
# State-Space Modeling with Correlated Measurements with Application to Small Area Estimation Under Benchmark Constraints

**Danny Pfeffermann and Richard Tiller**

## Abstract

The problem of Small Area Estimation is how to produce reliable estimates of area (domain) characteristics, when the sizes within the areas are too small to warrant the use of traditional direct survey estimates. This problem is commonly tackled by borrowing information from either neighboring areas and/or from previous surveys, using appropriate time series/cross-sectional models. In order to protect against possible model breakdowns and for other reasons, it is often required to benchmark the model dependent estimates to the corresponding direct survey estimates in larger areas, for which the survey estimates are sufficiently accurate. The benchmarking process defines another way of borrowing information across the areas.

This article shows how benchmarking can be implemented with the state-space models used by the Bureau of Labor Statistics in the U.S. for the production of the monthly employment and unemployment estimates at the state level. The computation of valid estimators for the variances of the benchmarked estimators requires joint modeling of the direct estimators in several states, which in turn requires the development of a filtering algorithm for state-space models with correlated measurement errors. No such algorithm has been developed so far. The application of the proposed procedure is illustrated using real unemployment series.

**S$^3$RI Methodology Working Paper M03/11**

# State-Space Modeling with Correlated Measurements with Application to Small Area Estimation Under Benchmark Constraints

*Danny Pfeffermann*
*Hebrew University, Jerusalem, Israel and University of Southampton, U.K*

and

*Richard Tiller*
*Bureau of Labor Statistics*, U.S.A.

## 1. INTRODUCTION

The Bureau of Labor Statistics (BLS) in the U.S.A uses state-space models for the production of all the monthly employment and unemployment estimates for the 50 states and the District of Columbia. The models are fitted to the direct sample estimates obtained from the Current Population Survey (CPS). The use of models is necessary because the sample sizes available for the states are too small to warrant accurate direct estimates, which is known in the sampling literature as a 'small area estimation problem'. The coefficients of variation (CV) of the direct estimates vary from about 8% in the large states to about 16% in the small states. For a recent review of small area estimation methods see Pfeffermann (2002, Section 6 considers the use of time series models). The new book by Rao (2003) contains a systematic treatment of the subject

The state-space models are fitted independently between states and combine a model for the true population values with a model for the sampling errors. The published estimates are the differences between the direct estimates and the estimates of the sampling errors as obtained under the combined model. At the end of each calendar year, the model dependent estimates are modified so as to guarantee that the annual mean estimate equals the corresponding mean sample estimate. This benchmarking procedure has, however, two major disadvantages:

1- The annual mean sample estimates are still unstable because the monthly sample estimates are highly correlated due to the large sample overlaps induced by the sampling design rotation pattern underlying the CPS

2- The benchmarking is 'postmortem', after that the monthly estimates have already been published so that they are of limited use, (its main use is for long term trend estimation)

It should be mentioned also in this respect that unlike in classical benchmarking that uses external (independent) data sources for the benchmarking process, (Hillmer and Trabelsi, 1987 ; Durbin and Quenneville, 1997), the procedure described above Benchmarks the monthly estimates to the mean of the same estimates. External data to which the monthly sample estimates can be benchmarked are not available even for single months.

In this article we study a solution to the benchmarking problem that addresses the two disadvantages mentioned with respect to the current procedure. The proposed solution is to fit the model jointly to several 'homogeneous states' (states with similar 'labor force behavior', about 12-15 states in each group, see Section 6), with the added constraints

$$\sum_{s=1}^{S} w_{st}\hat{Y}_{st,\,\mathbf{model}} = \sum_{s=1}^{S} w_{st}\hat{Y}_{st,\,\mathbf{cps}}\ , \qquad t=1,2,\ldots \tag{1.1}$$

The justification for the constraints in (1.1) is that the direct CPS estimators, which are unreliable in single states, can be trusted when averaged over different states. Note in this respect that by the sampling design underlying the CPS, the sampling errors are independent between states. The basic idea behind the use of the constraints is that if all the direct sample estimates in the same group jointly increase or decrease due to some external effects not accounted for by the model, the benchmarked estimators will reflect this change much quicker than the model dependent estimators. This property is illustrated very strikingly in the empirical results presented in this article using real data. Note also that by incorporating the constraints, the benchmarked estimators for any given time *t* 'borrow strength' both from past data and cross-sectionally, unlike the model dependent estimators in present use that only borrow strength from past data.

An important question underlying the use of the constraints in (1.1) is the definition of the weights $\{w_{st},\ s=1...S,\ t=1,2,...\}$. This question is still under consideration but possible definitions include

$$w_{1st}=1/S\ ;\ w_{2st}=N_{st}/\sum_{s=1}^{S}N_{st}\ ;\ w_{3st}=1/Var_{st}(CPS) \tag{1.2}$$

where $N_{st}$ and $Var_{st}(CPS)$ are respectively the total size of the labor force and the variance of the direct sample estimate in State *s* at time *t*. The use of the weights $\{w_{2st}\}$ is appropriate when the direct estimates are proportions. The use of the weights $\{w_{1st}\}$ or $\{w_{2st}\}$ guarantees that the global benchmarked estimates for the group of States are the same as the corresponding global direct estimates in every month *t*.

Application of the proposed solution to the state-space models employed by the BLS introduces a serious computational problem. The dimension of the state vector in the separate models is of length 30 (see next section), implying that the dimension of the state vector of the joint model fitted to a group of say 12 States would be 360. A possible solution to this problem investigated in the present article is to include the sampling errors as part of the observation (measurement) equation instead of the current practice of modeling their stochastic evolvement over time and including them in the state vector. Implementation of this idea reduces the dimension of each of the separate state vectors by half, because the sampling errors make up 15 elements of the state vector.

