Estimation for Detailed Publication Levels in the Current Employment Statistics Survey[1]

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

Abstract

One prerequisite for successful small domain modeling is the ability to form a "pool" of like "areas", such that "borrowing strength" across the areas is a viable concept. This may not always be possible for various reasons.

We consider estimation of employment from the Current Employment Statistics (CES) survey conducted by the U.S. Bureau of Labor Statistics. Detailed estimation cells are defined as intersections of industrial and geographic levels. Combining direct detailed level estimates across States is not always feasible due to possible substantial differences between States, intricacies of individual States' estimation structures, or simply due to logistics of the production procedures at State levels (for example, differences in the production timeframe.)

A simple area-level model is formulated for higher level estimates. We explore the possibility of applying the parameters from the higher-level setting to States' detailed levels.

Key Words: small area estimation, area-level model

1. Introduction

Estimates of employment from the Current Employment Statistics (CES) survey conducted by the U.S. Bureau of Labor Statistics are produced at various aggregation levels.

In this paper, we consider State and Area estimation at detailed levels defined by intersections of industry and geography. At finer levels, sample sizes are often small and direct sample based estimates are unreliable, providing motivation for application of small area estimation (SAE) methods (see Rao 2003). Here, the term "small area" refers generally to a domain of interest where the sample is scarce or even non-existent. Typically, a SAE model is based on a set of auxiliary variables and on certain assumptions about similarity of groups of areas. The similarity assumption allows borrowing information across areas, thus strengthening estimates in the participating domains.

Therefore, an important task in the model building is to form a suitable "pool" of areas. In CES, this would often stipulate grouping together domains from different States. However, this may be difficult to accomplish for several reasons. Each State has its own hierarchical estimation structure that may not be compatible with other States. Another obstacle is that the States have differences in the timeframe for producing their monthly

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

estimates; this would create a logistical problem in an attempt to combine estimates from different States. Most important, the assumption of similarity may not hold for a given detailed industrial level across the States. For these reasons, forming groups of areas from different States may not always be a viable solution.

For example, the currently used CES small domain model (Eltinge et al 2001) does not borrow information across States. For metropolitan statistical areas (MSA), the model relies on historical data for a given series and on sample based estimated statewide employment trends. This approach generally results in stable estimates. However, the model produces biased estimates if historical trends do not reflect current tendencies in the economy or whenever the MSA level employment trends differ from the statewide trends. Some details of the currently used small domain model are given in Section 2.

In Section 3, we describe a model for a higher level, the statewide industrial estimation supersectors (ESS). Recently, despite the aforementioned obstacles and reservations in combining the States estimates, CES has implemented a simple model for statewide ESS estimates. It must be noted that the sample at the statewide ESS series level is usually large enough and the direct probability based estimates there are considered reliable and suitable for publication. The model based estimates at this level are published only for a very limited pre-determined subset of series that do not have adequate sample. Essentially, the model is used to determine a multiplicative factor that adjusts individual historical predictions based on the current probability based estimates combined from all States. In Section 4, we propose using this adjustment factor, as a known parameter, for States' detailed level historical movements.

Data analysis is presented in Section 5.

2. The current small domain model

Let θ_d denote the true relative change in monthly employment, for a given month, in domain d of a State m and ESS i. A domain $d \in (im)$ may be a detailed statewide industry or an intersection of MSA and industry within a State. Suppose we can obtain p estimates $Y_d^{(1)},...,Y_d^{(p)}$ of θ_d using different sources of information. For example, for a domain defined as an industry in an MSA of a given State, let $Y_d^{(1)}$ be the direct sample based estimate, let $Y_d^{(2)}$ be the ARIMA forecast from the historical series, and let $Y_d^{(3)}$ be an estimate of the relative employment change for the entire State in the same industry. In a certain sense, each of these estimators is assumed to give a reasonable estimate of θ_d.

