio	390	-717	-0.45	4,167	4,823	2.81	3.91
Based on September 2010 benchmark							
LOESS	401	-961	-0.14	2,438	2,762	1.41	1.91
Robust	401	-946	-0.17	3,104	3,456	2.07	2.94
UnwRatio	401	-2,493	-0.83	4,028	4,464	2.32	3.04
WRatio	401	-339	-0.05	3,467	3,451	2.31	2.99
Based on September 2011 benchmark							
LOESS	401	-439	0.08	2,080	2,106	1.27	1.71
Robust	401	-893	-0.17	2,805	2,768	1.89	2.70
UnwRatio	401	-2,237	-0.52	3,644	3,542	2.14	2.51
WRatio	401	-1,100	-0.19	3,122	3,469	2.07	2.96

Table 2: Monthly differences from QCEW for 12 months of estimation. Summary over all MSAs at the Total Private Level and all 12 months

Estimator	**N**	c	$\overline{rel\_c}$	c^{a}	c^{75}	rel_c^{a}	rel_c^{75}
Based on September 2009 benchmark							
LOESS	4680	-53	-0.01	1,104	1,120	0.68	0.89
Robust	4680	-78	-0.03	1,377	1,516	0.96	1.29
UnwRatio	4680	-228	-0.08	1,693	1,695	1.11	1.31
WRatio	4680	-60	-0.04	2,025	1,716	1.83	1.46
Based on September 2010 benchmark							
LOESS	4812	-80	-0.02	985	996	0.62	0.82
Robust	4812	-79	-0.02	1,196	1,333	0.82	1.12
UnwRatio	4812	-208	-0.07	1,534	1,547	0.95	1.21
WRatio	4812	-28	-0.01	1,348	1,459	0.93	1.23
Based on September 2011 benchmark							
LOESS	4812	-37	0	1,002	991	0.60	0.77
Robust	4812	-74	-0.02	1,197	1,302	0.80	1.09
UnwRatio	4812	-186	-0.05	1,417	1,515	0.88	1.14
WRatio	4812	-92	-0.02	1,296	1,475	0.89	1.21

We present two examples of estimation over 12 months at an MSA supersector level (see Figures 2 and 3.) There are three lines on each plot. The black line corresponds to the true employment level at each of the 13 months, including the starting September. The blue line shows the Robust estimator (currently used estimator) and the magenta line is for the new LOESS based estimator. The first example (Figure 2) shows the situation where the new estimator results are smoother than the Robust estimator. Look especially at the change in employment between February and March and notice how the volatility of the Robust estimator was corrected in the new estimator.

