Model-Based Labor Force estimates for Sub-National Areas with large Survey Errors

Richard Tiller
Statistical Methods Staff
Office of Employment and Unemployment Statistics
Bureau of Labor Statistics
Washington, D.C.
Tiller_r@bls.gov
March 2006

Key Words: Small area estimation, Seasonal Adjustment, state-space models, survey error

Abstract

To produce monthly employment and unemployment estimates for all 50 States, the District of Columbia and selected metropolitan areas, the Bureau of Labor Statistics (BLS) uses time series models applied to estimates from the Current Population Survey (CPS). The CPS design raises two types of problems for time series modeling. The first, and most obvious, is the variability in the data due to small samples. Secondly, the CPS has an overlapping design that induces strong autocorrelations in the survey errors (SE).

When fitting time series models to CPS State data, it is important to explicitly account for the two important properties of the SE. This is done by using a signal plus-noise model where the monthly CPS estimates are treated as stochastically varying time series obscured by survey error. Given a model for the true labor force values (signal) and survey error (SE) variance-covariance information, we construct an estimator or filter that suppresses SE along with seasonal variation in the population. We also extend this model to a bivariate form to incorporate information in related series.

Introduction

To produce monthly employment and unemployment estimates for all 50 States and the District of Columbia, the Bureau of Labor Statistics (BLS) uses time series models applied to estimates from the Current Population Survey (CPS). While the CPS provides reliable estimates of national aggregates, its sample is spread too thinly geographically to provide acceptable reliability at the State and sub-State level. The time series approach provides a way of reducing variability by pooling survey data across time within a State. This approach treats the unknown true population values as stochastic and uses signal extraction techniques to estimate these true values from a noisy series. Because of the complex design of the CPS, the behavior of the observed sample estimates differs in important ways from the behavior of the hypothetical true values. An overlapping sample design and

changes in reliability induce strong positive autocorrelation and heteroscedasticity in the sampling errors. To account for the dynamic behavior of the true values and the errors induced by the survey design, a model of the population values is combined with a model of the survey errors. The population model is formulated to produce seasonally adjusted series and uses covariate series along with the CPS series in a bivariate model to estimate the CPS trend.

Under the Bureau of Labor Statistics (BLS) Federal-State cooperative Local Area Unemployment Statistics (LAUS) program, State model-based estimates provide controls for over 7,000 sub-State area estimates produced by various methods (Bureau of Labor Statistics, Handbook of Methods). BLS is responsible for the methodology, technical procedures, validation, and publication of the estimates. State employment security agencies are responsible for preparing the monthly estimates according to BLS standards. These State and area estimates are used by a wide variety of customers. Federal programs base allocations to States and areas on the data, as well as eligibility determinations for assistance. State and local governments use the estimates for planning budgetary purposes and to determine the need for local employment and training services. Private industry and individuals use the data to compare and assess local labor market developments.

History of State Labor Force Estimation

Historically, CPS samples have not been sufficiently large, in all but the most populous States, to produce reliable monthly estimates directly from the survey. As a result, indirect methods have been used to estimate employment and unemployment. As far back as 1960, statewide estimates were developed using the Handbook method. In 1978, following a series of sample expansions, BLS adopted the direct use of monthly CPS estimates in the 10 most populous States and later one additional State was added. In the remaining States the Handbook continued to be used but adjusted to a 6 month moving CPS average and at the end of the year benchmarked to the CPS annual average.

For 39 States and the District of Columbia, the Handbook method was replaced in 1989 by a time series approach to modeling State unemployment and employment series. For each State, a dynamic regression model was developed with CPS as the dependent variable and auxiliary series from unemployment insurance data and an employer survey of payroll employment as regressors with stochastic coefficients. To produce seasonally adjusted estimates, the model estimates were adjusted externally by X-11 ARIMA. At the end of the year, model estimates for each State were benchmarked to their respective CPS annual average.

In 1994 a major revision to the structure of the time series model was introduced to explicitly control for characteristics of the survey error. Known as the signal plus noise approach, a model of the survey errors based on design information is combined with a dynamic regression model of the true values to decompose the CPS into estimates of the true population values and the survey errors. In 1996 these models were extended to all States. The practice of external seasonal adjustment with X-11 and benchmarking to annual average State CPS estimates was continued.

Finally, in 2005, the dynamic regression model was replaced by a bivariate structural time series model of the true CPS values that directly produces seasonally adjusted estimates. In addition, the practice of benchmarking to the State CPS annual average was replaced by real time and historical benchmarking of estimates to the monthly national CPS.

State CPS series

The State series modeled consists of all 50 States and the District of Columbia. The two most populous States, California and New York are subdivided into two areas. California is divided into Los Angeles-Long Beach-Glendale metropolitan division and the balance of California, and New York is divided into New York City and balance of New York. In addition, five major metropolitan CPS series are also modeled (see details below).

