A Time Series approach to Small Area Estimation with Benchmarking

Richard Tiller
U.S. Bureau of Labor Statistics, Washington, DC, USA – tiller_r@bls.gov

Abstract

Model-based small area estimation depends on models that can be susceptible to misspecification arising from rapidly changing economic conditions. When periodic survey estimates are available, time series modelling with benchmark constraints provides an effective approach for making small area estimation more robust. This paper describes a state-space model for hierarchical grouping of areas, where the model for each area consists of a component which explicitly accounts for the survey design and a component that accounts for the behaviour of the population. The individual area models are combined into a multivariate state-space model and a benchmark constraint is added that forces the model-based predictors to agree with a reliable design-based estimator for an aggregate of the areas. As an empirical illustration, we use the U.S. Bureau of Labor Statistics (BLS) monthly unemployment estimates for 9 census divisions (CDs) and the 50 States and the District of Columbia, which are grouped within these divisions. We also compare the time series approach to benchmarking with cross-sectional approaches and demonstrate that the time series approach has special advantages.

Keywords: Benchmarking, State Space Models, Fay-Herriot estimator; correlated sampling error.

1. Introduction

The U.S. Bureau of Labor Statistics (BLS) uses state-space models for the production of all the monthly employment and unemployment estimates in the 50 States and the District of Columbia (Pfeffermann and Tiller, 2006). These models are fitted to the direct survey estimates obtained from the Current Population Survey (CPS). The model for each series combines an extended version of the Basic Structural Model (BSM, Harvey 1989) for the true series, with an AR (15) model for the sampling errors. A high order AR model is necessary to reflect strong autocorrelations in the sampling errors due to the use of a complicated rotating sampling scheme in the CPS. The direct survey estimates are the sums of these two unknown components. In order to protect against possible model breakdowns and to satisfy arithmetic consistency in publication, the separate State estimates are benchmarked to the corresponding direct CPS estimates for the entire nation in a two stage process discussed below.

We use a special type of benchmarking referred to as "internal" which guarantees consistency and robustness of the model estimates in real time.

Let,

$Y_{d,t}$ = true value in area d at time t, d = 1, 2,...,D
$y_{d,t}$ = direct survey estimate
$\hat{Y}_{d,t}$ = estimate obtained under a model for y_{dt}.

Benchmarking modifies model based estimates, $\hat{Y}_{dt}$, to satisfy the following constraint,

$$\sum_{d=1}^{D}\hat{Y}_{d,t}^{bmk} = B_t = \sum_{d=1}^{D} y_{d,t} \cong \sum_{d=1}^{D} Y_{d,t}$$

where B_t is sufficiently close to the aggregate true value. This is referred to as "internal" benchmarking because B_t is just a weighted sum of the area survey estimates. Since it provides no additional information it is suboptimal with respect to the model and therefore increases model variances. This contrasts with "external" benchmarks which are independent of the series being benchmarked and available only after a substantial time lag.

2. First Stage Benchmarking

When survey data are structured hierarchically, it may be desirable to benchmark the model-dependent estimators at each level of the hierarchy. For example, it may be computationally too costly or operationally inflexible to benchmark all of the area models in a single stage. Moreover, to the extent that the survey data have a natural hierarchy, the benchmarking process may be tailored more closely to area characteristics. We begin with first stage benchmarking.

Suppose there are D areas with direct sample estimates that sum to a reliable (low CV) national total. Let $y_t = (y_{1,t},\ldots,y_{D,t})'$ be a vector of direct survey estimates, $Y_t = (Y_{1,t},\ldots,Y_{D,t})'$ a vector of true values, where each follows an independent BSM, and $e_t = (e_{1,t},\ldots,e_{D,t})'$ a vector of survey sampling errors.