The use of this solution, however, introduces a new theoretical problem because as already mentioned, the sampling errors are highly correlated over time, requiring the development of an appropriate filtering algorithm for fitting the model. To the best of our knowledge, filtering of state-space models with correlated measurement errors has not been studied previously in the literature. It should be emphasized that the use of the constraints (1.1) invalidates the use of the classical Kalman filter irrespective of computational efficiency. This is so because the benchmark constraints contain the observations that depend on the sampling errors. If the sampling errors and the constraints are left in the state (transition) equations, the model consists of an observation equation and state equations with disturbances that are correlated concurrently and over time. Pfeffermann and Burck (1990) consider the incorporation of constraints of the form (1.1) in a state-space model and develop an appropriate filtering algorithm but in their model there are no sampling errors so that the measurement errors are independent cross-sectionally and over time.

The present article considers therefore three main research problems:

1- Develop a filtering algorithm for state-space models with correlated measurement errors

2- Incorporate the benchmark constraints defined by (1.1) and compute the corresponding benchmarked state estimates (estimates of the true employment or unemployment figures in the present application)

3- Compute the variances of the benchmarked estimators.

Notice with respect to the third problem that the computation of the variances is under the model without the benchmark constraints. As mentioned earlier, the benchmark constraints are imposed to protect against sudden external effects on the estimated values but they are not part of the model. Indeed, the incorporation of the constraints removes the bias of the model dependent estimators in abnormal periods but inflates the variance (only mildly, see the empirical results). This is different from the classical problem of fitting regression models under linear constraints where the constraints add new information on the estimated coefficients.

In section 2 we present the State BLS models in present use. Section 3 describes the filtering algorithm for state-space models with correlated measurement errors and discusses its properties. The filter is general and is not restricted to the benchmark problem considered in the remaining sections. Section 4 shows how to incorporate the benchmark constraints and compute the variances of the benchmarked estimators. The application of the proposed procedure is illustrated in Section 5 using real series of unemployment estimates. We conclude in Section 6 by discussing some outstanding problems that need to be addressed before the procedure can be implemented for routine use.

We assume throughout the paper that the model hyper-parameters are known. In practice, the hyper-parameters will be estimated by fitting the models separately for each State, see Tiller (1992) for the estimation procedures in present use. Application of the Bootstrap method developed by Pfeffermann and Tiller (2002) accounts for the use of hyper-parameter estimation in the estimation of the prediction variances of the state vector predictors.

2- THE BLS MODEL IN PRESENT USE

In this section we consider a single State and hence we drop the subscript *s* from the notation. The model employed by the BLS combines a model for the true (estimated) State values and a model for the sampling errors and is discussed in detail, including hyper-parameter estimation and model diagnostics in Tiller (1992). Below we provide a brief description. Let $y_t$ denote the direct sample estimate at time *t* and define by $Y_t$ the true population value such that $e_t = (y_t - Y_t)$ is the sampling error.

**2.1 Model assumed for population values**

$$Y_t = b_t X_t + L_t + S_t + I_t \quad , \quad I_t \sim N(0, s_I^2)$$

$$L_t = L_{t-1} + R_{t-1} + h_{Lt} \quad , \quad h_{Lt} \sim N(0, s_L^2) \; ; \; R_t = R_{t-1} + h_{Rt} \quad , \quad h_{Rt} \sim N(0, s_R^2)$$

$$S_t = \sum_{j=1}^{6} S_{j,t} \; ; \qquad (2.1)$$

$$S_{j,t} = \cos w_j S_{j,t-1} + \sin w_j S^*_{j,t-1} + n_{j,t} \quad , \quad n_{j,t} \sim N(0, s_S^2)$$

$$S^*_{j,t} = -\sin w_j S_{j,t-1} + \cos w_j S^*_{j,t-1} + n^*_{j,t} \quad , \quad n^*_{j,t} \sim N(0, s_S^2)$$

$$w_j = 2p \; j/12 \quad ; \quad j = 1...6$$

The model defined by (2.1) but without the covariate $X_t$ is known in the literature as the Basic Structural Model (BSM). In this model $L_t$ is a trend level, $R_t$ is the slope and $S_t$ is the seasonal effect operating at time *t*. The disturbances $I_t, h_{Lt}, h_{Rt}, n_{jt}, n^*_{jt}$ are independent white noise series. See Harvey (1989) for a detailed study of this kind of models. The covariate $X_t$ represents the 'number of persons in the State receiving unemployment insurance benefits' when modeling the total unemployment figures, and represents the 'ratio between the number of payroll jobs in business establishments and the population size in the State when modeli ng 'employment to population ratios'. The coefficient $b_t$ is modeled as a random walk. Note that the trend and seasonal effects only account for the 'remainder' trend and seasonality not accounted for by the trend and seasonality of the covariate.

### 2.2 Model assumed for the sampling errors

The model assumed for the sampling error is $e_t \sim AR(15)$, which is used as an approximation to the sum of an *MA(15)* process and an *AR(2)* process.

The *MA(15)* process accounts for the sample overlap implied by the CPS sampling design. By this design, households selected to the sample are surveyed for 4 successive months, they are left out of the sample for the next 8 months and then they are surveyed again for 4 more months. This rotation scheme induces sample overlaps of 75%, 50% and 25% for the first three monthly time lags and sample overlaps of 12.5%, 25%, 37.5%, 50%, 37.5%, 25%, 12.5% at lags 9 to 15. There is no sample overlap at lags 4-8 and 16 and over. A model accounting for these autocorrelations is $MA(15)$ with zero coefficients at the lags with no sample overlap. The *AR(2)* process accounts for autocorrelations not explained by the sample overlap. These autocorrelations account for the fact that households dropped from the survey are replaced by households from the same 'census tract'. The reduced ARMA presentation of the sum of the two processes is ARMA(2,17), which is approximated by an AR(15) model.

The separate models holding for the population values and the sampling errors are cast into a single state-space model for the observations $y_t$ (the direct sample estimates). The resulting state vector consists of the covariate coefficient, the trend level, the slope, 11 seasonal components accounting for the 12 month frequency and its five harmonics, the irregular term and the concurrent and 14 lags of the sampling errors, a total of 30 elements.