Small samples in each State result in unacceptably high variation in the monthly CPS estimates of State employment and unemployment. The table below gives the State sample sizes, and standard errors, and coefficient of variation (CVs) for the unemployment rate and the employment-to-population ratio assuming an unemployment rate of 6 percent for the States and the Nation as a whole.

Reliability of State CPS Monthly Estimators Under the 2000 Design

	Unemployment Rate (%)		Employment-to-Population Ratio (%)		Number of Households in sample
	CV%	SE	CV%	SE	
Nation	1.9	0.11	0.22	0.14	71,681
States					
median	17.51	1.05	2.03	1.32	1,204
range	13.00 - 20.35	0.78 - 1.22	0.92 - 3.47	0.58 - 1.92	700 - 5,344

To produce less variable labor force estimates as well as produce more stable seasonally adjusted estimates, BLS introduced time series models to ‘‘borrow strength’’ over time by using historical series of sample observations for a given State to increase its effective sample size. On average the variance of month-to-month change in the model estimates is about one third of the size of the CPS variance.

The models are based on the signal plus-noise approach to small area estimation where the monthly CPS estimates are treated as stochastically varying time series obscured by survey error (Tiller, 1992). Given a model for the true labor force values (signal) and survey error (SE) variance-covariance information, we construct an estimator or filter that suppresses SE along with seasonal variation in the population. We also extend this model to a bivariate form to incorporate information in related series. The estimator of the signal developed from this general model is optimal under the model assumptions and is design consistent under general conditions. This approach was first suggested by Scott and Smith (1974) and has

been more fully developed by Bell and Hillmer (1990), Binder and Dick (1990) and Pfefferman (1992).

Because the model estimates depend heavily on historical data they are slow to respond to sudden permanent shifts in CPS levels. To improve the robustness of the models, State estimates are constrained to sum to the National CPS estimates.

MODEL OF CPS SERIES

The CPS design raises two types of problems for time series modeling. The first, and most obvious, is the variability in the data due to small samples. Moreover, this variability changes systematically over time with changes in the population values, sample size, and sample design. Secondly, the CPS has an overlapping design that induces strong autocorrelations in the survey errors (SE).

When fitting time series models to CPS State data, it is important to explicitly account for these two properties of the SE. This is done by fitting separate (independent) time series models for the survey errors and the population values. An important target variable is the seasonally adjusted population series. Accordingly, we formulate the population model as a basic structural model that decomposes a single series into stochastic trend seasonal and irregular components (Harvey, 1989). SE is treated as an additional unobserved component of the time series, with the special advantage that its variance-covariance structure is objectively identified by design information. In effect this combined model filters out both survey error and seasonality.

Two models—one for unemployment and one for employment—are developed for each State. Let the direct survey estimator (CPS) for a given State, y_t (either unemployment or employment) be represented as the sum of two independent processes, the true population values, Y_t, and the survey errors, e_t, arising from survey only a portion of the total population

$$y_t = Y_t + e_t \tag{1}$$

Survey Error Model

The CPS survey's complex design induces heteroscedasticiy and autocorrelation in the survey errors, which is represented below,

$$\begin{aligned} & e_t \sim N\left(0, s^2_{e,t}\right); \quad E\left(e_{e,t}\ e_{e,t}\right) = s_{e,t\,t} \\ & s^2_{e,t} = D_y S^2_{y_t} \\ & S^2_{y_t} = \frac{N_t^2}{n_t} P_t\left[1-P_t\right], \quad P_t = \frac{Y_t}{N_t} \end{aligned} \tag{2}$$

where,

D_Y = design effect for Y_t, N_t = population size; n_t = sample size

We formulate the variance-covariance structure of each State's CPS series as a linear stochastic process since it can be easily put into a state-space form, which as explained later, has important advantages for estimation. Specifically, the SE model is specified as a ARMA(2,17) process with changing variance.

$$e_t = g_t e_t^* \tag{3}$$

where,

$$g_t = s_e / s_{e^*}$$
$$e_t^* = e_{1,t}^* + e_{2,t}^* ,$$
$$e_{1,t}^* = a_{1,t} + q_1 a_{1,t-1} + q_2 a_{1,t-2} + q_3 a_{1,t-3} + q_{12} a_{1,t-12} + q_{13} a_{1,t-13} + q_{14} a_{1,t-14} + q_{15} a_{1,t-15}$$
$$e_{2,t}^* = f_1 e_{2,t-1}^* + f_2 e_{2,t-2}^* + a_{2,t}$$

The a_t are uncorrelated independent disturbance terms with zero mean and fixed variance. The scale factor, g_t accounts for the changing CPS variance. The autocorrelated SE is modeled by e_t^* as the sum of a MA(15) process, $e_{1,t}^*$, and an AR(2) process, $e_{2,t}^*$, which results in an ARMA(2,17) process for the aggregate. Each of these processes represents specific characteristics of the survey design described below.

