Journal of Official Statistics
Vol. 8. No. 2, 1992, pp. 149–166

Time Series Modeling of Sample Survey Data from the U.S. Current Population Survey

Richard B. Tiller[1]

Abstract: The signal extraction approach to repeated sample surveys is potentially an effective way to reduce high variances in conventional sample estimators arising from small sample sizes. A signal-plus-noise model of labor force estimates from the U.S. Current Population Survey is formulated as a structural time series model with explanatory variables where variance-covariance information from the survey sample is used to place restrictions on the time series model. This model is fit to a statewide series. Model-based estimates are compared to the observed sample data and the effect of controlling for sampling error is explored.

Key words: Small area estimation; correlated sampling error; signal extraction; Kalman filter.

1. Introduction

In the United States and elsewhere there has been a long-standing demand for government agencies to produce reliable economic statistics below the national level. Often times a large scale sample survey is designed to produce reliable statistics for major geographic areas but because of budget constraints is spread too thinly across the country to produce reliable area specific data. In the small area estimation literature various model-based approaches have been suggested to improve the efficiency of the direct survey estimator. Most of this literature focuses on the situation where data exist for a large number of areas but only one or a few observations are available per area. Gains in efficiency are sought through the use of cross-sectional models to pool data across areas.

Many of the more important surveys conducted by government agencies are repeated at frequent intervals to provide decision makers with up-to-date information on the dynamic behavior of the phenomena being measured. The existence of these time series of sample estimates raises the possibility of achieving large reductions in variance by pooling data over time for a given area using signal extraction techniques developed in the time series literature. Such an approach originated in the work of Scott and Smith (1974) and Scott, Smith, and Jones (1977). The innovative feature of their approach was to demonstrate that if the underlying population values are treated as stochastic rather than fixed an improved

[1] Bureau of Labor Statistics, 441 G Street, N.W., Washington, D.C. 20212, U.S.A.

Acknowledgement: The author thanks Art Dempster and Steve Miller for providing error covariances; Mike Welch, Al Tupek, the associate editors, and three referees for their helpful comments; Tom Evans for preparation of the text and tables. The views expressed in this paper are those of the author and do not necessarily represent the policies of the Bureau of Labor Statistics.

estimator can be obtained by combining a time series model of the population with designed-based sampling error information.

In recent years there has been renewed interest in the time series approach to survey sample data as a potentially cost effective way of reducing variance in these estimates, see, e.g., Bell and Hillmer (1987), Binder and Dick (1989), Pfeffermann (1989), and Tiller (1990). This paper applies this basic approach to statewide labor force data from the U.S. Current Population Survey (CPS). The CPS is a nationwide monthly sample of about 59,000 households designed to produce estimates of employment and unemployment and other characteristics of the labor force status of the population. While acceptable variance estimates of key labor force variables are produced for the nation as a whole, at the state level these same variables have much higher variability. A simplified version of the model to be presented here was implemented by the U.S. Bureau of Labor Statistics in 1989 in 39 states and the District of Columbia (Tiller 1989).

In Section 2, a signal-plus-noise model of the CPS data is formulated. Section 3 discusses signal extraction and estimation of unknown parameters; Section 4 describes an application to unemployment rate data from the CPS sample for the state of Massachusetts; Section 5 discusses further research; and Section 6 provides a summary of results.

2. Signal-Plus-Noise Model

The observed CPS labor force estimate, $y(t)$, is represented as the sum of two independent processes, the true population or signal, $\theta(t)$, and the sampling error or noise, $e(t)$

$$y(t) = \theta(t) + e(t). \tag{2.1}$$

Given a model for $\theta(t)$ and design-based information on the covariance structure of $e(t)$, the observed sample series may be decomposed into its signal and noise components. The basic approach of this paper is to represent the signal by a structural time series model with explanatory variables (Harvey 1989) and to represent the noise as an ARMA model (Bell and Hillmer 1990). Nonsampling errors are not dealt with in this application.

2.1. *The signal*

The signal is modeled as a time series decomposed into the form

$$\theta(t) = M(t) + T(t) + S(t) + I(t) \tag{2.2}$$

where the terms on the right-hand side denote the regressor, trend, seasonal, and irregular components of the signal at time t. The first three components are allowed to drift slowly over time by subjecting them to mutually independent white noise disturbances. The variances of these disturbances constitute the hyperparameters of the signal and determine the stochastic properties of the individual components. A positive variance for a component implies that it is a stochastic process, possibly nonstationary, while a zero variance implies deterministic behavior. The irregular is treated as stationary. These components are described in more detail below.