Let us assume that vector $\mathbf{Y}_d = \left(Y_d^{(1)},...,Y_d^{(p)}\right)^T$ comes from the multivariate normal distribution:

$$\mathbf{Y}_d \sim N_p\left(\boldsymbol{\alpha}_d \theta_d, \boldsymbol{\Sigma}_d\right), \qquad (1)$$

where $\boldsymbol{\alpha}_d = \left(\alpha_d^{(1)},...,\alpha_d^{(p)}\right)^T$ is a vector of coefficients, that are interpreted as multiplicative bias corrections for each of the p estimators of θ_d; $\boldsymbol{\Sigma}_d$ is the variance-

covariance matrix. The values of $\alpha_d^{(j)}$ and matrix $\boldsymbol{\Sigma}_d$ are assumed known; thus, the best estimator of θ_d is found using generalized least squares (GLS) as

$$\hat{\theta}_d^{GLS} = \left(\boldsymbol{\alpha}_d^T \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\alpha}_d\right)^{-1} \boldsymbol{\alpha}_d^T \boldsymbol{\Sigma}_d^{-1} \mathbf{Y}_d . \qquad (2)$$

In practice, coefficients $\alpha_d^{(j)}$ are not known and further assumptions have to be made. Currently, each component of the vector of coefficients is set to 1, i.e., we assume that each $Y_d^{(j)}$ is an unbiased estimator of the truth. Components of matrix $\boldsymbol{\Sigma}_d$ also are not known and have to be estimated from the data. Estimates of variances $V_d^{(j)}$ are supplied with each of the component estimators. For example, a generalized variance function can be used as the variance of the sample based estimators; an estimated variance of the ARIMA prediction can be obtained from standard software. Further, the off-diagonal terms of $\boldsymbol{\Sigma}_d$ are assumed to be zeros (i.e., the estimators are viewed as coming from independent sources). Hence, the resulting weighted least squares (WLS) estimator has the form of a weighted average of p component estimators:

$$\hat{\theta}_d^{WLS} = w_d^{(1)} Y_d^{(1)} + \ldots + w_d^{(p)} Y_d^{(p)}, \qquad (3)$$

with weights $w_d^{(j)} = \dfrac{\left(V_d^{(j)}\right)^{-1}}{\left(V_d^{(1)}\right)^{-1} + \ldots + \left(V_d^{(p)}\right)^{-1}}$.

The appeal of estimator (3) is its simplicity. Several estimators of the same quantity are combined together to form the optimal estimator. This approach differs from the area-level small domain methods based on the mixed model or empirical Bayes in that there is no assumption that true parameters of interest are similar for a group of domains. Thus, there is no need to group together a set of domains. The assumption about "similarity" of domains is replaced by the assumption about existence of a set of estimators of the same truth.

However, the currently used assumption of unbiasedness of each estimator $Y_d^{(j)}$ may be overly strong. Failure of this assumption would lead to problematic results. For example, using the statewide estimate as an unbiased estimator for the MSA level may be misleading. Viewing predictions from historical data as unbiased estimates of the current event is also risky and may lead to significant biases at turning points in the economy.

Unfortunately, the model postulated by (1) does not contain a prescription on how to obtain the true vector $\boldsymbol{\alpha}_d = \left(\alpha_d^{(1)}, \ldots, \alpha_d^{(p)}\right)^T$ of bias adjustments. In the following, we consider the empirical Bayes approach, which makes use of similar information as intended by the current model.

3. The higher level model

We now describe a higher level model. It is a special case of the Fay-Herriot (FH) model (Fay and Herriot 1979). The areas in this case are defined as the statewide ESS levels.

The models are formulated separately for each ESS at a given month. Let θ_{im} denote the true relative change in monthly employment in ESS i in State m, $m=1,...,M$. The auxiliary variable $Y_{im}^{(2)}$ is a relative change in employment as forecasted from the historical data. The **model H1** assumptions are

$$Y_{im}^{(1)} \overset{ind}{\sim} N\left(\theta_{im}, V_{im}^{(1)}\right), \tag{4}$$

$$\theta_{im} \overset{ind}{\sim} N\left(\beta_i Y_{im}^{(2)}, A_i\right), \tag{5}$$

where $Y_{im}^{(1)}$ is the direct sample based estimate of θ_{im}, the coefficient β_i and variance A_i are unknown parameters of the model. Variance $V_{im}^{(1)}$ of $Y_{im}^{(1)}$ is assumed to be known. In practice, a generalized variance function is used to approximate the variances.