The monthly employment and unemployment estimates published by the BLS are obtained under the model (2.1) as,

$$\hat{Y}_t = (y_t - \hat{e}_t) = \hat{b}_t X_t + \hat{L}_t + \hat{S}_t + \hat{I}_t \qquad (2.2)$$

3. FILTERING OF STATE-SPACE MODELS WITH CORRELATED MEASUREMENT ERRORS

In this section we assume the following state-space model

$$y_t = Z_t a_t + e_t \;\; ; \;\; E(e_t) = 0 \;\; , \;\; E(e_t e_t') = \Sigma_t \;\; ; \;\; E(e_t e_t') = \Sigma_{tt} \tag{3.1a}$$

$$a_t = T a_{t-1} + h_t \; ; E(h_t) = 0 \;\; , \;\; E(h_t h_t') = Q \;\; , \;\; E(h_t h_{t-k}') = 0 \quad k > 0 \tag{3.1b}$$

It is also assumed that $E(h_t e_t') = 0$ for all $t$ and $t$ . Clearly, what distinguishes this model from the classical state-space model is that the measurement errors $e_t$ are correlated over time. Below we propose a filtering algorithm to take account of the covariances $\Sigma_{t\,t}$ .

*At time* 1

Let $\hat{a}_1 = (\mathrm{I} - K_1 Z_1) T \hat{a}_0 + K_1 y_1$ be the filtered (updated) state estimator at time 1 where $\hat{a}_0$ is a starting estimator with covariance matrix $P_0 = E[(\hat{a}_0 - a_0)(\hat{a}_0 - a_0)']$, assumed for convenience to be independent of the observations and $K_1 = P_{1|0} Z_1' F_1^{-1}$ is the 'Kalman gain' with $P_{1|0} = T P_0 T' + Q$ and $F_1 = Z_1 \tilde{P}_{1|0} Z_1' + \Sigma_1$. The matrix $P_{1|0}$ is the covariance matrix of the prediction errors $(T\hat{a}_0 - a_1) = (\hat{a}_{1|0} - a_1)$ and $F_1$ is the covariance matrix of the innovations $n_1 = (y_1 - \hat{y}_{1|0}) = (y_1 - Z_1 \hat{a}_{1|0})$. Since $y_1 = Z_1 a_1 + e_1$ ,

$$\hat{a}_1 = (\mathrm{I} - K_1 Z_1) T \hat{a}_0 + K_1 Z_1 a_1 + K_1 e_1 \tag{3.2}$$

*At time* 2

Let $\hat{a}_{2|1} = T\hat{a}_1$ define the predictor of $a_2$ at time 1 with covariance matrix $P_{2|1} = E[(\hat{a}_{2|1} - a_2)(\hat{a}_{2|1} - a_2)']$. An unbiased estimator $\hat{a}_2$ of $a_2$ [$E(\hat{a}_2 - a_2) = 0$] based on $\hat{a}_{2|1}$ and $y_2$ is the Generalized Least Square (GLS) estimator of the random coefficient $a_2$ in the regression model

$$\begin{pmatrix} T\hat{a}_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mathrm{I} \\ Z_2 \end{pmatrix} a_2 + \begin{pmatrix} u_{2|1} \\ e_2 \end{pmatrix} \quad (u_{2|1} = T\hat{a}_1 - a_2) \tag{3.3}$$

that is,

$$\hat{a}_2 = \left[ (\mathrm{I}, Z_2^{'}) V_2^{-1} \begin{pmatrix} \mathrm{I} \\ Z_2 \end{pmatrix} \right]^{-1} (\mathrm{I}, Z_2^{'}) V_2^{-1} \begin{pmatrix} T\hat{a}_1 \\ y_2 \end{pmatrix} \tag{3.4}$$

where

$$V_2 = Var \begin{pmatrix} u_{2|1} \\ e_2 \end{pmatrix} = \begin{bmatrix} P_{2|1} & C_2 \\ C_2^{'} & \Sigma_2 \end{bmatrix} \tag{3.5}$$

and $C_2 = Cov[u_{2|1}, e_2] = TK_1\Sigma_{12}$ (follows straightforwardly from (3.2) and the previous assumptions). Notice that $V_2$ is the covariance matrix of the errors $u_{2|1}$ and $e_2$, and not of the predictors $T\hat{a}_1$ and $y_2$. By Pfeffermann (1984), the estimator $\hat{a}_2$ is the best linear unbiased predictor (BLUP) of $a_2$ based on $T\hat{a}_1$ and $y_2$, with covariance matrix

$$E[(\hat{a}_2 - a_2)(\hat{a}_2 - a_2)'] = \left[ (\mathrm{I}, Z_2^{'}) V_2^{-1} \begin{pmatrix} \mathrm{I} \\ Z_2 \end{pmatrix} \right]^{-1} = P_2 \tag{3.6}$$

*At Time* 3

Let $\hat{a}_{3|2} = T\hat{a}_2$ define the predictor of $a_3$ at time 2 with covariance matrix $E[(\hat{a}_{3|2} - a_3)(\hat{a}_{3|2} - a_3)'] = TP_2T' + Q_3 = P_{3|2}$. Denote $(\mathrm{I}, Z_2')\ V_2^{-1} = B_2 = (B_{21}, B_{22})$ such that $\hat{a}_2 = P_2 B_2 \begin{pmatrix} T\hat{a}_1 \\ y_2 \end{pmatrix} = P_2(B_{21}T\hat{a}_1 + B_{22}y_2)$. Since $y_2 = Z_2 a_2 + e_2$, it follows from (3.2) that

$$C_3 = Cov[T\hat{a}_2, e_3] = Cov[TP_2B_{21}TK_1e_1 + TP_2B_{22}e_2, e_3] = (TP_2B_{21}TK_1\Sigma_{13} + TP_2B_{22}\Sigma_{23}) \tag{3.7}$$