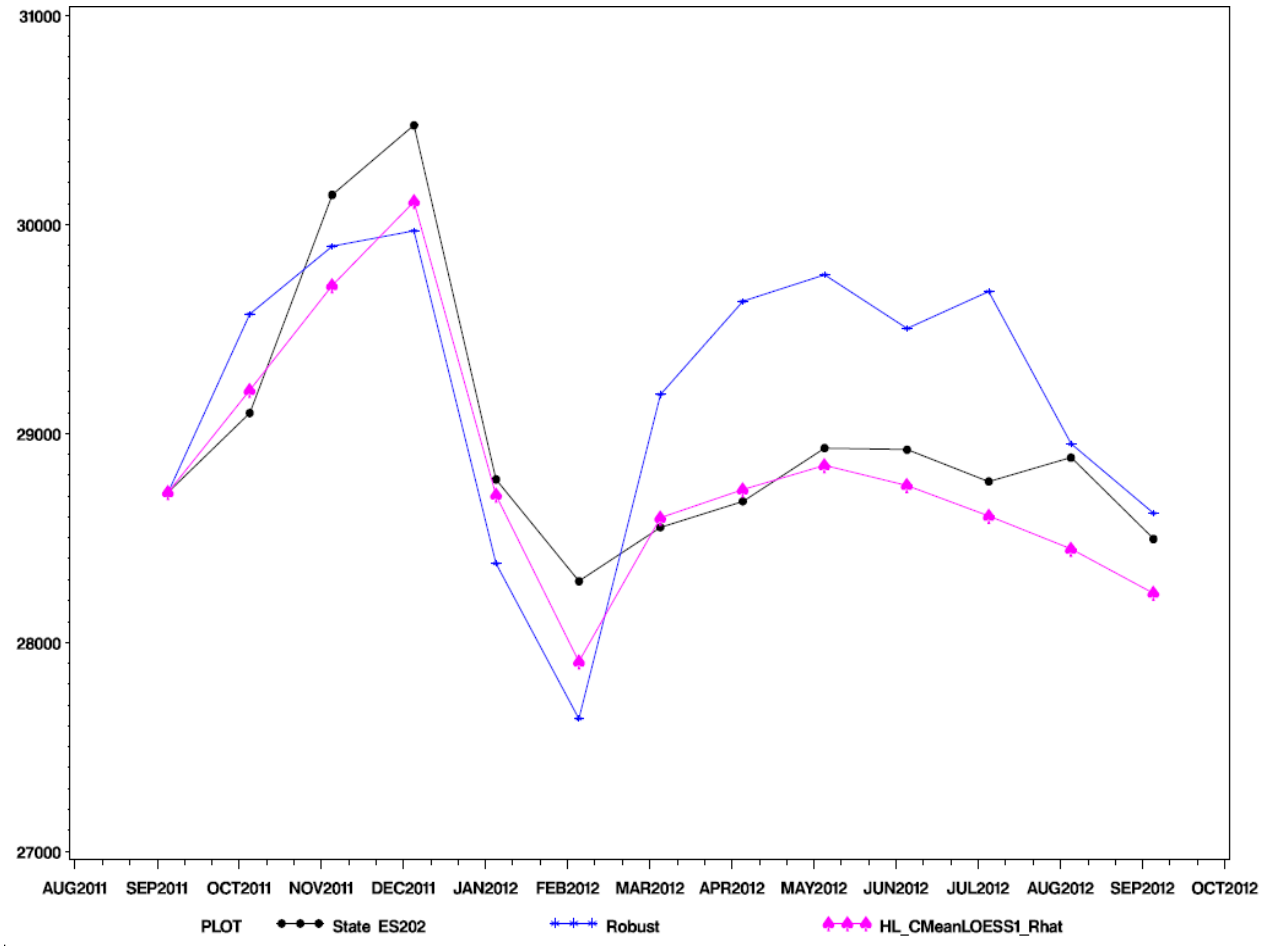

Figure 2: Example of estimation where new estimator works well. Results from two competing estimators (Robust: blue stars; LOESS-based: magenta spides) and the employment levels from QCEW.

The second example (Figure 3) demonstrates that there exist instances where the new method is not working as expected. Notice April to May and May to June changes.