These series are represented by a multivariate linear state-space model,

$$y_t = Y_t + e_t,\ Y_t = Z_t\alpha_t,\ E(e_t)=0,\ E(e_\tau e_t{}') = \Sigma_{\tau t} \tag{1}$$

$$\alpha_t = T\alpha_{t-1} + \eta_t;\ E(\eta_t)=0,\ \ E(\eta_t\eta_t{}') = Q,\ E(\eta_t\eta_{t-k}{}')=0,\ k>0,\ E(\eta_t e_\tau{}')=0, \forall\ t,\tau \tag{2}$$

where Z_t and $\Sigma_{\tau t}$ are known matrices. In the first stage, model-based predictors, $\hat{Y}_{d,t}^{bmk}$ of the true values, Y_t, are forced to add up to the aggregate direct survey estimate,

$$\sum_{d=1}^{D}\hat{Y}_{d,t}^{bmk} = \sum_{d=1}^{D} y_{d,t} = \sum_{d=1}^{D} Y_{d,t} + \sum_{d=1}^{D} e_{d,t}. \tag{3}$$

To satisfy (3) we augment (1) as follows,

$$\tilde{y}_t = \tilde{Z}_t\alpha_t + \tilde{e}_t; \qquad \alpha_t = \left(\alpha'_{1,t},\cdots\alpha'_{1,D}\right) \tag{4}$$

$$\tilde{y}_t = \left(y'_{t,} \sum_{d=1}^{D} y_{d,t}\right)',\ \tilde{Z}_t = \begin{bmatrix} Z_t \\ Z_{1t},\ldots,Z_{Dt}\end{bmatrix},\ \tilde{e}_t = \left(e'_{t,} \sum_{d=1}^{D} e_{d,t}\right)'.$$

The autocovariance matrix of the augmented survey error vector, $\tilde{e}_t$ is given by,

$$\tilde{\Sigma}_{\tau,t} = E\left(\tilde{e}_\tau\tilde{e}'_t\right) = \begin{bmatrix}\Sigma_{\tau,t} & h_{\tau,t} \\ h'_{\tau,t} & v_{\tau,t}\end{bmatrix};\ \Sigma_{\tau,t} = E\left(e_\tau e'_t\right),\ v_{\tau,t} = Cov\left(\sum_{d=1}^{D} e_{d,\tau}, \sum_{d=1}^{D} e_{d,t}\right),$$
$$h_{\tau,t} = Cov\left(e_\tau, \sum_{d=1}^{D} e_{d,t}\right). \tag{5}$$

Predictions are produced from a random coefficients model,

$$\begin{pmatrix} T\hat{\alpha}_{t-1}^{bmk} \\ \tilde{y}_t \end{pmatrix} = \begin{pmatrix} I \\ \tilde{Z}_t \end{pmatrix}\alpha_t + \begin{pmatrix} u_{t|t-1} \\ \tilde{e}_t \end{pmatrix} \tag{6}$$

where $T\hat{\alpha}_{t-1}^{bmk}$ is the predictor of α_t with prediction error, $u_{t|t-1} = T\hat{\alpha}_{t-1}^{bmk} - \alpha_t$ and covariance matrix $E\left(u_{t/t-1}u'_{t/t-1}\right) = P_{t/t-1}^{bmk}$.

The covariance matrix of the regression errors in (6) is given by,

$$V_t = E\left[\begin{pmatrix} u_{t/t-1} \\ \tilde{e}_t \end{pmatrix}(u'_{t/t-1}, \tilde{e}_t')\right] = \begin{bmatrix} P_{t/t-1}^{bmk} & C_t^{bmk} \\ C_t^{bmk\prime} & \tilde{\Sigma}_{t,t} \end{bmatrix},\ C_t^{bmk} = E\left[\left(T\hat{\alpha}_{t/t-1}^{bmk} - \alpha_t\right), \tilde{e}_t'\right]. \quad (7)$$

The presence of correlated survey errors results in contemporaneous covariances, C_t^{bmk}, between prediction errors and survey errors which invalidates the use of the classic Kalman filter. To account for these sampling error covariances, we develop a GLS filter (Pfeffermann and Tiller, 2006) described below.

For the benchmark constraint to be binding, it is necessary to initially set the aggregate benchmark sampling error to zero in (4). This is done by setting the variance of the benchmarked error and all its covariances with the sampling errors for the first stage areas to zero.

$$V_{t,0} = \begin{bmatrix} P_{t/t-1}^{bmk} & C_{t,0}^{bmk} \\ C_{t,0}^{bmk\prime} & \tilde{\Sigma}_{t,0} \end{bmatrix} \quad (8)$$

where,

$$\tilde{e}_{t,0} = (e_t', 0)', \tilde{\Sigma}_{t,0} = E\left(\tilde{e}_{t,0}\tilde{e}_{t,0}'\right) = \begin{bmatrix} \Sigma_{t,t} & 0 \\ 0 & 0 \end{bmatrix},\ C_{t,0}^{bmk} = E\left[\left(T\hat{\alpha}_{t/t-1} - \alpha_t\right), \tilde{e}_{t,0}'\right]$$

Setting the benchmark errors to zero is a device for satisfying the constraints but it is not correct for computing variances. Later we make the appropriate adjustments to compute the correct variances.