2.1.1. Regressor component

This component represents that part of the signal that can be explained by a set of observable economic variables, largely independent of the sampling error in the observed series

$$M(t) = x(t)\beta(t) \tag{2.3}$$

where $x(t)$ is a $1 \times k$ vector of the known explanatory variables and $\beta(t)$ a $k \times 1$ coefficient vector. The coefficients may be treated as either fixed or stochastic. In the

latter case $\beta(t)$ is modeled as a random walk where $\nu_\beta(t)$ is a vector of mutually independent random shifts

$$\beta(t) = \beta(t-1) + \nu_\beta(t)$$
$$E[\nu_\beta(t)\nu_\beta'(t)] = \text{Diag}\,(\sigma^2_{\beta_1}, \ldots, \sigma^2_{\beta_k}). \tag{2.4}$$

2.1.2. Trend component

This component is represented as a local approximation to a linear trend

$$T(t) = T(t-1) + R(t-1) + \nu_T(t)$$
$$R(t) = R(t-1) + \nu_R(t). \tag{2.5}$$

The trend level, $T(t)$, is shifted by the white noise variable, $\nu_T(t)$, and its first difference or growth rate is shifted by $\nu_R(t)$. The two disturbances are mutually independent with mean zero variances $\sigma^2_{v_T}$ and $\sigma^2_{v_R}$, respectively. A variety of common forms emerge as special cases. If $R(t) = 0$, the trend follows a simple random walk in levels. A fixed linear trend results if both variances are zero.

2.1.3. Seasonal component

The seasonal component is the sum of six trigonometric terms associated with the 12-month frequency and its five harmonics

$$S(t) = \sum_{j=1}^{6} S_j(t) \tag{2.6.a}$$

where each of the individual terms $\{S_j(t)\}$ is subject to a white noise shock, $\nu_{S_j}(t)$, assumed to have a common variance, σ^2_S

$$S_j(t) = \cos(\omega_j)S_j(t-1) + \sin(\omega_j)S_j^*(t-1) + \nu_{s_j}(t) \tag{2.6.b}$$

$$S_j^*(t) = -\sin(\omega_j)S_j(t-1) + \cos(\omega_j)S_j^*(t-1) + \nu^*_{s_j}(t)$$

$$\omega_j = \frac{\pi j}{6}. \tag{2.6.c}$$

Over a 12-month period the expected seasonal effects add to zero

$$E\left(\sum_{l=0}^{11} S_{t-l}\right) = 0.$$

A positive value for σ^2_S permits the seasonal pattern to evolve over time while a zero value results in a fixed seasonal pattern.

2.1.4. Irregular component

The irregular is a residual not explained by the regression or time series components. It is assumed to be stationary. Sometimes it is appropriate to treat it as white noise, although serial correlation may be allowed for by modeling it as an ARMA process. In practice, a low-order AR model is usually sufficient when $I(t)$ is not white noise

$$\alpha_I(L)I(t) = \nu_I(t)$$
$$E[\nu_I^2(t)] = \sigma_I^2 \tag{2.7}$$

where

$\alpha_I(L) = 1 - \alpha_{I,1}L - \cdots - \alpha_{I,p}L^p$ is a stationary operator.

While the general model of the signal, just described, is very flexible, it need not involve a large number of parameters to be estimated. Some of the components may drop out in practice. Often, the regressor variables will be able to explain a substantial amount of variation in the observed series with fixed coefficients. If the regressors were fully successful, the trend component would reduce to a fixed intercept and the seasonal component would drop out. In general, it is unlikely that the regression component will account for all of the systematic variation in the signal since its behavior is likely to be influenced by variables that are difficult to measure conceptually or too costly to collect. Stochastic time series components can be very effective in controlling for changes in the extent and influence of these unmeasured explanatory variables. Even if the regressors are only partially successful,

the trend and seasonal components need not take very complicated forms, although it is important to assume that the time series components are stochastic at the outset.

2.2. *Noise*

The noise component of the observed CPS estimate represents error that arises from sampling only a portion of the total population. Its structure depends upon the CPS design and population characteristics. For our purposes, we focus on those design features that are likely to have a major effect on the variance-covariance structure of $e(t)$.

One of the most important features of the CPS is the large overlap in sample units from month to month. The sample is divided into eight independent panels or rotation groups. Units are partially replaced each month according to a 4-8-4 rotating panel. When new households are introduced into the sample, they are included for four consecutive months, dropped out for eight months, and then returned for four months. Since this system provides large overlaps between samples one month and one year apart, we can expect $e(t)$ to be strongly autocorrelated. Also, there is likely to be some correlation between nonidentical units in the same rotation group because of the way in which new samples are generated. When a cluster of housing units permanently drops out of a rotation group, it is replaced by nearby units. Since the new units will have characteristics similar to those being replaced, this will result in correlations between nonidentical households in the same rotation group (Train, Cahoon, and Makens 1978).

Finally, the dynamics of the sampling error will also be affected by the composite estimator. This is a weighted average of an estimate based on the entire sample for the current month only and an estimate which is a sum of the prior month composite and change that occurred in the six rotation groups common to both months (Bureau of the Census 1978). In effect, this estimator takes a weighted average of sample data from the current and all previous months.

Another important feature of the CPS is its changing variance over time. This variance may be expressed in compact form as

$$\sigma_e^2 = D_y S_y^2 \tag{2.8}$$

where

D_y = ratio of the variance of the CPS estimator to the variance of the simple random sample estimator (design effect)

$S_y^2 = N^2(t)\theta(t)[1 - \theta(t)]/n(t)$

$n(t)$ = sample size

$N(t)$ = total population size.

The variance, S_y^2, is derived from the formula for a proportion.