It is assumed that $Y_{im}^{(2)}$ is a good predictor of the truth. There is a certain level of belief that the monthly trends have limited tendency to change from one year to another, for the same month of a year. We consider regression through the origin. Coefficient β_i can be viewed as an adjustment factor that "corrects" area specific historical information (represented by $Y_{im}^{(2)}$) based on the current tendency across all areas.

A slight variation of assumption (5) is

$$\theta_{im} \overset{ind}{\sim} N\left(\beta_i Y_{im}^{(2)}, a_i V_{im}^{(2)}\right), \tag{6}$$

where $V_{im}^{(2)}$ is a known factor (e.g., the variance of forecast $Y_{im}^{(2)}$). We refer to assumptions (4) and (6) as **model H2**.

The best linear unbiased predictor (BLUP) has the form

$$\hat{\theta}_{im}^{BLUP} = \hat{\beta}_i Y_{im}^{(2)} + \gamma_{im}\left(Y_{im}^{(1)} - \hat{\beta}_i Y_{im}^{(2)}\right), \tag{7}$$

where $\hat{\beta}_i$ is the best linear unbiased estimator (BLUE) of β_i, $\hat{\beta}_i = \left(\sum_{m=1}^{M} w_{im}\left(Y_{im}^{(2)}\right)^2\right)^{-1} \sum_{m=1}^{M} w_{im} Y_{im}^{(1)} Y_{im}^{(2)}$. Under model H1, $w_{im} = \frac{1}{A_i + V_{im}^{(1)}}$ and $\gamma_{im} = \frac{A_i}{A_i + V_{im}^{(1)}}$; under model H2, $w_{im} = \frac{1}{a_i V_{im}^{(2)} + V_{im}^{(1)}}$ and $\gamma_{im} = \frac{a_i V_{im}^{(2)}}{a_i V_{im}^{(2)} + V_{im}^{(1)}}$.

4. The proposed models for detailed estimation cells

Using the notation of Section 2 and 3, the assumptions for domain d are

$$Y_d^{(1)} \overset{ind}{\sim} N\left(\theta_d, V_d^{(1)}\right), \tag{8}$$

$$\theta_d \overset{ind}{\sim} N\left(\beta_d Y_d^{(2)}, A_d\right), \qquad (9)$$

where $d \in (im)$, $i = 1,...,I$, $m = 1,...,M$.

As with the current model, variance $V_d^{(1)}$ of direct sample estimate $Y_d^{(1)}$ is assumed to be known.

Parameters β_d and A_d are unknown and have to be estimated from the data. To make estimation possible, we are bound to make further assumptions. In what follows, we assume that $\beta_d = \beta_m \beta_i$ for domains $d \in (im)$; β_m and β_i are factors applied for sublevels of State m and ESS i, respectively. The estimates of β_i are obtained from the H1 or H2 model for a given ESS i; they are plugged into (9) and henceforth are viewed as known parameters.

Assume $A_d = A_m$, common for all domains in State m. Thus, we fit a separate model for each State m, where condition (9) becomes

$$\theta_d \overset{ind}{\sim} N\left(\beta_m Y_d^{(*)}, A_m\right), \qquad (10)$$

with $Y_d^{(*)} = \beta_i Y_d^{(2)}$ playing the role of auxiliary variable; β_m and A_m are unknown parameters. We refer to conditions (8) and (10) as **model M1**.

Alternatively, assume A_d to be proportional to variance $V_d^{(2)}$. Condition (9) becomes

$$\theta_d \overset{ind}{\sim} N\left(\beta_m Y_d^{(*)}, a_m V_d^{(2)}\right), \qquad (11)$$

with unknown parameters β_m and a_m. Conditions (8) and (11) are referred to as **model M2**.

The empirical best linear unbiased predictor (EBLUP) of θ_d has the form

$$\hat{\theta}_d^{EBLUP} = \hat{\beta}_m \beta_i Y_d^{(2)} + \gamma_d \left(Y_d^{(1)} - \hat{\beta}_m \beta_i Y_d^{(2)}\right), \qquad (12)$$

where, for model M1, $\gamma_d = \dfrac{\hat{A}_m}{\hat{A}_m + V_d^{(1)}}$ and $\hat{A}_m$ is an estimate of A_m; for M2, $\gamma_d = \dfrac{\hat{a}_m V_d^{(2)}}{\hat{a}_m V_d^{(2)} + V_d^{(1)}}$ and $\hat{a}_m$ is an estimate of a_m.