An unbiased estimator $\hat{a}_3$ of $a_3$ is obtained as the GLS estimator of the random coefficient $a_3$ in the regression model

$$\begin{pmatrix} T\hat{a}_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \mathrm{I} \\ Z_3 \end{pmatrix} a_3 + \begin{pmatrix} u_{3|2} \\ e_3 \end{pmatrix} \quad (u_{3|2} = T\hat{a}_2 - a_3) \tag{3.8}$$

that is,

$$\hat{a}_3 = \left[ (\mathrm{I}, Z_3^{'}) V_3^{-1} \begin{pmatrix} \mathrm{I} \\ Z_3 \end{pmatrix} \right]^{-1} (\mathrm{I}, Z_3^{'}) V_3^{-1} \begin{pmatrix} T\hat{a}_2 \\ y_3 \end{pmatrix} \tag{3.9}$$

where

$$V_3 = Var\begin{pmatrix} u_{3|2} \\ e_3 \end{pmatrix} = \begin{bmatrix} P_{3|2} & , & C_3 \\ C_3' & , & \Sigma_3 \end{bmatrix} \tag{3.10}$$

The estimator $\hat{a}_3$ is the BLUP of $a_3$ based on $T\hat{a}_2$ and $y_3$ with covariance matrix

$$E[(\hat{a}_3 - a_3)(\hat{a}_3 - a_3)'] = \left[ (\mathrm{I}, Z_3')V_3^{-1}\begin{pmatrix} \mathrm{I} \\ Z_3 \end{pmatrix} \right]^{-1} = P_3 \tag{3.11}$$

*At time* t

Let $\hat{a}_{t|t-1} = T\hat{a}_{t-1}$ define the predictor of $a_t$ at time ($t$-1) with covariance matrix $E[(\hat{a}_{t|t-1} - a_t)(\hat{a}_{t|t-1} - a_t)'] = TP_{t-1}T' + Q_t = P_{t|t-1}$ where $P_{t-1} = E[(\hat{a}_{t-1} - a_{t-1})(\hat{a}_{t-1} - a_{t-1})']$. Set the random coefficient regression model

$$\begin{pmatrix} T\hat{a}_{t-1} \\ y_t \end{pmatrix} = \begin{pmatrix} \mathrm{I} \\ Z_t \end{pmatrix} a_t + \begin{pmatrix} u_{t|t-1} \\ e_t \end{pmatrix} \quad (u_{t|t-1} = T\hat{a}_{t-1} - a_t) \tag{3.12}$$

and define

$$V_t = Var\begin{pmatrix} u_{t|t-1} \\ e_t \end{pmatrix} = \begin{bmatrix} P_{t|t-1} & , & C_t \\ C_t' & , & \Sigma_{tt} \end{bmatrix} \tag{3.13}$$

The computation of $C_t = Cov[T\hat{a}_{t-1}, e_t]$ is carried out as follows: Let, $[B_{j1}, B_{j2}] = [\mathrm{I}, Z_t']V_j^{-1}$ where $B_{j1}$ contains the first $q$ columns and $B_{j2}$ the remaining columns with $q = \dim(a_j)$. Define, $A_j = TP_jB_{j1}$, $\tilde{A}_j = TP_jB_{j2}$, $j$=2…$t$-1 ; $\tilde{A}_1 = TK_1$. Then,

$$C_t = Cov[T\hat{a}_{t-1}, e_t] = A_{t-1}A_{t-2}....A_2\tilde{A}_1\Sigma_{1t} + A_{t-1}A_{t-2}....A_3\tilde{A}_2\Sigma_{2t} + ... + A_{t-1}\tilde{A}_{t-2}\Sigma_{t-2,t} + \tilde{A}_{t-1}\Sigma_{t-1,t} \tag{3.14}$$

The BLUP of $a_t$ based on $T\hat{a}_{t-1}$ and $y_t$ and the covariance matrix of the prediction errors are obtained from (3.12)-(3.14) as,

$$\hat{a}_t = \left[ (\mathrm{I}, Z_t')V_t^{-1}\begin{pmatrix} \mathrm{I} \\ Z_t \end{pmatrix} \right]^{-1} (\mathrm{I}, Z_t')V_t^{-1}\begin{pmatrix} T\hat{a}_{t-1} \\ y_t \end{pmatrix};\ P_t = E[(\hat{a}_t - a_t)(\hat{a}_t - a_t)'] = \left[ (\mathrm{I}, Z_t')V_t^{-1}\begin{pmatrix} \mathrm{I} \\ Z_t \end{pmatrix} \right]^{-1} \tag{3.15}$$

The filtering algorithm defined by (3.15) has the following properties:

1- At every time point *t*, the filter produces the BLUP of $a_t$ based on the predictor $\hat{a}_{t|t-1} = T\hat{a}_{t-1}$ from time (*t*-1) and the new observation $y_t$ (follows from Pfeffermann, 1984).

2- Unlike the Kalman filter that assumes independent measurement errors, the filter (3.15) does not produce the BLUP of $a_t$ based on all the observations $y_{(t)} = (y_1 ... y_t)$. Computation of the latter requires joint modeling of all the observations (see comment below).

3- Empirical evidence so far suggests that the loss in efficiency from using the proposed algorithm instead of the BLUP that is based o n all the observations is mild.

Comment: For arbitrary covariances $\Sigma_{tt}$ between the measurement errors, it is impossible to construct an optimal filtering algorithm that combines the predictor from the previous time point with the new observation. By an optimal filtering algorithm we mean an algorithm that yields the BLUP of the state vector at any given time t based on the observations $y_{(t)}$. To see this, consider the simplest case of 3 observations $y_1, y_2, y_3$ with common mean $m$ and variance $s^2$. If the three observations are independent, the BLUP of $m$ based on the first 2 observations is $\bar{y}_{(2)} = (y_1 + y_2)/2$ and the BLUP based on the three observations is $\bar{y}_{(3)} = (y_1 + y_2 + y_3)/3 = (2/3)\bar{y}_{(2)} + (1/3)y_3$. The BLUP $\bar{y}_{(3)}$ is the Kalman filter predictor for time 3.