Equation 2.8 illustrates three major sources of heteroscedasticity: (1) sample redesigns as reflected by changes in D_y; (2) changes in the sample size $n(t)$; and (3) changes in the true value of $\theta(t)$. The first two cause discrete shifts in the sample variance. For example, the CPS is redesigned each decade to make use of decennial census data to update the sampling frame and estimation procedures. Most recently, a state-based design was phased in during 1984/85 along with improved procedures for noninterviews, ratio adjustments, and compositing. Changes in state sample sizes have occurred more frequently than redesigns and have had a major effect on variances at the state level. Even with a fixed design and sample size, the error variance will be changing because it is a function of the size of the true labor force. Since the labor force is both highly cyclical and

seasonal, we can expect the variance to follow a similar pattern.

To capture the autocorrelated and heteroscedastic structure of $e(t)$, we may express it in multiplicative form (see Bell and Hillmer 1990) as

$$e(t) = \gamma(t)e^*(t) \qquad (2.9.a)$$

with $e^*(t)$ reflecting the autocovariance structure, assumed to follow an ARMA process and $\gamma(t)$ representing a changing variance over time. More explicitly

$$e^*(t) = \phi^{-1}(L)\theta(L)v_e(t)$$

$$\gamma(t) = \frac{\sigma_e(t)}{\sigma_e^*} \qquad (2.9.b)$$

where

$\theta(L)$ = a stationary moving-average operator of order q_e

$\phi(L)$ = a stationary autoregressive operator of order p_e

$$\sigma_{e^*}^2 = \sigma_{v_e}^2 \sum_{k=0}^{\infty} g_k.$$

The weights $\{g_k\}$ are computed from the generating function

$$g(L) = \phi^{-1}(L)\theta(L).$$

The autocovariance structure may also change over time with redesigns of the sample. However, since the most important source of autocorrelation is the 4-8-4 rotation scheme, which has not changed, it seems reasonable to treat this structure as stable, at least, between sample designs.

3. Signal Extraction and Estimation

This section briefly describes how signal extraction and estimation of the unknown parameters are performed. For more details, consult Tiller (1990). For estimation and signal extraction the component signal and noise models are put into state-space form. Given the parameters of the system, the Kalman filter (KF) is then used to optimally decompose the current sample observation into its signal and noise components. The structure of the noise process is given by survey design information and the unknown hyperparameters of the signal process are estimated by maximum likelihood.

A general state-space model is defined in terms of two equations: a transition equation (3.1) that describes the behavior of the state vector, consisting of the unobserved components of the signal and noise, as a first-order vector autoregressive process and an observation equation (3.2) that relates the observed data to the state vector. Mathematically

$$Z_t = FZ_{t-1} + V_t \qquad (3.1)$$

$$y_t = H_t Z_t \qquad (3.2)$$

where Z is the state vector, F a fixed transition matrix, V a vector containing the white noise disturbances of the model, and H a vector that converts the unobserved components of the state vector to the observed sample data.

The problem is to find the mean vector, given the observed sample values, denoted by

$$E[Z_t | Y_t] = Z_{t|t} \qquad (3.3)$$

and the covarance matrix

$$COV(Z_t | Y_t) = P_{t|t} \qquad (3.4)$$

where E denotes the expectation operator, and Y_t is a vector of current and past values of y_t.

The solution, due to Kalman (1960), takes the form of a set of updating equations to calculate $Z_{t|t}$ and $P_{t|t}$ recursively from $Z_{t|t-1}$ and $P_{t|t-1}$ by using the current observation y_t. The resulting estimator has the minimum-mean-square error property and is optimal if V_t is normal. The estimator of the signal, obtained as a linear combination of

the elements of the state vector that are associated with the signal, has the following recursive form

$$E(\theta_t|Y_t) = E(\theta_t|Y_{t-1}) + h_t[y_t - E(y_t|Y_{t-1})]. \quad (3.5)$$

A corresponding expression exists for the sampling error component

$$E(e_t|Y_t) = E(e_t|Y_{t-1}) + (1 - h_t) \times [y_t - E(y_t|Y_{t-1})]. \quad (3.6)$$

The first term on the right of the update equation for θ is the model prediction of the signal, given sample data up to $t - 1$, to which is added a portion, h_t, of the error in predicting the observed sample estimate at time t. This has a simple interpretation as a composite-type estimator that combines a model estimate based on past data with current sample information to obtain an improved estimate.

The quantity h_t, which varies between zero and one, determines how much weight is placed on the current sample estimate. It is a function of the ratio of the variance in the signal to the sampling error variance. As discussed by Bell and Hillmer (1990), this illustrates an important characteristic of the time series approach: it provides a design-consistent estimator, in the sense that full weight is given to the sample estimate as its variance goes to zero. While this is a reassuring property, the greatest potential gains from a model-based approach come when the sampling error variance is large. Nevertheless, this consistency property has important practical implications even for relatively small sample sizes, as will be illustrated below.

From an implementation point of view, the structure of the KF is particularly convenient for the preparation of monthly labor force estimates. Since it is a recursive data processing algorithm, it does not require all previous data to be kept in storage and reprocessed every time a new sample observation becomes available. All that is required is an estimate of the state vector and its covariance matrix for the previous month. However, estimates prior to the current period are not updated as new sample data become available.