The next two models make use of parameters estimated from the higher level model (H1 or H2) without any alteration.

Assume that for the detailed level $d \in (im)$, the following **model L1** holds for transformed variables:

$$\tilde{Y}_d^{(1)} \sim N\left(\tilde{\theta}_d, V_{im}^{(1)}\right), \tag{13}$$

$$\tilde{\theta}_d \sim N\left(\beta_i \tilde{Y}_d^{(2)}, A_i\right), \tag{14}$$

where $\tilde{Y}_d^{(1)} = Y_d^{(1)} \sqrt{\left(V_d^{(1)}\right)^{-1} V_{im}^{(1)}}$, $\tilde{Y}_d^{(2)} = Y_d^{(2)} \sqrt{\left(V_d^{(1)}\right)^{-1} V_{im}^{(1)}}$; $\tilde{\theta}_d$ is an unknown population parameter and $\tilde{\theta}_d = \theta_d \sqrt{\left(V_d^{(1)}\right)^{-1} V_{im}^{(1)}}$.

Assume the parameters β_i and A_i are known and are the same as in the H1 model. BLUP for $\tilde{\theta}_d$ based on model (13)-(14) is

$$\tilde{\theta}_d^{BLUP} = \beta_i \tilde{Y}_d^{(2)} + \gamma_d^{(L1)} \left(\tilde{Y}_d^{(1)} - \beta_i \tilde{Y}_d^{(2)}\right). \tag{15}$$

Going back to the original scale, the BLUP for the population parameter θ_d is

$$\hat{\theta}_d^{(L1)} = \beta_i Y_d^{(2)} + \gamma_d^{(L1)} \left(Y_d^{(1)} - \beta_i Y_d^{(2)}\right), \tag{16}$$

with $\gamma_d^{(L1)} = \dfrac{A_i}{A_i + V_{im}^{(1)}}$. Notice that $\gamma_d^{(L1)}$ is the same as γ_{im} from the higher level model H1. This makes model L1 especially easy to apply, as it does not require any additional calculation. On the other hand, the fact that, in the weighted average (16), any domain inside a State in a given ESS would receive the same weight γ_{im} for its direct estimator, regardless of the size of the domain, is somewhat disquieting.

Finally, **model L2** makes use of parameters obtained from model H2 by assuming that, in (9), $\beta_d = \beta_i$ and $A_d = a_i V_d^{(2)}$, where estimates of β_i and a_i are obtained from fitting model H2 and are viewed as the known parameters for model L2; BLUP from model L2 has the form

$$\hat{\theta}_d^{(L2)} = \beta_i Y_d^{(2)} + \gamma_d^{(L2)} \left(Y_d^{(1)} - \beta_i Y_d^{(2)}\right), \tag{17}$$

with $\gamma_d^{(L2)} = \dfrac{a_i V_d^{(2)}}{a_i V_d^{(2)} + V_d^{(1)}}$.

5. Data Analysis

We use data for a set of MSAs where, based on the available sample size, it has been determined that direct sample based estimation is only feasible at highly aggregated industrial levels, Goods Producing or Private Service Providing industries. These high level industries cannot be broken down for direct sample based estimation at detailed levels. A model is required for nearly all standard ESS levels for these MSAs.

We present results based on 2008, 2009, and 2010 benchmark years, with September levels used as starting points for estimation; estimates for consecutive months are derived

by multiplying previous month's estimated level by the estimate of the current month relative employment change. For the $Y_d^{(2)}$ component in the L1, L2, M1, and M2 models, we used a "typical" movement of a series at a given month of the year. We used an average over several previous years as the definition of "typical". For the composite estimator (the current method), we used predictions from the time series as $Y_d^{(2)}$.