Suppose, however, that $Cov(y_1, y_2) = Cov(y_2, y_3) = s^2 r_{12}$ and $Cov(y_1, y_3) = s^2 r_{13} \neq s^2 r_{12}$. The BLUP of $m$ based on the first 2 observations is again $\bar{y}_{(2)} = (y_1 + y_2)/2$, but the BLUP of $m$ based on the 3 observations is in this case $\bar{y}^c_{(3)} = ay_1 + by_2 + ay_3$ where $a = \dfrac{(1 - r_{12})}{3 - 4r_{12} + r_{13}}$ and $b = \dfrac{(1 - 2r_{12} + r_{13})}{3 - 4r_{12} + r_{13}}$. Clearly, since $a \neq b$, the predictor $\bar{y}^c_{(3)}$ cannot be written as a linear combination of $\bar{y}_{(2)}$ and $y_3$. For example, if $r_{12} = 0.5,\ r_{13} = 0.25 \Rightarrow \bar{y}^c_{(3)} = 0.4y_1 + 0.2y_2 + 0.4y_3$.

4. INCORPORATION OF THE BENCHMARK CONSTRAINTS

### 4.1 Joint modeling of S concurrent sample estimates and their weighted mean

In this section we model jointly the direct estimates in S States and their weighted mean. We follow for convenience the BLS modeling practice and assume that the true population values and their direct sample estimates are independent between States. In Section 6 we consider extensions of the joint model to allow for cross-sectional correlations between components of the separate state vectors operating in the various States.

Suppose that the separate State models are written as in (3.1) with the sampling errors placed in the observation equation. Below we add the subscript *s* to all the model components to distinguish between the various States. Note that the observations $y_{st}$ (the direct sample estimates) and the measurement errors $e_{st}$ (the sampling errors) are scalars and $Z_t$ is a row vector (denoted hereafter as $z_t'$). Let $\tilde{y}_t = (y_{1t}...y_{St}, \sum_{s=1}^{S} w_{st} y_{st})'$ define the concurrent estimates in the S States (belonging to the same 'homogeneous group') and their weighted mean (the right hand side of the benchmark equations (1,1)). The corresponding vector of sampling errors is $\tilde{e}_t = (e_{1t}...e_{St}, \sum_{s=1}^{S} w_{st} e_{st})'$. Let $Z_t^* = \mathrm{I}_S \oplus z_{st}'$(block diagonal matrix with $z_{st}'$ in the s[th] block), $T_t^* = \mathrm{I}_S \oplus T$ , $\tilde{Z}_t = \begin{bmatrix} Z_t^* \\ w_{1t} z_{1t}' ... w_{St} z_{St}' \end{bmatrix}$, $a_t = (a_{1t}'..a_{St}')'$ and $h_t = (h_{1t}'..h_{St}')'$. By (3.1) and the independence of the state vectors and sampling errors between the States, the joint model holding for $\tilde{y}_t$ is,

$$\tilde{y}_t = \tilde{Z}_t \tilde{a}_t + \tilde{e}_t \; ; \; E(\tilde{e}_t) = 0 \; , \quad E(\tilde{e}_t \tilde{e}_t') = \tilde{\Sigma}_{tt} = \begin{bmatrix} \Sigma_{tt}, & h_{tt} \\ h_{tt}, & n_{tt} \end{bmatrix} \tag{4.1a}$$

$$\tilde{a}_t = \tilde{T} \tilde{a}_{t-1} + \tilde{\hbar}_t \; ; E(\tilde{\hbar}_t) = 0 \; , \; E(\tilde{\hbar}_t \tilde{\hbar}_t') = \mathrm{I}_S \oplus Q_{st} \; , \; E(\tilde{\hbar}_t \tilde{\hbar}_{t-k}') = 0, \; k > 0 \tag{4.1b}$$

$$\Sigma_{tt} = Diag[s_{1tt}...s_{Stt}) \; ; \; s_{stt} = Cov[e_{st}, e_{st}) \quad , \quad n_{tt} = \sum_{s=1}^{S} w_{st} \, w_{st} s_{stt} = Cov[\sum_{s=1}^{S} w_{st} e_{st}, \sum_{s=1}^{S} w_{st} e_{st}]$$

$$h_{tt} = (h_{1tt}...h_{Stt})'; h_{stt} = w_{st} s_{stt} = Cov[e_{st}, \sum_{s=1}^{S} w_{st} e_{st}]$$

Comment: The model (4.1) is the same as the separate models defined by (3.1). There is no new information in the observation equation by adding the model holding for $\sum_{s=1}^{S} w_{st} y_{st}$ .

### 4.2 Incorporating the benchmark constraints

Under the model (3.1) with the sampling errors in the observation equation, the model dependent estimator for State *s* at time *t* takes the form $\hat{Y}_{st,\text{model}} = z_{st}'\hat{a}_{st}$ (see equations 2.1 and 2.2). Thus, the benchmark constraints (1.1) can be written as,

$$\sum_{s=1}^{S} w_{st} z_{st}'\hat{a}_{st} = \sum_{s=1}^{S} w_{st} y_{st} \quad , t=1,2,\ldots \tag{4.2}$$

where $y_{st} = \hat{Y}_{st,\text{cps}}$ defines as before the direct sample estimate. By (4.1a) $\sum_{s=1}^{S} y_{st} = \sum_{s=1}^{S} z_{st}'a_{st} + \sum_{s=1}^{S} w_{st} e_{st}$. Hence, a simple way of incorporating the benchmark constraints is by imposing $\sum_{s=1}^{S} w_{st} y_{st} = \sum_{s=1}^{S} w_{st} z_{st}'a_{st}$, or equivalently, by setting

$$Var[\sum_{s=1}^{S} w_{st} e_{st}] = Cov[e_{st}, \sum_{s=1}^{S} w_{st} e_{st}] = 0 \ , \quad t=1,2,\ldots \tag{4.3}$$