The MA (15) model accounts for the overlap of identical households in the CPS sample induced by the 4-8-4 rotation pattern. This scheme results in a sample overlap of 75%, 50% and 25% for the first three monthly time lags, overlaps of 12.5%, 25%, 37.5%, 50%, 37.5%, 25% and 12.5% at lags 9 to 15 and no overlap at lags 4-8 and 16 and over. Accordingly, $e_{1,t}^*$ has non-zero autocorrelations at only those lags corresponding to a sample overlap.

Another important feature of the CPS survey design is that panels rotating out of the sample are replaced by panels sampled from the same census block, implying that the survey errors are actually correlated even outside the sample overlap period. A model accounting for this source of autocorrelations is the AR (2) model, $e_{2,t}^*$. The two roots of the AR operator are real, which captures the slow decay in the autocorrelations with increasing lag length.

Model of the Signal

We refer to the true population values as the "signal". Since a seasonally adjusted series is an important target, we specify the signal in terms of the classical time series decomposition,

$$Y_t = T_{Y,t} + S_{Y,t} + I_{Y,t} \tag{4}$$

where $T_{Y,t}$ is the trend-cycle, $S_{Y,t}$ the seasonal, and $I_{Y,t}$ the irregular component.

The trend-cycle and seasonal components have mutually independent normal disturbance terms that cause them to drift slowly over time. The variances of these disturbances constitute the "hyperparameters" of the signal and determine the properties of the individual components. A positive variance for a component implies that it is stochastic (not perfectly

predictable from past history), while a zero variance implies deterministic behavior (a fixed pattern over time). The irregular is treated as an uncorrelated zero mean disturbance with fixed variance (white noise process).

Trend-Cycle

The trend-cycle is represented as a local approximation to a linear trend-cycle with a random level, $T_{Y,t}$ and slope, $R_{Y,t}$

$$\begin{aligned} T_{Y,t} &= T_{Y,t-1} + R_{Y,t-1} + n_{T_Y,t}, \quad n_{T_Y,t} \sim NID\left(0, s^2_{T_Y,t}\right) \\ R_{Y,t} &= R_{Y,t-1} + n_{R_Y,t}, \qquad\quad n_{R_Y,t} \sim NID\left(0, s^2_{R_Y,t}\right) \end{aligned} \tag{5}$$

This simple trend-cycle model can accommodate patterns ranging from an irregular cyclical series to a linear trend with a fixed rate of growth. Shifts up or down in the level give the trend-cycle a jagged appearance while changes in slope are inherently more gradual, causing acceleration, deceleration or change in direction. Overall smoothness, therefore, depends on the magnitude of the level variance relative to the slope variance. A small level variance relative to the slope variance implies a smooth trend (i.e. few turning points). In contrast, if the slope variance is small relative to the level variance, the trend will frequently change direction. In general, the trend-cycle is a combination of a long run trend and more variable cyclical fluctuations.

Seasonal

The seasonal component is specified in terms of 6 trigonometric terms associated with the 12 month periodicity and its harmonics (periodicities of 6, 4, 3, 2.4, and 2 months).

$$S_{Y,t} = \sum_{j=1}^{6} S_{Y,j,t} \tag{6}$$

Each frequency component is represented by a pair of stochastic variables expressed in recursive form as a vector autoregressive process,

$$\begin{aligned} S_{Y,j,t} &= \cos w_j S_{Y,j,t-1} + \sin w_j S^*_{Y,j,t-1} + h_{Y,j,t}, \quad h_{Y,j,t} \sim NID(0, s^2_{S_Y}) \\ S^*_{Y,j,t} &= -\sin w_j S_{Y,j,t-1} + \cos w_j S^*_{Y,j,t-1} + h^*_{Y,j,t}, \quad h^*_{Y_j,t} \sim NID(0, s^2_{S_Y}) \\ w_j &= 2p_j / 12, \quad j = 1...6 \end{aligned}$$

The disturbance terms, $h_{Y,j,t}$ and $h^*_{Y_j,t}$ allow the seasonal effects to evolve stochastically over time but in a way that guarantees that the expectation of the sum of 12 successive seasonal effects will be zero.

$$E\left(\sum_{i=0}^{11} S_{Y,t+i}\right) = 0$$

The seasonal disturbances are assumed to have a common variance, $s^2_{S_Y}$, thus the change in the seasonal pattern depends upon a single parameter. If the common variance is zero then the seasonal pattern is fixed over time (Harvey 1989).

Irregular

The irregular component is a residual not explained by the components discussed above. This component is specified as consisting of a single white noise disturbance with a zero mean and constant variance

$$I_t = n_{I,t}, \quad n_{I,t} \sim NID\left(0, s_I^2\right) \tag{7}$$

If the variance for this component is zero, then the irregular is identically zero and can be dropped from the model. From a diagnostic point of view it is useful to start with an irregular component since it is sensitive to outliers and therefore useful for identifying the presence of the latter in the series.