Standard GLS yields the recursive predictors,

$$\tilde{\alpha}_t^{bmk} = \left[(I, \tilde{Z}_t')[\tilde{V}_{t,0}]^{-1}\begin{pmatrix} I \\ \tilde{Z}_t \end{pmatrix}\right]^{-1}(I, \tilde{Z}_t')[\tilde{V}_{t,0}]^{-1}\begin{pmatrix} \tilde{T}\tilde{\alpha}_{t-1}^{bmk} \\ \tilde{y}_t \end{pmatrix},\ \hat{Y}_{d,t}^{bmk} = \tilde{Z}_t\tilde{\alpha}_t^{bmk} \quad (9)$$

which can be computed more efficiently without having to invert the V_t matrix using an algorithm similar to the Kalman Filter (KF).

$$\begin{aligned} &\hat{Y}_t^{bmk} = \tilde{Z}_t\hat{\alpha}_t^{bmk} = \tilde{Z}_t\left[T\hat{\alpha}_{t-1}^{bmk} + K_t\hat{d}_t^{bmk}\right];\ \hat{d}_t^{bmk} = \tilde{y}_t - \tilde{Z}_t T\hat{\alpha}_{t-1}^{bmk} \\ &K_t = \left(P_{t/t-1}^{bmk}\tilde{Z}_t' - C_{t,0}^{bmk}\right)Var^{-1}\left(\hat{d}_t^{bmk}\right), Var\left(\hat{d}_t^{bmk}\right) = \tilde{Z}_t P_{t/t-1}^{bmk}\tilde{Z}_t' - \tilde{Z}_t C_{t,0}^{bmk} - C_{t,0}^{bmk\prime}\tilde{Z}_t' + \tilde{\Sigma}_{t,t,0} \end{aligned} \quad (10)$$

Notice that if $C_{t,0}^{bmk}$ in (10) contains all zeroes the GLS filter is identical with the KF.

Now we have to correct the V-C matrix for the benchmarked predictor which ignored the benchmarking errors when imposing the constraint.

$$\begin{aligned} &P_t^{bmk} = E[(\hat{\alpha}_t^{bmk} - \alpha_t)(\hat{\alpha}_t^{bmk} - \alpha_t)'] = G_t P_{t/t-1}^{bmk} G_t' + K_t\tilde{\Sigma}_{t,t}K_t' + G_t C_t^{bmk} K_t' + K_t C_t'^{bmk} G_t' \\ &G_t = I - K_t\tilde{Z} \end{aligned} \quad (12)$$

The true variances and covariances $\tilde{\Sigma}_{\tau,t} = E\left(\tilde{e}_\tau\tilde{e}_t'\right)$, $C_t^{bmk} = Cov\left[T\hat{\alpha}_{t-1}^{bmk} - \alpha_t, \tilde{e}_t\right]$ are reflected in (12).

The first stage Benchmarked Predictors, $\hat{Y}_t^{bmk} = \tilde{Z}_t \hat{\alpha}_t^{bmk}$ are unbiased under correct model specification. They are also design consistent in the sense that as sample size in area d at time t increases, the benchmarked predictor for that area and time converges to the true value even under the wrong model specification. This property follows from the property that the sampling error for area d goes to zero as the sample size increases.