This is implemented by replacing the covariance matrix $\tilde{\Sigma}_{tt}$ in the observation equation (4.1a) by the matrix $\tilde{\Sigma}_{tt}^{*} = \begin{bmatrix} \Sigma_{tt} & , & 0_{(S)} \\ 0_{(S)}' & , & 0 \end{bmatrix}$. Thus, the benchmarked estimator takes the form,

$$\hat{a}_t^{bmk} = \left[ (\mathrm{I}, \tilde{Z}_t')[V_t^{*}]^{-1} \begin{pmatrix} \mathrm{I} \\ \tilde{Z}_t \end{pmatrix} \right]^{-1} (\mathrm{I}, \tilde{Z}_t')[V_t^{*}]^{-1} \begin{pmatrix} \tilde{T}\tilde{a}_{t-1}^{bmk} \\ \tilde{y}_t \end{pmatrix} \tag{4.4}$$

where $V_t^{*} = Var\begin{pmatrix} \tilde{a}_t - \tilde{T}\tilde{a}_{t-1}^{bmk} \\ e_t^{*} \end{pmatrix} = \begin{bmatrix} P_{t|t-1}^{bmk} & , & C_t^{bmk} \\ C_t^{bmk'} & , & \tilde{\Sigma}_{tt}^{*} \end{bmatrix}$; $P_{t|t-1}^{bmk} = \tilde{T}P_{t-1}^{bmk}\tilde{T}' + \tilde{Q}$ and $C_t^{bmk} = Cov[\tilde{T}\tilde{a}_{t-1}^{bmk}, \tilde{e}_t]$.

Note that $P_{t|t-1}^{bmk}$ is the true covariance matrix of $(\tilde{a}_t - \tilde{T}\tilde{a}_{t-1}^{bmk})$ under the model. Similarly, $C_t^{bmk} = Cov[\tilde{T}\tilde{a}_{t-1}^{bmk}, \tilde{e}_t]$ is the covariance under the model. See below for the computation of $P_t^{bmk}$ and $C_t^{bmk}$.

**4.2 Computation of** $P_t^{bmk} = Var(\tilde{a}_t^{bmk} - a_t)$ and $C_t^{bmk} = Cov[\tilde{T}\tilde{a}_{t-1}^{bmk}, \tilde{e}_t]$

Let $P_t^* = \left[ (\mathrm{I}, \tilde{Z}_t^{'})[V_t^*]^{-1} \begin{pmatrix} \mathrm{I} \\ \tilde{Z}_t \end{pmatrix} \right]^{-1}$ such that $\tilde{a}_t^{bmk} = P_t^*(\mathrm{I}, \tilde{Z}_t^{'})[V_t^*]^{-1} \begin{pmatrix} \tilde{T}\tilde{a}_{t-1}^{bmk} \\ \tilde{y}_t \end{pmatrix} = P_t^* B_{t1}^{bmk} \tilde{T}\tilde{a}_{t-1}^{bmk} + P_t^* B_{t2}^{bmk} \tilde{y}_t$

$= P_t^* B_{t1}^{bmk} \tilde{T}\tilde{a}_{t-1}^{bmk} + P_t^* B_{t2}^{bmk} \tilde{Z}_t \tilde{a}_t + P_t^* B_{t2}^{bmk} \tilde{e}_t$.

By definition of $P_t^*$, $B_{t1}^{bmk}$ and $B_{t2}^{bmk}$, $P_t^* B_{t1}^{bmk} + P_t^* B_{t2}^{bmk} \tilde{Z}_t = P_t^* [P_t^*]^{-1} = \mathrm{I}$. Hence,

$\tilde{a}_t = P_t^* B_{t1}^{bmk} \tilde{a}_t + P_t^* B_{t2}^{bmk} \tilde{Z}_t \tilde{a}_t$ and

$$(\tilde{a}_t^{bmk} - \tilde{a}_t) = P_t^* B_{t1}^{bmk} (\tilde{T}\tilde{a}_{t-1}^{bmk} - \tilde{a}_t) + P_t^* B_{t2}^{bmk} \tilde{e}_t \tag{4.5}$$

It follows that,

$$\begin{aligned} P_t^{bmk} = E[(\tilde{a}_t^{bmk} - \tilde{a}_t)(\tilde{a}_t^{bmk} - \tilde{a}_t)'] = {} & P_t^* B_{t1}^{bmk} P_{t|t-1}^{bmk} B_{t1}^{bmk}{}' P_t^{*} + P_t^* B_{t2}^{bmk} \tilde{\Sigma}_{tt} B_{t2}^{bmk}{}' P_t^{*} + \\ & + P_t^* B_{t1}^{bmk} C_t^{bmk} B_{t2}^{bmk}{}' P_t^{*} + P_t^* B_{t2}^{bmk} C_t^{bmk}{}' B_{t1}^{bmk}{}' P_t^{*} \end{aligned} \tag{4.6}$$

The computation of $C_t^{bmk} = Cov[\tilde{T}\tilde{a}_{t-1}^{bmk}, \tilde{e}_t]$ is carried out by use of formula (3.14), with $\tilde{T}$, $P_j^*, (B_{j1}^{bmk}, B_{j2}^{bmk})$ replaced by $T, P_j, (B_{j1}, B_{j2})$ in the definitions of $A_j$ and $\tilde{A}_j$, $j$=2…$t$-1, and defining $\tilde{A}_1 = \tilde{T}P_1^* B_{1,2}^{bmk}$.

5. EMPIRICAL ILLUSTRATIONS

For the empirical illustrations we fitted the BLS model defined in Section 2 but without the covariate $X_t$, to the direct (CPS) unemployment estimators in the 9 Census divisions of the U.S.A. The observation period is January, 1976 – December, 2001. The last year is of special interest since it is affected by a start of a recession in March and the bombing of the New York World Trade Center in September. These two events provide an excellent test for the performance of the proposed benchmarking procedure.

The individual Division models, along with their estimated hyper-parameters, are combined into the joint model (4.1). The benchmark constraints are as defined in (1.1) with $w_{st} = 1$, so that the model dependent estimators of the Census Divisions

unemployment are benchmarked to the total national unemployment. The CV of the CPS estimator of the total national unemployment is 2%, which is considered to be sufficiently precise.