The suboptimality of previous period estimates is easily remedied through a process called smoothing. This process can be described conceptually as combining a KF running forward from initial time to terminal time and a separate filter running backward from terminal time to initial time (Maybeck 1979). Smoothing, in contrast to filtering, requires that the entire data series be processed in batch. In the actual implementation, monthly estimates for the current year are produced using the KF and revised at the end of the year, along with previous years, with a smoothing algorithm.

In practice, knowledge of the underlying models comprising the signal and noise is incomplete since neither are observable. However, since $y(t)$ comes from a survey, the covariance structure of $e(t)$ is known, or at least can be estimated independently of the signal with conventional design-based procedures. Holding the noise component fixed, the signal may be estimated using standard model fitting and diagnostic techniques.

The parameters of the signal, the variances of the white noise disturbances and the coefficients of the irregular component are estimated by maximum likelihood. The innovation form of the likelihood is formed (Harvey 1989) and maximized with respect to the unknown parameters using a quasi-Newton approach as implemented in the IMSL subroutine, DUMINF (IMSL 1987).

4. A State Unemployment Rate Example

This section describes an application of the signal-plus-noise model to an unemployment rate series collected from the Massachusetts CPS sample. Specifically, we discuss the methods used to estimate the sampling error structure, the specification of the signal, the results of fitting the complete model to the state data, and an assessment of the importance of directly modeling the sampling error.

4.1. Modeling the signal

In modeling the unemployment rate series at the state level, the following three explanatory variables were chosen for inclusion:

i. UI claims rate: The number of unemployed workers claiming unemployment insurance (UI) benefits as a percent of total nonagricultural employment.
ii. EP ratio: Total nonagricultural payroll employment as a percent of the population.
iii. Entrant rate: The number of unemployed entrants into the labor force as a percent of the labor force for the nation as a whole.

The first two variables are state specific and are developed from non-CPS data sources. The claims data are an administrative by-product of the federal-state UI program. The nonagricultural employment data come from the Current Employment Statistics (CES) program, a payroll survey of employers. While the entrant rate is a CPS statistic, it is taken from the entire national sample. Therefore its sampling error can be treated as largely independent of a given state's unemployment rate.

The rationale for including the above variables is discussed by Tiller (1989). A brief summary is provided here. The UI claims rate reflects those unemployed workers who have passed their state's requirements for benefit eligibility. For various reasons, the claims data do not fully reflect the cyclical behavior of job losers. To partially control for this, the EP ratio is included as a general measure of labor market tightness. In addition, unemployment due to labor force entry, which can account for as much as 40% of the total unemployed, is not accounted for at all in the UI statistics. The national entrant rate is included to account for the distinctive behavior of labor force entrants. Figures 1–4 present graphs of the unemployment rate and the three explanatory variables. The first three figures refer to data specific to Massachusetts and Figure 4 contains national data.

Since the choice of the explanatory variables is constrained by the availability and the limitations of the data, there is no guarantee that they will account for all the variation in the signal. For example, the payroll employment variable, while highly correlated with the CPS household survey estimates, is known to have some important seasonal and cyclical differences. Moreover, the national entrant rate may not fully reflect the seasonal behavior of entrants in a specific state. For these reasons, trend and seasonal components were added to the model.

4.2. Modeling the noise

The application of the signal-plus-noise approach requires information on the variance-covariance structure of the sampling error. The most obvious approach is to estimate this structure directly from the sample unit data using the sample design information. Scott, Smith, and Jones (1977), hereafter SSJ, referred to this approach as a primary analysis. It has the advantage of providing efficient estimators that impose few restrictions on the error covariances.