Outliers

Time series are occasionally influenced by exogenous disturbances that shift the level of the series. We model this type of behavior as either temporary or permanent without stochastic disturbance terms,

$$O_{Y,t} = \sum_j l_{Y,j} z_{Y,j,t} \tag{8}$$

where $z_{Y,j,t}$ is an indicator variable identifying when the outlier effect first occurred and its duration. The coefficient $l_{Y,j}$ is the change in the level of the series at time j. For an outlier that affects only one observation,

$$z_{Y,j,t} = \begin{cases} 1 & \text{if } t = j \\ 0 & \text{if } t \neq j \end{cases}$$

and for a permanent shift in level,

$$z_{Y,j,t} = \begin{cases} 1 & \text{if } t \geq j \\ 0 & \text{if } t < j \end{cases}.$$

Covariate Model

The above model uses information in a single State CPS time series. A natural extension of the structural model is to allow one or more of the unobserved components of the signal to be related to corresponding components in another series. A common core of State specific monthly covariates have been developed from auxiliary data sources – unemployment insurance claims from the Federal-State Unemployment Insurance System is used for the unemployment model, and nonagricultural payroll employment estimates from the Current Employment Statistics (CES) program for the employment model.

The model for the covariate, X_t, follows the same basic structural form as for Y_t, with stochastic trend, seasonal and irregular components,

$$X_t = T_{X,t} + S_{X,t} + I_{X,t} \tag{9}$$

with hyperparameters $\Omega_X = \left\{ s^2_{T_X}, s^2_{R_X}, s^2_{S_X}, s^2_{I_X} \right\}$.

The two series, Y_t and X_t are treated as related in a bivariate time series model with contemporaneous correlations between their respective trend disturbances.

$$\begin{aligned} E\left(n_{T_Y,t}, n_{T_X,t}\right) &= s_{T_Y,T_X} \\ E\left(n_{R_Y,t}, n_{R_X,t}\right) &= s_{R_Y,R_X} \end{aligned} \tag{10}$$

Correlations between irregular and seasonal components could also be allowed, but because they are very weak it is not worth the additional complexity.

This model is a special case of the seemingly unrelated time series equations (SUTSE) model (Harvey, 1989). It allows for a cointegrating relationship between the two trends such that some linear combination of the two is either stationary in levels or first differences. That is, this model allows for a stable relationship between either levels or slopes or both. For the most part the empirical correlations are not strong enough to imply the presence of cointegration.

ESTIMATION

Our combined CPS model with covariates has a complex form with a very large number of parameters to estimate. Two major simplifications make it feasible to implement. First, the availability of design based information allows us to estimate the SE parameters independently of the time series parameters. Secondly, estimation of the unobserved component series is simplified by casting the model into state-space form.

Estimation of Unknown Model Parameters

When fitting time series models to CPS data, our objective is to account for the variances and correlations of the survey errors. Our combined model consists of up to 17 SE ARMA coefficients, SE variance estimates for each time point and up to 10 hyperparameters associated with the time series components. Estimation is complicated since neither the population series nor the SE is observable and the only data available for fitting the time series model is the observed survey series. Estimating simultaneously the parameters for the combined model is problematic because of their large number and the identification problems arising when the SE model, because it is autocorrelated, is partially confounded with the model for the population values.

To overcome the identification problem as well as to simplify estimation we use a two step process. First we estimate the survey error model parameters from design based information independently of the time series model of the signal. In the second step we estimate the parameters of the signal from the survey series, holding the parameters of the SE model

fixed. The model holding for the signal is thus identified and estimated from the observed series. See Bell and Hilmer (1990), Tiller (1992), Pfeffermann, Feder and Signorelli (1998) and Harvey and Hwang (2000) for application of this modeling paradigm.

The estimation of the two sets of parameters—those related to the SE model and to the time series model—are discussed below.

Variance estimates

To assess the reliability of national statistics on an ongoing basis, the Census Bureau uses the method of generalized variance functions (GVF). This approach fits variance curves to groups of statistics for which variances have been estimated directly from the survey replicate variances. This curve is then generalized over time and to other statistics not used in the fit but with similar coefficients of variations. The form of the GVF is

$$V_{y_t}^2 = a + \frac{b}{y_t} \tag{11}$$

where $V_{y_t}^2$ is the rel-variance of the estimate y_t and a and b are estimated parameters.

This approach raises problems at the State level where sample sizes are small and non-self representing (NSR) samples often have a substantial contribution to total variance. Since only one PSU is selected per NSR stratum, there is no direct way to estimate the NSR sampling variance without collapsing strata. This approach creates an artificial between stratum variance component that inflates the variance estimate.

At the State level, the variance parameters are computed indirectly as described below.

$$b_t = k_t \frac{N}{n} D_y, \quad a_t = -\frac{b_t}{N} \tag{12}$$

where,

D_y = within PSU design effect
N = total population size
n = sample size
k_t, = ratio of total to between PSU variance.

The State design effect is based on the national design effect adjusted for State differences in non-interview rates. The between PSU variance (incorporated in k_t,) is computed directly from the latest Census data and adjusted to current labor force levels.