Figure 1 compares the sum of the model dependent predictions over the 9 Divisions without the benchmark constraint with the CPS national unemployment estimator. In the first part of the observation period the sum of the model predictors are close to the CPS estimator. In 2001 there is evidence of systematic model underestimation. This is better illustrated in Figure 2, which plots the difference between the total of the model predictors and the CPS estimator. As can be seen, starting in March, 2001, all the differences are negative and in some months the absolute difference is larger than twice the standard deviation of the CPS estimator.

Figures 3-11 display the model dependent predictors, the benchmarked predictors and the direct CPS estimators from January 2000 for each of the 9 Census divisions. Except for New England, the Benchmarked estimators are seen to correct the underestimation of the model dependent estimators in the year 2001. The reason that this bias correction does not occur in New England is that in this division, the model dependent predictors are actually higher than the CPS estimators, which serves as an excellent illustration for the need to apply the benchmarking in 'homogeneous groups' (see Section 6).

Table 1 shows the means of the monthly ratios between the benchmarked predictors and the model dependent predictors for each of the 9 Census divisions in the year 2001. The means are computed separately for the estimation of the total unemployment figures and for estimation of the trend levels ($L_t$ in equation 2.1). As can be seen, the means of the ratios are all greater than one but the largest means are about 4% indicating that the effect of the benchmarking is generally mild.

## 6. CONCLUDING REMARKS, OUTLINE OF FUTURE RESEARCH

Benchmarking of small area model dependent estimators to agree with the direct sample estimates in 'large areas' is a common requirement by statistical agencies producing official statistics. This article shows how this requirement can be implement with state-space models. When the direct estimates are obtained from a survey with correlated sampling errors like in labor Force surveys, the benchmark constraints cannot be incorporated within the framework of the Kalman filter, requiring instead the development of a filter with correlated measurement errors. This filter is needed to allow the computation of the variances of the benchmarked estimators under the model. Unlike the Kalman filter, filtering with correlated measurement errors does not produce the BLUP predictors based on all the observations but empirical evidence obtained so far indicates that the loss of efficiency by use of the proposed filtering algorithm is mild. Further empirical investigation is needed to ascertain this property.

An important condition for the success of the benchmarking procedure is that the small areas (States in the present application) are 'homogeneous' with respect of the behavior of the true (estimated) quantities of interest (the true employment or unemployment figures in the present application). The need for the fulfillment of this condition is illuminated in the empirical illustrations where the benchmarking of the Census Division estimates to the direct (CPS) national estimate increased the model dependent predictors in New England instead of decreasing them. This happened because unlike in all the other divisions, the model dependent predictors in New England were already higher than the corresponding CPS estimators. Since the benchmarking of the employment and unemployment estimates In the U.S.A. is currently planned for the State estimates, our next major task is to classify the 50 States and the District of Columbia into homogeneous groups.

Several factors need to be taken into account when defining the groups. Geographic proximity to account, for example, for weather conditions, breakdown of the Labor Force into the major categories of employment (percentages employed in manufacturing, services, farming etc.) and the size of the States (to avoid the possibility that large States will dominate the benchmarking in small States) are obvious candidate factors that should

be considered. Obviously, the behavior of past estimates and their components like the trend and seasonal effects should be investigated for a successful classification of the States. Accounting for all the factors mentioned above for the grouping process might result in very small groups but it should be emphasized that the groups must be sufficiently large to justify the benchmarking to the corresponding global CPS estimate in the group. Thus, the sensitivity of the benchmarking process to the definition of the groups needs to be investigated.

Another area for future research is the development of a smoothing algorithm that accounts for correlated measurement errors. Clearly, as new data accumulate it is desirable to modify past predictors, which is particularly important for trend estimation. Last, the present BLS models assume independence between the state vectors operating in separate States. It can be surmised that changes in the trend or seasonal effects are correlated between homogeneous States and accounting for these correlations might improve further the efficiency of the predictors. In fact, the existence of such correlations underlies implicitly the use of the proposed benchmarking procedure. Accounting explicitly for the existing correlations is simple within the joint model defined by (4.1) and may reduce quite substantially (but not eliminate) the effect of the benchmarking on the model dependent predictors.

## REFERENCES

Durbin, J. and Quenneville, B. (1997). Benchmarking by State Space Models. *International Statistical Review*, **65**, 23-48.

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

Hillmer, S.C., and Trabelsi, A. (1987). Benchmarking of Economic Time Series, *Journal of the American Statistical Association*, **82**, 1064-1071.

Pfeffermann, D. (1984). On Extensions of the Gauss-Markov Theorem to the case of stochastic regression coefficients. *Journal of the Royal Statistical Society*, Series B, **46**, 139-148.

Pfeffermann, D. (2002). Small area estimation- new developments and directions. *International Statistical Review*, **70**, 125-143.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, **16**, 217-237.

Pfeffermann, D., and Tiller, R. B. (2002). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. Working Paper, Department of Statistics, Hebrew University, Jerusalem, Israel.

Rao, J. N. K. (2003), *Small Area Estimation*. New York: Wiley.

Tiller, R. B. (1992). Time series modeling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, **8**, 149-166.

**Means of Ratios Between Benchmarked and Model Dependent Predictors of Total Unemployment and Trend in Census Divisions, 2001**

| Division | Prediction of Unemployment | Prediction of Trend |
|---|---|---|
| *New England* | 1.015 | 1.015 |
| Middle Atlantic | 1.011 | 1.012 |
| East North Central | 1.036 | 1.036 |
| West North Central | 1.020 | 1.020 |
| South Atlantic | 1.030 | 1.030 |
| East South Central | 1.040 | 1.040 |
| West South Central | 1.043 | 1.043 |
| Mountain | 1.016 | 1.016 |
| Pacific | 1.038 | 1.038 |

Figure 1. **Monthly Total Unemployment**
National CPS and Sum of Division Model Estimates
(100,000)

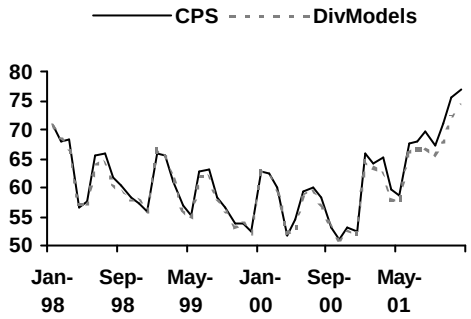


Figure 2. **Monthly total Unemployment**
Difference between Sum of Division Model Estimates and CPS SD(CPS) ≈ 1.35 (100,000)