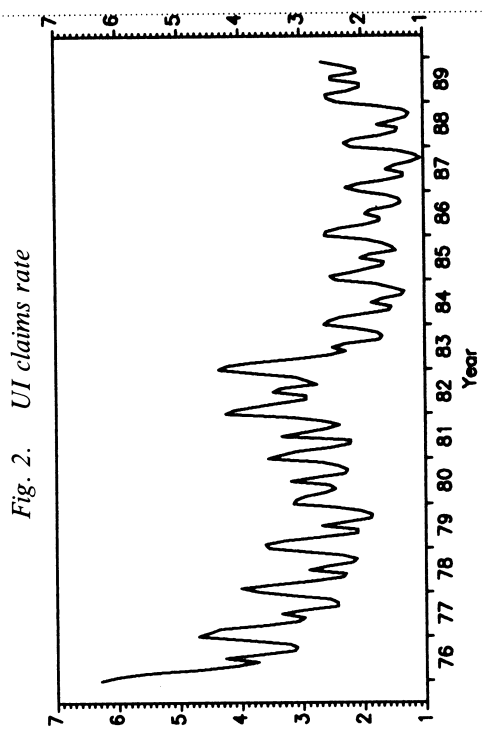

Fig. 1. CPS unemployment rate

Fig. 2. UI claims rate

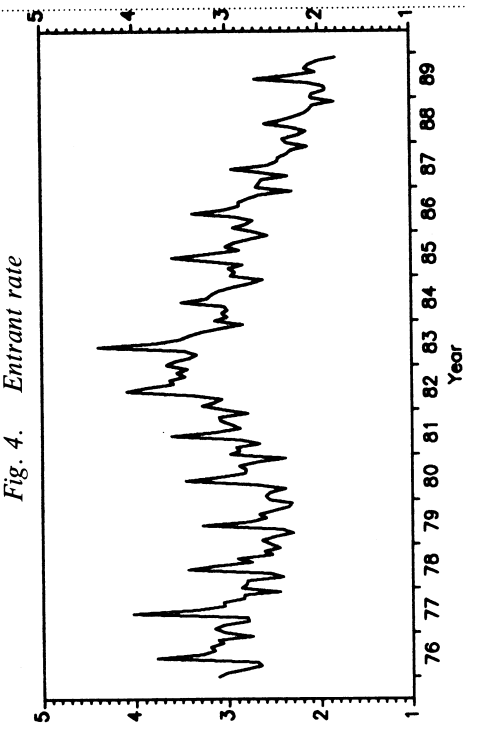

Fig. 3. EP ratio

Fig. 4. Entrant rate

In practice, a primary analysis is seldom feasible in large scale surveys, where variance estimation involves complex computations on huge microdata files. In these circumstances, SSJ suggest a secondary analysis in which the error structure is modeled directly from the aggregate data. While this avoids the cost of a primary analysis, it does require more assumptions.

Limited research has been conducted using either approach for CPS data. Hausman and Watson (1985) developed an ARMA (1, 15) model of the error process for the national teenage unemployment rate series through a secondary analysis. Bell and Hillmer (1987) using teenage data, but for a different time period, developed an ARMA(1, 1) model as an approximation to the design-based autocovariances estimated by Train, Cahoon, and Makens (1978).

The approach followed in this study uses data more aggregated than sample unit data, but does so in a way that does not require strong assumptions on the error structure. The methods used to estimate variances and autocorrelations are discussed below.

4.2.1. Variance estimates

Because of high costs of computation, the Census Bureau directly estimates variances once every ten years for selected characteristics at the national level only. 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 microdata. This curve is then generalized over time and to other statistics not used in the fit but with similar design effects. The form of the GVF is

$$V_y^2 = a + \frac{b}{y(t)} \tag{4.1}$$

where V_y^2 is the relvariance of the estimate $y(t)$ and a and b are empirically determined parameters (Bureau of the Census 1978).

At the state level, the variance parameters are not directly computed. Instead, these parameters were developed indirectly from the following assumed relationship between the b parameter and certain sample quantities

$$b = \begin{cases} k \dfrac{N}{n} D_y, & \text{after 1986} \\ \dfrac{N}{n} D_y + k', & \text{prior to 1987} \end{cases} \tag{4.2}$$

$$a = -\frac{b}{n}$$

where

D_y = design effect for the y statistic

N = total population size

n = sample size

k, k' = adjustments for between PSU variance.

The above equations follow from representing the CPS variance as the product of a simple random sample variance and a design effect (Wolter 1985, ch. 5).

The GVF for the unemployment rate, $y(t)$, may be derived by making use of the approximation for the relvariance of the ratio of two statistics x and y (Hansen, Hurwitz, and Madow 1953, p. 576),

$$V_{x/y}^2 = V_x^2 - V_y^2. \tag{4.3}$$

It follows that the estimate of the sampling error variance for the unemployment rate is given by

$$\sigma_e^2(t) = y(t)^2 \left[a + \frac{b}{[y(t) * CLF(t)]/100} - a' - \frac{b'}{CLF(t)} \right] \tag{4.4}$$

where

$y(t)$ = CPS unemployment rate (in percent)

a,b = GVF parameters for y

$CLF(t)$ = CPS civilian labor force

and

a', b' = GVF parameters for CPS employment.

While these variance estimates have been developed indirectly, they do reflect important known changes in the sample designs at the state level. Moreover, recent research supports their accuracy. Using the method of generalized replication for 1987 state data, Lent (1991) concluded that the method described above yields good results for unemployment.

4.2.2. Autocovariance estimates

In principle, autocovariances can be directly computed using the same design-based techniques as for variances. As with the case for the variance estimation, this is a very costly process and has only rarely been done even at the national level. Much of what is known empirically about the CPS covariance structure is based upon a study by Train, Cahoon, and Makens (1978). Using the Keyfitz paired difference method, autocovariances were estimated for national level statistics, both composited and uncomposited. This study, however, was limited to 13 months of data, December 1974 through December 1975. Bell and Hillmer (1987) used results from this study to model national teenage unemployment. Although these covariance estimates cover a different geographical level, time period, and sample design, it is nonetheless useful to compare them to state estimates.

This study draws upon autocovariances, specific to a state, developed from preliminary work that uses state-level time series data for the eight rotation groups, previously described in Section 2.2. Each of these groups may be treated as independent subsamples. Variability across subsamples, when averaged over time, provides the basis for estimating the error covariances.

State data from January 1981 to July 1989 (103 observations) were used. To control for rotation group bias, mean differentials by time in sample were subtracted out. Eight time series of errors were constructed from deviations of each group's adjusted estimate about the overall mean. Assuming each of the error series is stationary, autocovariances were averaged across time and across groups. These estimates were not adjusted for compositing.