Survey Error Autocorrelations

The other key set of survey error parameters are the survey error autocorrelations (SEA). The SEA could be estimated from replicates but we do not have a very long time series of replicates. To over come this problem we use a different approach to produce SEA

estimators. This approach is based on a simple method for estimating the SEA from the separate panel (rotation group) estimates (see Zimmerman and Robison,1995 and Pfeffermann, Feder and Signorelli 1998,). A panel is defined as the set of sampling units joining and leaving the sample at the same times. The CPS sample consists of 8 such panels in every month, where each panel is a representative sample of the population.

The "direct" sample estimator before compositing is an average of the panel estimates. Pseudo- errors may be computed as the deviation of each panel estimates around the mean of the panel estimates for each month (after correcting for time in sample bias). SEA may be estimated directly from the pseudo errors. For further details, see Pfeffermann, Tiller and Zimmerman (2000).

Originally we fit an ARMA(2,17) model to the autocorrelations, but because it took up considerable computer time to estimate the parameters we approximated it with an AR(15). The AR coefficients were quickly computed using the Yule-Walker equations (Box and Jenkins, 1976). We chose the order 15 because it reproduces the empirical SEA for the first 15 lags, and because the partial autocorrelations beyond that lag computed from the empirical SEA were found to be very low (usually below 0.05). Having a model that produces the 'correct' first 15 SEA for the CPS series is useful because these are the more important correlations and hence the AR(15) structure is robust against possible departures from the true underlying model.

Time Series Parameters

The second step is to estimate the hyperparameters of the time series model. Our complete model is represented in compact form as,

$$
\begin{aligned}
& y_t = Y_t + e_t \\
& Y_t = T_{Y,t} + S_{Y,t} + I_{Y,t}, \quad e_t = g_t e_t', \quad e_t' = f_1 e_{t-1}' + \ldots + f_{15} e_{t-15}' + n_{e(t)}, \quad s_{n_e}^2 = 1, \\
& X_t = T_{X,t} + S_{X,t} + I_{X,t} \\
& \text{hyperparameters: } \Omega = \left\{ s_{T_Y}^2, s_{R_Y}^2, s_{S_Y}^2, s_{I_Y}^2, s_{T_X}^2, s_{R_X}^2, s_{S_X}^2, s_{I_X}^2, s_{T_Y,T_X}, s_{R_Y,R_X} \right\}
\end{aligned}
\tag{13}
$$

To simplify estimation, this model is put into the state-space form as described below. Once in that form, the Kalman Filter (given below) computes the conditional density for a single observation, y_t, which is normal with mean, $y_{t/t\text{-}1}$ and variance, $f_{t/t\text{-}1}$. The joint density of the T sample observations is the product of these individual densities. Given estimates of the SE model parameters, $\hat{s}_{e,t}^2$, $\hat{f}_i$, and initial values for the state vector, a_0 , and its covariance matrix, P_0 , the unknown hyperparameters of the time series components are estimated by maximum likelihood using the prediction error decomposition of the likelihood function, L (Harvey, 1989),

$$
L\left[\Omega \middle| (y_{l+1}, X_{l+1}), \ldots, (y_t, X_t); \hat{s}_{e,t}^2, \hat{f}_i, a_0, P_0 \right] = \sum_{t=l+1}^{T} \left[\ln \left| f_{t/t-1} \right| + u_{t/t-1}' f_{t/t-1}^{-1} u_{t\,t-1} \right]. \tag{14}
$$

In the equation above, l is the number of non-stationary elements in the state vector which determines the number of observations required to form priors of these elements and $u_{t/t-1}$ is the one-step ahead prediction error described below. The estimators of the variance components are obtained by maximizing L with respect to W using a quasi-Newton routine to search the parameter space.

Estimation of Time Series Components

The key to simplifying estimation of the unobserved components of the CPS series is to represent the model in state-space form. The nice thing about this set up is that estimation of complex time series models is implemented with very simple algorithms while other estimation approaches, such as regression, are very difficult to implement. The fundamental algorithm is the Kalman filter (KF) for estimation in real time as new observations become available each month and a smoother algorithm which updates the KF for estimation in historical time when the number of observations is fixed.

Consider the multivariate case where y_t is now a vector of the observed series at time t and e_t a corresponding vector of "measurement" errors. Let the unobserved components be contained in the "state" vector, a_t. The standard state-space form consists of a transition equation and an observation equation. The transition equation describes the dynamic behavior of the state vector as a first order vector autoregressive process,

$$a_t = F_t a_{t-1} + G_t h_t \;; \quad h_t \sim NIID(0, Q), \; Q = Diag\left(s^2_{h_j}\right) \tag{15}$$

The observation equation represents the observed data as a linear combination of the state variables plus uncorrelated "measurement error".

$$\tilde{y}_t = Z_t a_t + \tilde{e}_t; \quad E(\tilde{e}_t) = 0, \quad E(\tilde{e}_t \tilde{e}_t') = \begin{cases} \Sigma_{tt} & t = t \\ 0 & t \neq t \end{cases} \tag{16}$$

The observation matrix, Z_t, and transition matrix, F_t, and disturbance matrix G_t are known non-stochastic matrices, and h_t is a vector containing the white noise disturbances of the component models with a diagonal covariance matrix, Q.