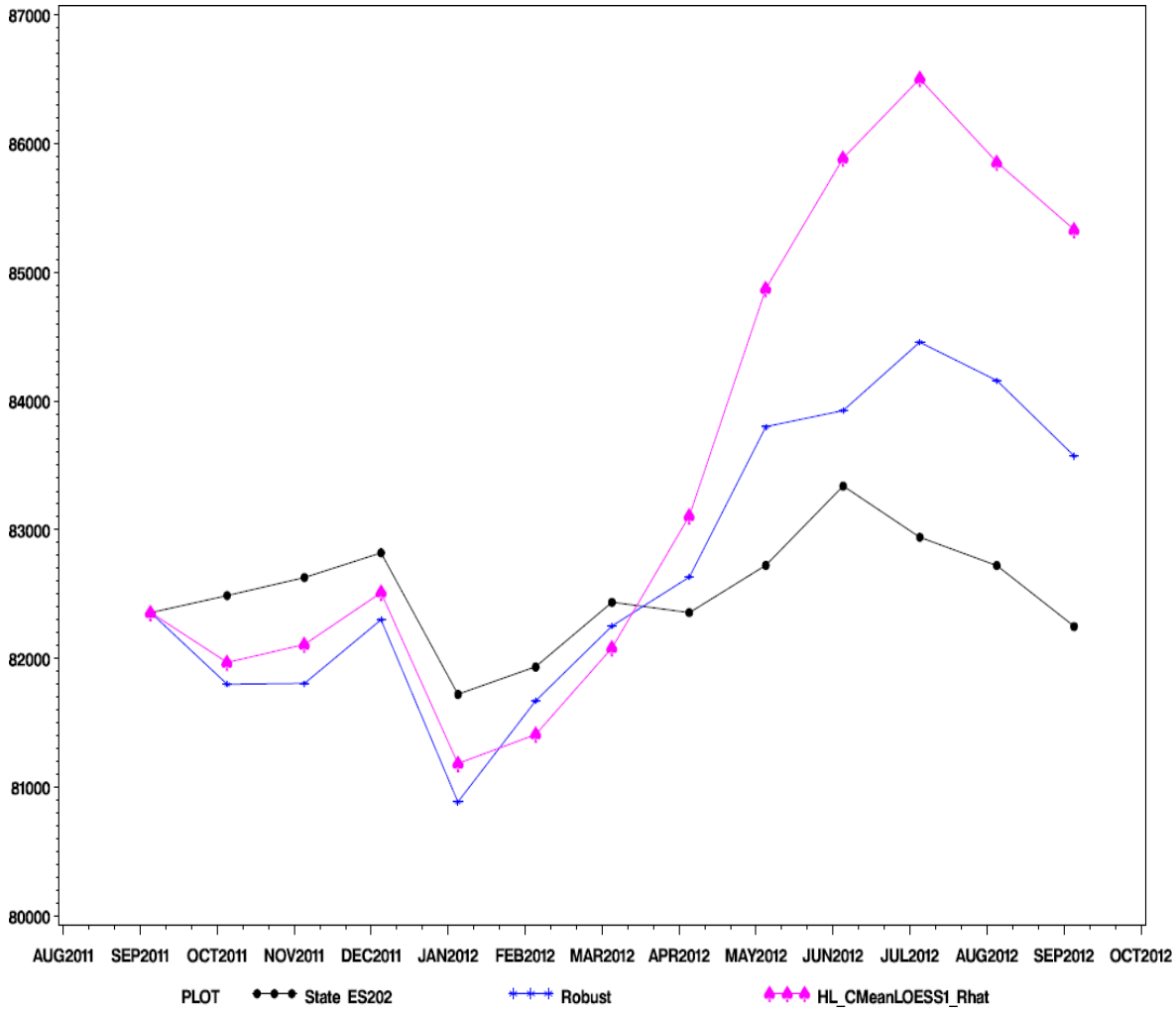

Figure 3: Example of estimation where new estimator is not working well. Results from two competing estimators (Robust: blue stars; LOESS-based: magenta spides) and the employment levels from QCEW.

In Figure 4, we plot weights against the residuals for the problem month (April to May change). Notice that the smoothed weights for the 4 points on the right are obtained by nearly linear interpolation. As a result, we have hugely exaggerated “smooth” weights for two of these points. This tells us that there is room for improvement in the nonparametric method we use.