3. Second Stage Benchmarking

In two stage benchmarking, each first stage area is subdivided into second stage sub-areas. The D model- dependent first stage estimates are benchmarked to the reliable national aggregate direct survey estimate. Within each first stage area, the second stage model-based estimates are benchmarked to the first stage benchmarked estimates. Let $\hat{Y}_{s,d,t}^{bmk}$ be the benchmarked model estimate for the s^{th} sub-area in the d^{th} area. As in the first stage, each sub-area is modeled as a BSM plus sampling error component and combined into a multivariate state-space model with a benchmark constraint. The 2nd stage benchmark constraint is given by,

$$\sum_s \hat{Y}_{s,d,t}^{bmk} = Y_{d,t}^{bmk}\ ,\ s = 1,\ldots,s_d \tag{13}$$

Notice that the 2nd stage benchmarking is a model-based estimate, rather than a direct survey estimator, and is correlated with the model errors and the sampling errors in a complicated way. For more details see Pfeffermann, Sikov, and Tiller (2014).

4. Empirical Example

We use as an example State unemployment for the U.S. over the period of 2008-12 which reflects the effect of the Great Recession (December 2007 to June 2009) on the labor market. The 50 States are grouped into 9 Census Divisions. The CPS state unemployment estimates are aggregated by Census Division where each of the 9 Divisions are modeled and constrained to sum to the national CPS values (CV of about 2%). States within each Division are modeled and constrained to sum to their respective benchmarked Division model estimates. This two stage process guarantees that all State benchmarked model estimates will sum to the national CPS aggregates.

Figure 1compares the direct aggregate national CPS estimators with the sum of the not benchmarked Division models. From March 2008 to July 2009 the models were systematically underestimating the rapidly rising unemployment levels. Figure 2 plots the model-CPS differences which in a number of consecutive months exceeded twice the standard error of the CPS estimator. The first stage benchmarking of the Census model estimates eliminate these differences.

In the second stage the State models are benchmarked to their Division benchmarked models. Figure 3 shows for the State of Illinois the direct CPS estimators, the not benchmarked model predictors, and benchmarked predictors. Benchmarking corrects in real time the under estimation of the not benchmarked predictors, with the largest corrections occurring in 2008-09. Figure 4 highlights the bias reduction by plotting the differences between the not benchmarked and benchmarked model predictors.

5. Cross-Sectional Benchmarking

Another approach to benchmarking, used extensively in the small area literature, is based on cross-sectional modeling. This approach models differences among areas independently for each time period and then benchmarks each area to an aggregate of all area survey estimates, again independently for each time period. Pfeffermann, Sikov, and Tiller (2014) demonstrated with a simulation that the time series approach has a number of important advantages when compared with an equivalent Fay-Herriot model. With a sufficiently long time series (50 or more time series observations per area), benchmarking of TS estimators yields more accurate predictors than benchmarking of the Fay-Herriot estimators.

In the case of a model breakdown, cross-sectional benchmarking only corrects the bias in some of the areas, but not in all of them. While it may reduce the average MSE across all the areas, for some areas it will increase the MSE. In contrast, cross-sectional benchmarking of time series estimators will correct a bias induced by a model breakdown in every area, if the breakdown is similar in all areas. As shown in our empirical example, this is likely to occur when there are nationwide shocks to the economy.

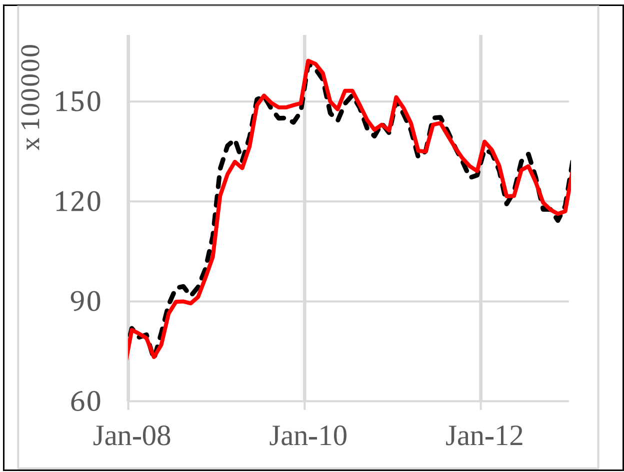

Figure 1. U.S. CPS Unemployment and Sum of Not Benchmarked Division Models
----- U.S. CPS;—— Division Models