While the transition equation may appear to be restrictive, a surprisingly wide range of models can be transformed to an AR form by constructing artificial state variables. For all processes that can be given a state-space representation, the KF algorithm provides a simple unified approach to prediction and estimation in real time. The trend, seasonal, and irregular components easily fit into the state-space form since they are already expressed as first order AR processes.

For our application, the restriction that measurement error (in our case survey error), e_t, be independent of its previous values with fixed variance requires modification of the observation and transition equations. When measurement error is correlated, as is the case with CPS survey error, the usual practice is to remove it from the observation equation and add it to the state vector. This is possible provided the measurement error can be represented by a

model that has a state-space form. Our CPS survey error model is an ARMA process which is easily translated into a vector autoregressive form and, then, included in the state vector along with the other components. This modification results in a transition equation that is augmented with the SE model and an observation equation with no measurement error.

$$\tilde{y}_t = Z_t a_t \tag{17}$$

The first row of the observation matrix contains mostly zeros and ones to select the relevant components that sum to the observed CPS values as well as outlier regression variables and the CPS variance inflation factors. The second row selects out the components that sum to the covariate values.

We are interested in the signal or its non-seasonal component which are linear combinations of the state vector, which we represent as,

$$y_{Signal,t} = z_t a_t \tag{18}$$

The vector z_t retrieves the signal from the state vector. It is identical to Z_t except that the components of the state vector not associated with the signal are zeroed out.

Given the sample data $(y_1,\ldots,y_n)$, the problem is to predict the state vector. Assuming the transition equation disturbances are normal, the predictor, $\hat{a}_{t/n}$, is the conditional expectation given the data, which is the minimum variance unbiased predictor with covariance matrix $P_{t/n}$

$$\hat{a}_{t/n} = E\left(a_t / y_n\right), \quad P_{t/n} = E\left[\left(a_t - \hat{a}_{t/n}\right)^2 / y_n\right] \tag{19}$$

E denotes the expectation operator and n indexes the latest period for which data are available. The value we predict may refer to the present $(t = n)$, past $(n > t)$ or future $(n < t)$.

Without the normality assumptions, this predictor is only a linear projection rather than a conditional expectation and therefore is a best linear unbiased predictor (BLUP).

Filtering

The KF provides a recursive formula for calculating the conditional mean of the state vector at time t, $\hat{a}_{t/t}$, and its covariance matrix by means of updating the estimator, $\hat{a}_{t/t-1}$. It is constructed from two sets of equations derived from the state-space equations – the prediction equations and update equations. The prediction equations compute the mean vector and covariance matrix for the conditional density based on sample data prior to time t, given the variance parameters, and the initial state vector $\hat{a}_0$,

$$\hat{a}_{t/t-1} = F_t \hat{a}_{t-1/t-1}, \quad P_{t/t-1} = F_t P_{t-1/t-1} F_t' + G_t Q G_t' \tag{20}$$

the mean $y_{t/t-1}$, and variance $f_{t/t-1}$ of the conditional density of the sample observation is given by

$$\hat{y}_{t/t-1} = Z_t \hat{a}_{t/t-1}, \quad f_{t/t-1} = Z_t P_{t/t-1} Z_t' \tag{21}$$

Once an additional observation, y_t, becomes available, the update equation revises the conditional moments with the new information in the latest observation.

$$\hat{a}_{t/t} = \hat{a}_{t/t-1} + k_t u_{t/t-1}, \quad P_{t|t} = \left(1 - k_t Z_t\right) P_{t|t-1} \tag{22}$$

where,

$$k_t = f_{t/t-1}^{-1} P_{t/t-1} Z_t', \quad u_{t/t-1} = y_t - Z_t \hat{a}_{t/t-1}$$

The quantity, k_t, is the gain of the KF and $u_{t/t-1}$ is the one-step-ahead error in predicting y_t with its conditional mean, $\hat{y}_{t/t-1}$. The estimator of the signal and its covariance matrix are given by

$$\hat{y}_{Signal,t/t} = z_t \hat{a}_{t/t}, \quad P_{Signal,t/t} = z_t P_{t/t} z_t' \tag{23}$$

To initialize the KF, it is necessary to specify starting values for the conditional moments. Those elements of the state vector that are stationary, i.e., survey error and the irregular, are initialized with their unconditional moments. The variances for the non-stationary and non-stochastic state variables are initialized with diffuse priors.

Filtering is tailored to real time processing of one observation at a time as it first becomes available. Each period the KF makes a prediction, $\hat{y}_{t/t-1}$ of the next observed value, y_t, using only the previous period estimates of the state vector, $\hat{a}_{t-1/t-1}$, and covariance matrix, $P_{t-1|t-1}$, and calculates updated estimates from the prediction error, $u_{t/t-1}$, thereby incorporating information from the latest available data at time t. Since the predicted values reflect all the past data up to time t -1, the corrected estimates for time t reflect all the available information from both the historical and current values of the series. After each prediction and update step, the prediction update process is repeated. The significant point is that the KF does no more work to process the last observation than it does for the first. The net result is an algorithm tailored to real-time applications, where data are continually coming in and information about the current value of the unobserved components is needed immediately.