Table 1 presents the autocorrelation estimates. The second column shows the sample overlap that arises from the 4-8-4 rotation schedule. The next two columns give the autocorrelations from the national study (Train, Cahoon, and Makens 1978) for the composited and uncomposited unemployment statistics. Estimates from the state study are presented in the fifth column.

The state estimates show some strong similarities with the national estimates. The autocorrelations are strongest at the first three lags and decline sharply from lags four to eight, where there is no overlap of housing units. Even with no overlap, there is still some dependency between nonidentical units in the same rotation group since they were selected from the same neighborhood. The autocorrelations begin to rise at the higher lags where the samples overlap again. The state estimates show a peak at the 12-month lag which corresponds to a local peak in the sample overlap. The national estimates do not show this peak, but this may be due to the fact that the 12-month lag correlation was estimated

Table 1. Sampling error autocorrelations

Lag	% Overlap of identical housing units	National CPS		State CPS	State Model
		Composited	Uncomp.	Uncomposited	
1	75.0	0.50	0.45	0.39	0.39
2	50.0	0.33	0.28	0.30	0.30
3	25.0	0.23	0.17	0.22	0.22
4	0.0	0.17	0.08	0.07	0.07
5	0.0	0.12	0.07	0.09	0.09
6	0.0	0.07	0.05	0.08	0.08
7	0.0	0.07	0.05	0.07	0.07
8	0.0	0.09	0.08	0.05	0.05
9	12.5	0.09	0.10	0.07	0.07
10	25.0	0.12	0.14	0.09	0.09
11	37.5	0.09	0.09	0.08	0.08
12	50.0	0.07	0.11	0.13	0.13
13	37.5	–	–	0.04	0.04
14	25.0	–	–	−0.01	0.01
15	12.5	–	–	−0.01	0.00
16	0.0	–	–	−0.01	0.00
17	0.0	–	–	−0.03	0.00

from only one observation. Finally, we note that the negative autocorrelations from lag 14 and up are a reflection of low reliability in these estimates.

Given the state autocorrelations, the next step is to develop an ARMA approximation. As Table 1 indicates, there are certain features of the CPS design that suggest some complexity in the ARMA representation. The peak at the 12-month lag implies that a model with high-order lags will be necessary to pickup the autocorrelation due to the rotating panel. An ARMA (1, 12) model was specified, resulting in estimated parameters which exactly reproduce the autocorrelations up to lag 13. (See column six of Table 1.) Alternatively, a more parsimonious model might have been developed by minimizing a sum of squares function which could have been helpful if the correlations failed to dampen out quickly after lag 12.

4.3. Estimation results

This section presents the results of applying the signal-plus-noise model to monthly statewide CPS unemployment rate data covering the period from January 1976 to December 1989 (168 observations). To assess the importance of modeling the noise component, an alternative model was estimated that did not explicitly take it into account.

Part A of Table 2 presents the specification and parameter estimates for the basic unemployment rate model with and without accounting for the CPS error structure. Identical regressor variables were used in each case with fixed coefficients since the variance of their white noise disturbances were estimated to be very close to zero. Accounting for sampling error does affect the values of the coefficients but not by a

Table 2. Parameter estimates and test diagnostics

	Ignoring sampling error	*With sampling error*
A. Parameter estimates		
Regression coefficients (abs. t-values)		
UI claims rate	0.592 (6.8)	0.610 (7.1)
EP	−0.314 (6.1)	−0.286 (6.2)
Entrant rate	1.207 (11.2)	0.987 (8.9)
Time series components		
Trend level ($\sigma^2_{v_T}$)	0.20	0.013
Seasonal ($\sigma^2_{v_S}$)	0.412×10^{-3}	0.337×10^{-3}
Irregular variance ($\sigma^2_{v_I}$)	0.224	0
Irregular coefficient ($\alpha_{I,1}$)	0.358	–
Likelihood	−143	−111
B. Diagnostics		
Test statistics		
Ljung-Box [12]	8.51	9.07
Ljung-Box [24]	18.40	14.33
Heteroscedasticity w/time	*3.55	1.17
Bera-Jarque normality	*10.01	2.59
Skewness	−0.14	0.32
Excess kurtosis	1.20	0.15
Post-sample prediction	0.27	0.45
Post-sample bias	0.04	0.08

*significant at the 5% level

substantial amount. Binder and Dick (1989) reported similar results in a related study.

Both models have a trend level that follows a simple random walk, a stochastic growth rate not being necessary with the presence of regression variables. Also, both models have a stochastic seasonal component of the same general form. When sampling error is accounted for, the variance of the irregular component goes to zero and it drops out of the model. When sampling error is ignored, it is necessary to include a first-order autoregressive term to account for residual autocorrelation.