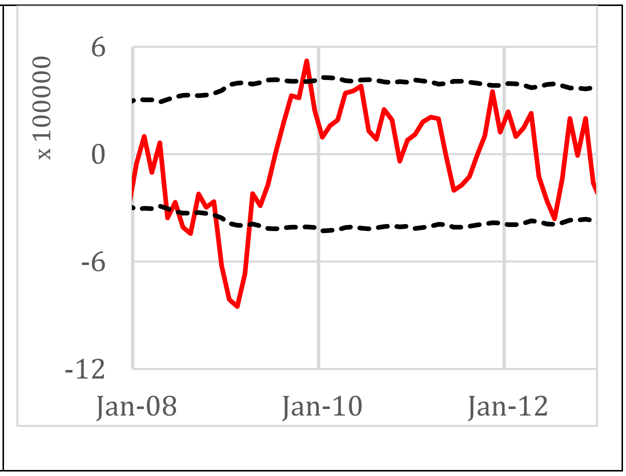

Figure 2. Difference between Sum of Not Benchmarked Division Models and U.S. CPS
—— Difference;----- ±2×SE(CPS)

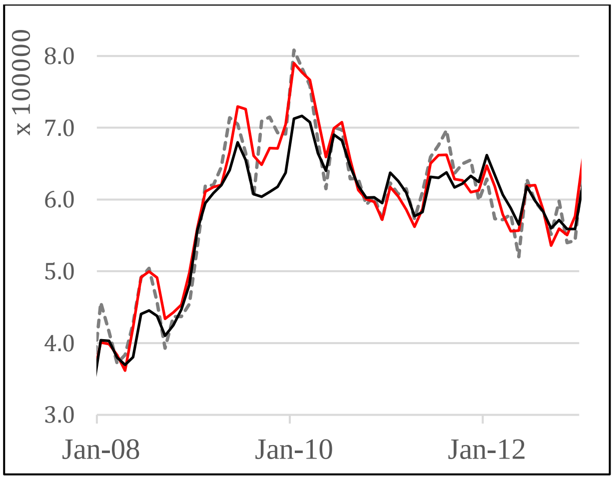

Figure 3. Illinois Unemployment
--- CPS;— BMK Model; — Not BMK Model

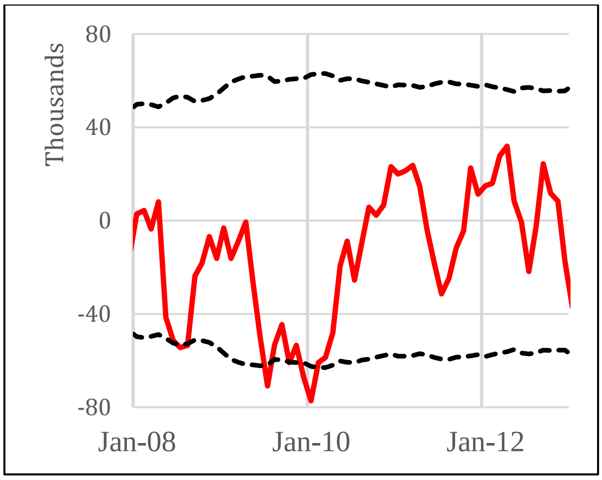

Figure 2. Difference between Not BMK and BMK Illinois Model
— Difference; --- ±2×SE(BMK)

5. Conclusions

A state-space model for hierarchical grouping of areas was developed, where the model for each area combines a model of the sampling error with a model of the population. The individual area models were combined together with a multivariate state-space model and an internal benchmark constraint added that forces the model-based predictors to agree with a reliable design-based estimator for an aggregate of the areas. This approach was illustrated with the U.S. Bureau of Labor Statistics (BLS) monthly unemployment estimates for 9 census divisions (CDs) and the 50 States and the District of Columbia. By reducing bias during the Great Recession this internal benchmarking improved the robustness of the real time unemployment estimates. Unlike cross-sectional benchmarking, time series benchmarking may correct a bias induced by model breakdown in every area, if the breakdown is similar in all areas.

References

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press. Cambridge.

Pfeffermann, D. and Tiller, R. (2006), "Small-area estimation with state–space models subject to benchmark constraints," *Journal of the American Statistical Association*, 101, 1387-1397

Pfeffermann, D., Sikov, A., and Tiller, R, (2014), "Single and Two-Stage Cross-Sectional and Time Series Benchmarking Procedures For Small Area Estimation," *Test: Journal of the Spanish Society of Statistics and Operations Research*, 23, 631-666