Smoothing

The KF, however, is not well suited for producing historical estimates for a fixed set of data observations since it is designed to produce a current period estimate only and not to revise any earlier estimates. Observations following time t, $y_{t+1}, y_{t+2}, \ldots$, convey information about unobserved components at time t which can supplement the information available at time t.

The retrospective improvement of the KF estimates using ex post information is achieved by a process conveniently described as “smoothing”. This process revises each of the KF estimates for a period running from t = 1, to the last available observation at $t = n$. These “retrospective” estimate are obtained from the “Kalman Smoother,” which runs the KF

recursion backwards from t=n to t = 1 through the earlier data revising the estimates produced by filtering at each time point. Smoothing is batch processing in the sense that it operates on all of the data at once in contrast to the KF which processes one observation at a time.

A number of smoothing algorithms are available in the literature. We use a fixed interval smoother developed by DeJong(1988),

$$\begin{aligned} \hat{a}_{t/n} &= \hat{a}_{t/t-1} + P_{t/t} r_{t-1} \\ P_{t/n} &= P_{t/t}\left(I - N_{t-1} P_{t/t}\right) \end{aligned} \tag{24}$$

where

$$\begin{aligned} r_{t-1} &= Z_t' f_{t/t-1}^{-1} u_{t/t-1} + L_t r_t \\ N_{t-1} &= Z_t' f_{t/t-1}^{-1} Z_t + L_t' N_t L_t \\ L_t &= F_t\left(1 - K_t\right) Z_t \end{aligned}$$

for $t = n,\ldots,1$ initialized with $r_n = 0$ and $N_n = 0$.

Not surprisingly, the estimates from the smoother typically look "smoother" than those from the filter. This is because the variances of the smoothed estimates are never larger than the variances for the filtered estimates and usually much smaller towards the center of the series. But it is important to note that since these smoothed estimates use data from the entire sample, they do not correspond to estimates that would have been available to data users in real time. In practice, smoothing is done once a year. While smoothing could be performed each month, there is an obvious disadvantage to revising previous month estimates each month since it is likely to confuse data users.

Our model involves a large number of estimated parameters, many of which are estimated outside the time series model. This raises the question as to how errors in the parameters can be accounted for in the variances of the model estimates. Pfeffermann and Tiller, (2005a) develop a bootstrap method that is unbiased and computationally feasible for very complex models.

DIAGNOSTIC TESTING

A model should adequately represent the main features of movements in the CPS. An analysis of the model's prediction errors is the primary tool for assessing goodness of fit. This is an indirect test of the model. The actual model error is the difference between the true value of the signal and the model's estimate of that value. Since we do not observe the true values we cannot compute the actual model error. The overall model, however, provides an estimate of the signal and survey error, which sum to an estimate of the CPS. We may, therefore, use the model to predict new CPS observations.

The prediction errors are computed as the difference between the current values of the CPS and the predictions of the CPS made from the model, based on data prior to the current period. Since these errors represent movements not explained by the model, they should not contain any systematic information about the behavior of the signal or noise component of

the CPS. Specifically, the prediction errors, when standardized, should approximate a randomly distributed normal variate with zero mean and constant variance. The models are subjected to a battery of diagnostic tests to check the prediction errors for departure from these properties.

SEASONAL ADJUSTMENT

Data users are primarily interested in the underlying trend movements in the labor force series. The seasonal adjusted series is therefore an important target variable. The model directly produces an estimate of the non-seasonal component of the CPS free of survey error. To estimate the signal, the model in effect constructs a survey error filter, which attenuates the effect of survey error in the CPS and then removes seasonality with a conventional type of seasonal filter.

Conventional non-model based approaches to seasonal adjustment, such as X-12, ignore survey error and produce a trend, seasonal, and irregular decomposition that is very different from the classical decomposition. Much of the correlated survey error is absorbed into the trend, which produces spurious long run fluctuations. SE also tends to cause seasonal patterns to look less stable than they really are (Tiller, 1996).

BENCHMARKING

While the use of models produces estimators with much smaller variances than the survey estimates, it raises the question of how to protect against model breakdowns. The most dramatic type of breakdown occurs when there is an unexpected external shock occurring in real time that results in a large shift in the level of the series. Since this shift is unrelated to the historical past, the model will be slow to adapt to the new level. Monitoring prediction errors in real time is a common practice for detecting model breakdowns of this sort, but even when large prediction errors are detected, prior information about the nature of the outlier is rarely available. In these circumstances it is not possible to determine the appropriate outlier specification until additional data become available. Therefore, it is desirable to have a "built-in mechanism" to ensure the robustness of the estimators when the model fails to hold.