Part B of Table 2 presents the results of diagnostic testing performed on the errors in the one-step-ahead predictions of the actual CPS estimates generated from the KF. Conditional on the parameters, these prediction errors should behave as normally distributed white noise variables. For a discussion of the individual tests, see Harvey (1989). Examination of the test results gives no reason to question the adequacy of the model when the CPS error structure is explicitly accounted for. If the CPS error is ignored, one might expect the prediction errors to be both autocorrelated and heteroscedastic. In fact, Table 2 indicates the presence of heteroscedasticity and non-normality in the prediction errors. A time series analyst unfamiliar with the CPS may be tempted to transform the data in an attempt to stabilize the variance. Frequently, a power transformation is used but this is not likely to be very helpful since both the CPS variance and coefficient of variation change over time, sometimes independently

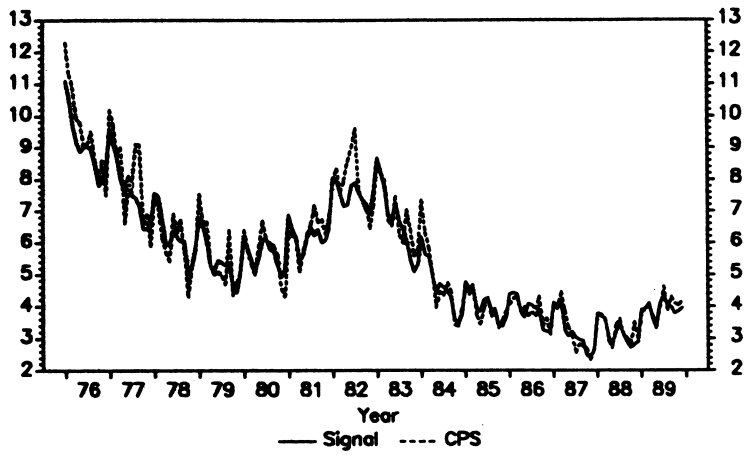

Fig. 5. Unemployment rates

of the population variance. This point will be discussed in further detail below. That there is no evidence of autocorrelation when sampling error is ignored is not surprising since conventional time series modeling is flexible enough to absorb the autocorrelated portion of the error into the irregular and possibly into the seasonal component as well. Of course, confounding the source of the autocorrelation could lead to inappropriate inferences about the behavior of the time series.

Figure 5 compares the CPS to the smoothed signal from the model that accounts for sampling error. Plots of the regressor, seasonal, and sampling error components appear in Figures 6–8. The signal is considerably smoother than the CPS. This is highlighted in Table 3 which shows the decomposition of the variance of change in the CPS over selected time spans. Elimination of the sampling error from the CPS by signal extraction removes about 46% of the variation at the one-month span.

The smoothed estimates of sampling error are plotted in Figure 8. This series represents the difference between the CPS and the smoothed signal taking into account the changing variance and autocorrelation structure of the sample design. Prior to 1985 large differences are occasionally evident. In particular, for July 1982, the CPS estimate is 9.6% compared with 7.6% for the signal. The recursive structure of the KF provides a useful diagnostic for examining the plausibility of such a large difference by generating one-step ahead predictions of the actual sample estimates each month conditional on the model. Since these are true predictions, made prior to incorporating information from the current sample values into the estimation process, they provide a way of assessing how compatible the model is with the observed sample estimates. For July 1982 the model-based prediction of the CPS value was 9.0% with a standard deviation of 0.71. Thus, the large difference between the sample and the signal is compatible with the model.

There is another instance in which the model did not predict the CPS well. This occurred in September 1977 which shows up as a large positive spike in the sampling error in Figure 8. For this observation the prediction error was 3.5 times its standard error. However, the CPS observation has the appearance of an additive outlier, being unusually high at 9.1% and then falling to

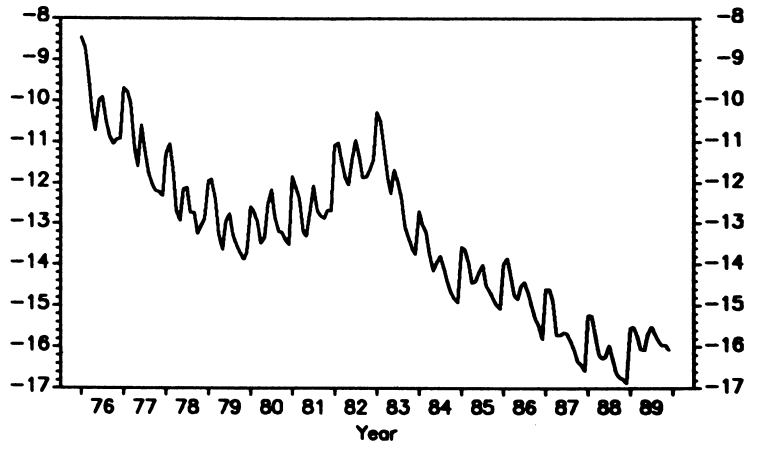

Fig. 6. Regressor component of the signal

6.6% the next month and remaining at a lower level thereafter. Accordingly, it is not surprising the model was unable to predict this observation and it seems reasonable to assign a large portion of this prediction error to the sampling error component.

Figure 9 plots the GVF standard errors for the CPS (dashed line) and the standard errors for the smoothed signal accounting for sampling error (solid line) and ignoring sampling error (dotted line). The CPS standard error shows a considerable amount of variation, with a peak of about 0.7 percentage points in the recession years of the early 1980s and dropping to around 0.4 percentage points in recent years. While a declining unemployment rate accounted for part of this drop, the most important factor was a 62% expansion in the number of assigned households for the state during 1984–85.