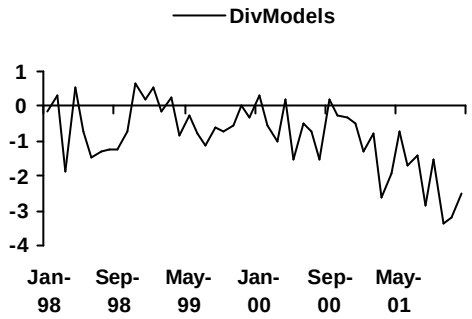


Figure 3. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**New England** (10,000)

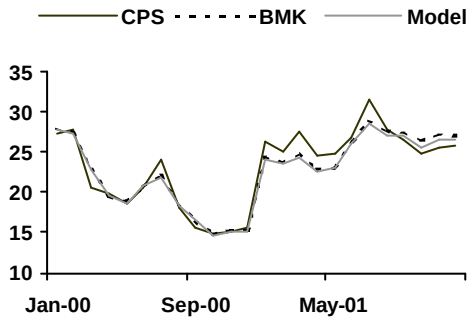


Figure 4. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**Middle Atlantic** (100,000)

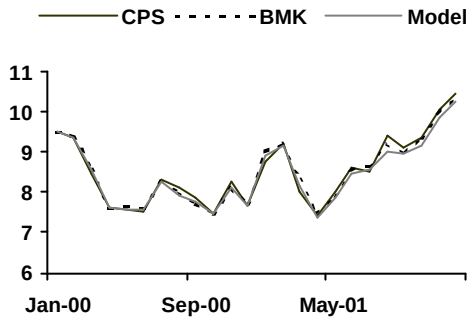


Figure 5. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**East North Central** (100,000)

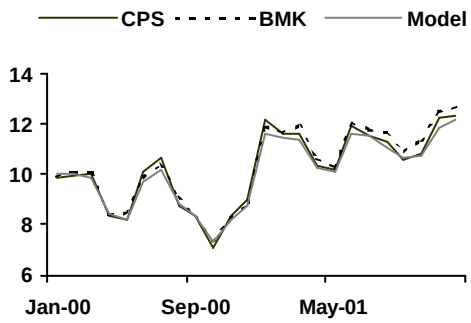


Figure 6. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**West North Central** (10,000)

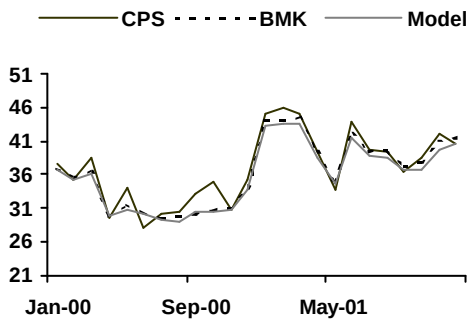

Figure 7. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**South Atlantic** (100,000)

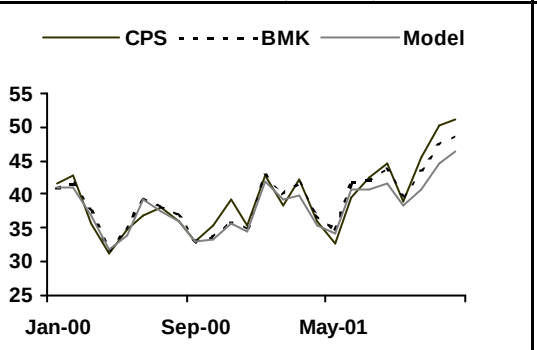


Figure 8. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**East South Central** (10,000)

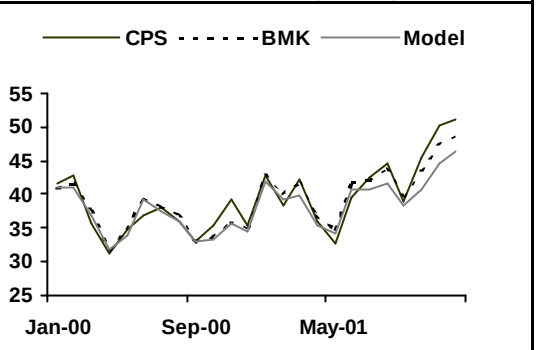


Figure 9. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**West South Central** (100,000)

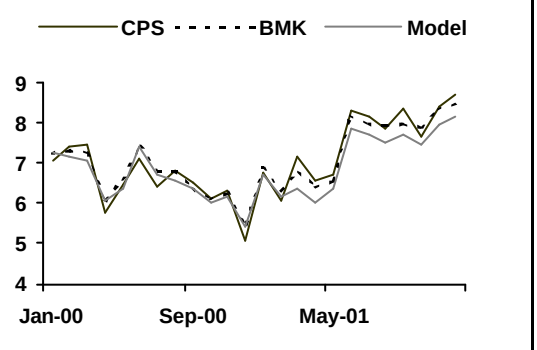


Figure 10. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**Mountain** (10,000)

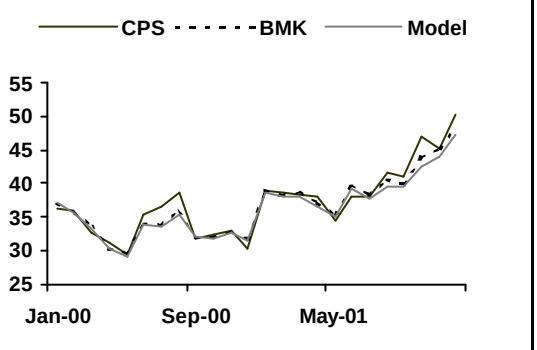


Figure 11. CPS, Model and Benchmark Estimates of Monthly Total Unemployment
**Pacific** (100,000)