To provide protection against nationwide shocks, we constrain the sum of the State model estimates to equal the national CPS values. The justification for using the CPS national data is that the direct CPS estimators, which are unreliable in single States, can be trusted when aggregated over many States. The basic idea behind the use of the constraints is that if there is a nationwide shock that affects most States, the benchmarked estimators will reflect this change much faster than the model dependent estimators.

Benchmarking actually takes place in two stages. First, States are grouped into the 9 Census Divisions. The aggregate CPS division employment and unemployment series are modeled and then constrained to add up to the monthly National CPS estimates. These adjusted Division model estimates serve as benchmarks for constraining the sum of the State estimates to add to their respective adjusted Division estimates. In this way all of the State model estimates are constrained to sum to the National CPS estimates.

An approach under research for taking into account the errors in the benchmarks is discussed in Pfeffermann and Tiller (2005b).

METRO-AREA MODELS

Time series models are also developed for the following metropolitan areas,

Chicago-Naperville-Joliet, IL metropolitan division

Cleveland-Elyria-Mentor, OH metropolitan area

Detroit-Warren-Livonia, MI metropolitan area

Miami-Miami Beach-Kendall, FL metropolitan division

Seattle-Belvue-Everett, WA metropolitan division.

For each State with a metro model, a model is also developed for an aggregate balance of State area. The sum of the metro area and balance of State models estimates are forced to equal the corresponding benchmarked State estimates.

OPERATIONS OF THE SYSTEM

As part of the Federal-State cooperative Local Area Unemployment Statistics program, staffs in the 50 State and District of Columbia Employment Security agencies prepare their respective official monthly estimates using the State Time Series Analysis and Review System (STARS) software developed by BLS. A web-based interface allows State users to access this software on BLS servers to create, review, update and download labor force estimates.

During monthly processing, State users are queried for their latest UI and CES data which is combined with CPS data to produce model based estimates using the Kalman filter. At the end of the year, preliminary data are revised and the filtered estimates are revised with the smoothing algorithm and the smoothed State estimates are benchmarked to the monthly national CPS estimates.

REFERENCES

Bell, W.R. and Hillmer, S.C. (1990), ‘‘The Time Series Approach to Estimation for Repeated Surveys,’’ **Survey Methodology**, 16, 195-215.

Binder, D.A. and Dick, J.P. (1989),‘‘Modeling and Estimation for Repeated Surveys,’’ **Survey Methodology,** 15,29-45.

Box, G.E.P. and Jenkins, G.M. (1976), **Time Series Analysis Forecasting and Control**, San Francisco, CA, Holden-Day, Inc.

Bureau of Labor Statistics. **BLS Handbook of Methods**. www.bls.gov

DeJong, P. (1989), “Smoothing and interpolation with the State Space Model”, **Journal of the American Statistical Association**, 84, 1085-88

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

Harvey, A.C. and Hwang, C. (2000), “Estimating the Underlying Change in UK Unemployment,” (with discussion), **Journal of the Royal Statistical Society** (forthcoming)

Pfeffermann, D. (1992), ‘‘Estimation and Seasonal Adjustment of Population Means Using Data From Repeated Surveys,’’ **Journal of Business and Economic Statistics,** 9, 163-175.

Pfeffermann, D., Feder, M., and Signorelli, D. (1998), “Estimation of Autocorrelations of Survey Errors with Application to Trend Estimation in Small Areas”, **Journal of Business and Economic Statistics**, 16, 339-348.

Pfeffermann, D., Tiller, R., and Zimmerman, T. (2000), “Accounting for Sampling Error Autocorrelations Towards Signal Extraction from Models with Sampling Error”, **ASA Proceedings**, Business and Economics Statistic Section

Pfeffermann, D., and Tiller, R., (2005a), “Bootstrap Approximation to Prediction MSE for State-Space Models with Estimated Parameters”, **Journal of Time Series Analysis**, 26,6, 893-916.

Pfeffermann, D., and Tiller, R. (2005b), “Model-Based Seasonal Adjustment of Survey Series Subject to Benchmark Constraints with a State-Space Smoothing Algorithm”, **ASA Proceedings**, Business and Economics Statistic Section

Scott, A.J., and Smith, T.M.F. (1974), ‘‘Analysis of Repeated Surveys Using Time Series Methods,’’ **Journal of the American Statistical Association,** 69, 674-678.

Scott, A.J., Smith, T.M.F., and Jones, R.G. (1977), ‘‘The Application of Time Series Methods to the Analysis of Repeated Surveys,’’ **International Statistical Review**, 45, 13- 28.

Tiller, R. (1992), ‘‘An Application of Time Series Methods to Labor Force Estimation Using CPS Data,’’ **Journal Of Official Statistics,** 8, 149-166.

Tiller, R. (1996), “Time Series Decomposition of Periodic Survey Data with Autocorrelated Errors”, Invited paper presented at the Business and Economic Section of the American Statistical Association.

Zimmerman, T.S., and Robison, E. (1995), “Estimation of Autocorrelations for Current Population Survey Labor Force Characteristics,” 1995 ASA Proceedings Survey, 414-419