Summary statistics for the MSA ESS estimates at the 12th month after the corresponding benchmark are presented in Tables 1-3. For each series s, estimates E_s are compared to the true population values T_s available on a lagged basis (6 to 9 months after the publication of the CES estimates) from the Quarterly Census of Employment and Wages program. The statistics presented in the tables are

$$\text{Mean Revision} = \frac{1}{S}\sum_{s=1}^{S} R_s, \qquad \text{Mean Rel Revision} = \frac{1}{S}\sum_{s=1}^{S} relR_s,$$

$$\text{Mean Abs Revision} = \frac{1}{S}\sum_{s=1}^{S} |R_s|, \qquad \text{Mean Abs Rel Revision} = \frac{1}{S}\sum_{s=1}^{S} |relR_s|,$$

75th percentile of $|R_s|$, and 75th percentile of $|relR_s|$,

where $R_s = E_s - T_s$ is revision and $relR_s = 100\dfrac{E_s - T_s}{T_s}$ is percent relative revision at 12th month after the benchmark month.

The direct estimator has the largest mean absolute revision in all three years. The composite estimator based on the model described in Section 2 is an improvement over the direct estimator, although it is susceptible to a bias. Notice that the bias is positive in the 2008 benchmark year (ending in September 2009) and negative in the other two years. This reflects inflexibility of the time series forecasts based on historical data. The forecasts cannot show contemporaneous unexpected changes in the economy, yet they are used in the composite estimator without bias adjustment.

The proposed models performed well on the series considered in our data analysis. The proposed estimators have lower revisions than the direct or composite estimators. Results for L1 and L2 estimators are very close and slightly better than for M1 and M2.

6. Summary and future research

Based on the results of the data analysis, L1 and L2 estimators perform well.

Although more testing is due, the models have the potential to be an improvement over the currently used method.

The current proposal treats each month independently. In the future, it may be beneficial to consider using information across time (as well as cross-sectional). This can be achieved by imposing additional structure on the model parameters (such as an assumption that an area random effects are correlated over time) and by exploiting possible monthly correlations of the sampling errors.

Table 1: The 2008 benchmark year results, 1325 MSA ESS series

	Mean Revision	*Mean Rel Revision, %*	*Mean Abs Revision*	*Mean Abs Rel Revision, %*	*75th pct of Abs Revision*	*75th pct of Abs Rel Revision, %*
L1	136	1.97	325	5.46	403	7.35
L2	189	2.51	367	5.93	442	7.76
M1	188	2.45	386	6.01	475	7.92
M2	168	2.22	383	6.09	502	7.98
Direct	134	2.65	530	9.21	703	11.45
Composite	212	2.48	459	7.53	565	8.97

Table 2: The 2009 benchmark year results, 1359 MSA ESS series

	Mean Revision	*Mean Rel Revision, %*	*Mean Abs Revision*	*Mean Abs Rel Revision, %*	*75th pct of Abs Revision*	*75th pct of Abs Rel Revision, %*
L1	2	0.55	290	4.65	363	5.94
L2	-12	0.39	281	4.58	354	5.83
M1	-6	0.37	295	4.70	378	6.19
M2	-11	0.26	295	4.75	373	6.19
Direct	-14	-0.17	463	7.71	603	9.76
Composite	-264	-1.87	417	7.38	471	8.04

Table 3: The 2010 benchmark year results, 1385 MSA ESS series

	Mean Revision	*Mean Rel Revision, %*	*Mean Abs Revision*	*Mean Abs Rel Revision, %*	*75th pct of Abs Revision*	*75th pct of Abs Rel Revision, %*
L1	22	0.66	282	4.19	346	5.70
L2	24	0.63	285	4.27	359	5.88
M1	22	0.55	312	4.48	388	6.10
M2	12	0.44	313	4.60	393	6.27
Direct	13	0.39	500	8.19	656	10.12
Composite	-101	-1.77	390	6.07	484	7.22

References

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. http://www.bls.gov/opub/hom/pdf/homch2.pdf

Eltinge, J.L., Fields, R.C., Gershunskaya, J., Getz, P., Huff, L., Tiller, R., and Waddington, D. (2001). Small Domain Estimation in the Current Employment Statistics Program *Unpublished Background Material for the FESAC Session on Small Domain Estimation at the Bureau of Labor Statistics.*

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

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