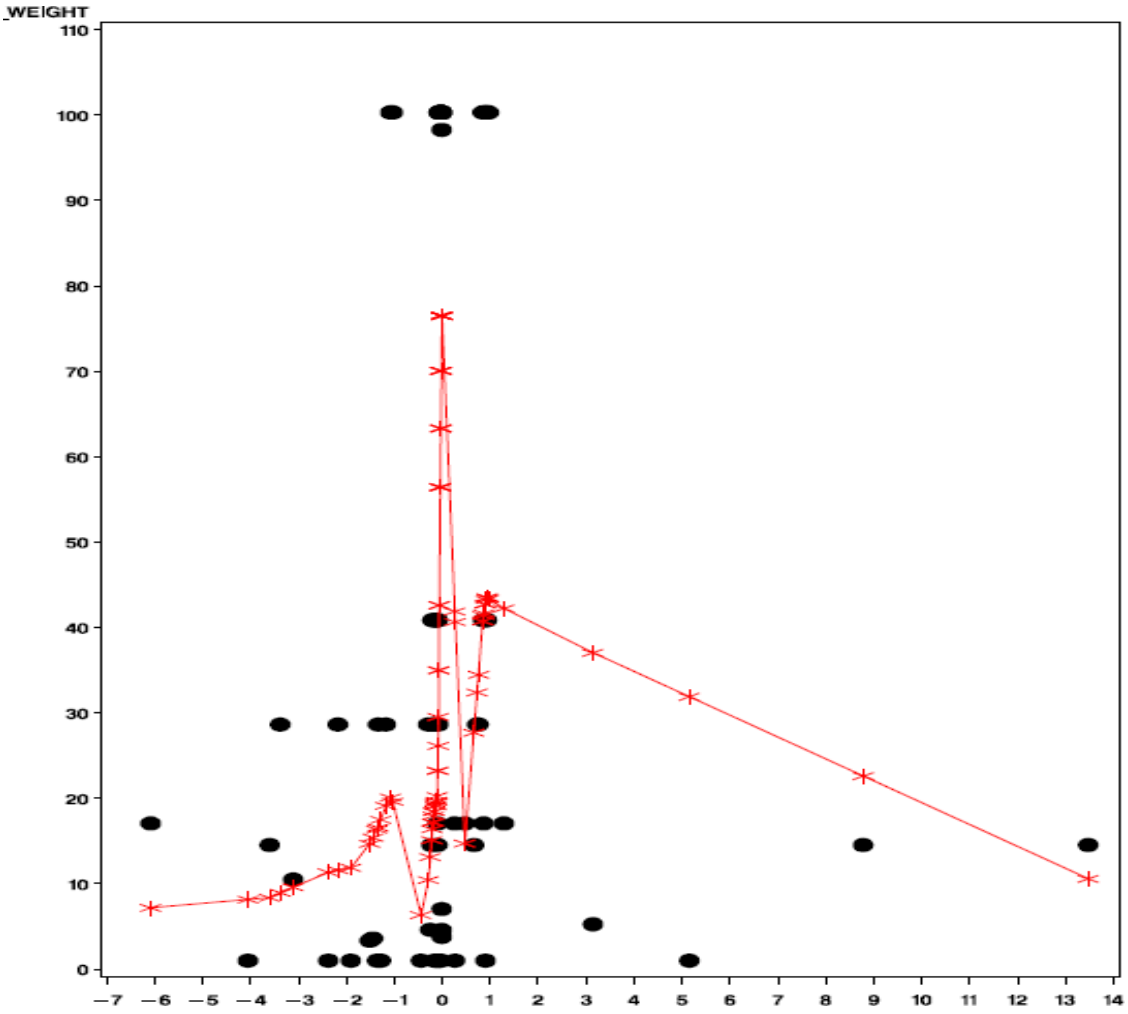

Figure 4: Illustration for the April-May change in Figure 3. Robust sample weights w_j^R against estimated residuals $y_{j;t} - \hat{R}_t^{PWR} y_{j;t-1}$ (black) and SAS LOESS procedure estimates for regression $E(w_j^R \mid y_{j;t} - R_t y_{j;t-1}, j \in S_t)$ against estimated residuals for a particular MSA (red).

Summary:

The overall results are promising but there are certain cases where the new method is not working properly. Some tuning is needed of the nonparametric method we used.

One minor practical inconvenience is that the smooth weights change every month.

References

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, **95**, 3, *pp*. 539–553

Beaumont, J.-F., and Rivest, L.-P (2009). Dealing with outliers in survey data. Chapter 11 in D. Pfeffermann and C. R. Rao (eds.), *Handbook of Statistics. No. 29A, Sample Surveys: Inference and Analysis*, pp. 247-280

Bureau of Labor Statistics (2011). Employment, hours, and earnings from the establishment survey. Chapter 2 of BLS Handbook of Methods, U.S. Department of Labor, http://www.bls.gov/opub/hom/pdf/homch2.pdf

Gershunskaya, J. (2011) Treatment of influential observations in the Current Employment Statistics survey. Unpublished doctoral dissertation, University of Maryland, College Park.

Kokic, P. N., and Bell, P. A. (1994), "Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator," *Journal of Official Statistics*, 10, 419-435.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data, *Sankhya* B, **61**, Pt.1, 166 – 186

Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. Chapter 12 in R. L. Chambers and C. Skinner (eds.), *Analysis of Survey Data*, Chichester: Wiley , pp. 175 – 195

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. Chapter 39 in D. Pfeffermann and C. R. Rao (eds.), *Handbook of Statistics. No. 29B, Sample Surveys: Inference and Analysis*, pp. 455-487