Looking at the behavior of the standard error for the smoothed signal estimated from the model accounting for CPS error, we see that it has been considerably below the CPS, averaging about 50% and has shown much less variability. However, the

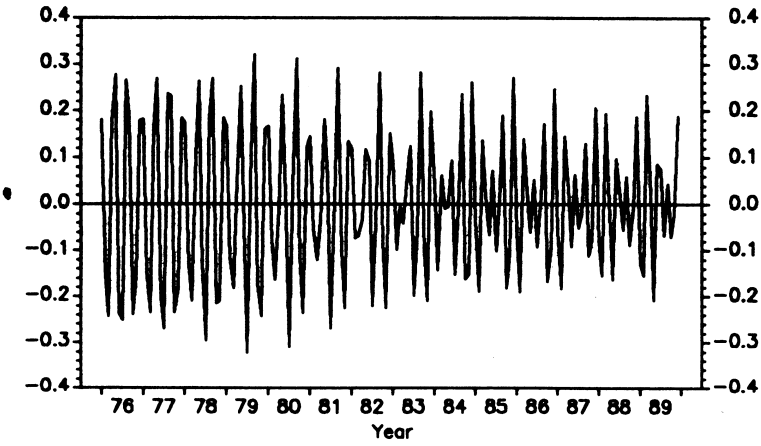

Fig. 7. Seasonal component of the signal

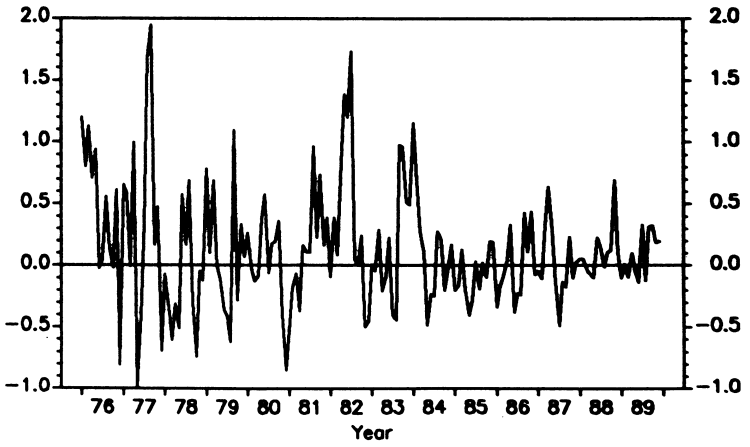

Fig. 8. CPS sampling error

size of the CPS standard error relative to the signal has clearly declined primarily due to the sample expansion.

The direct effect of sample expansion on the model estimates may be illustrated by the behavior of the weight given an individual CPS observation in the KF update of the signal estimate (see equation 3.5). Again, focusing on the model that includes sampling error, Figure 10 shows that these weights (solid line) increased about 40% or so since 1984. Putting more weight on more precise sample estimates is a reflection of the design-consistency property of the estimator.

When a model ignoring sampling error is used to estimate the regression, trend, and seasonal components of the signal, major inefficiencies occur. As can be seen from Figure 9, the standard error of the smoothed signal (dotted line) is almost constant except at the end points. It lies below the signal estimated from the model accounting for sampling error prior to sample expansion and above afterwards. Turning to Figure 10, we see that estimating the signal from a model ignoring sampling error (dotted line) produces a very stable weighting pattern for the individual CPS observations. The model overweights the CPS in the early years and underweights it in the later years.

While the signal extraction approach appears to result in substantial gains over the sample estimator, certain limitations must be kept in mind. The model-based variances do not account for uncertainty in the estimated signal parameters. Also, the sampling error structure is estimated outside of the time series model and is treated as if it were known. Finally, the model of the signal is only an approximation, and hence subject to misspecification bias.

Table 3. Contribution of components to variance in observed CPS series

Span in months	Sampling error	Signal	Regressors	Trend	Seasonal
	Percent				
1	46.3	53.7	41.3	0.3	12.2
3	29.9	70.1	66.5	0.7	2.8
12	29.0	71.0	64.5	6.4	0.1

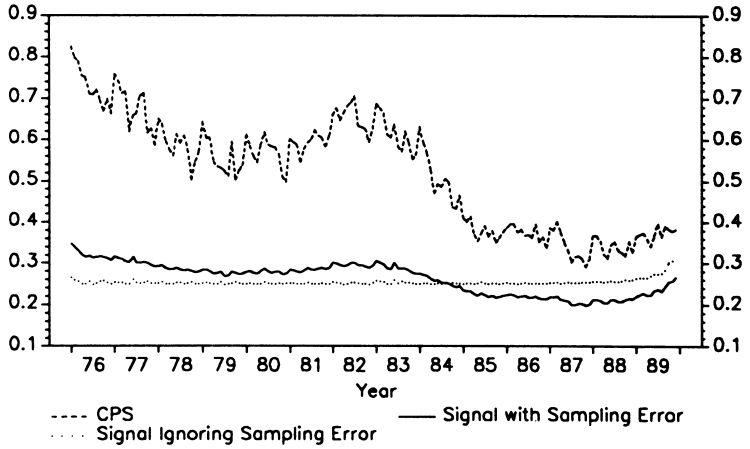

Fig. 9. Standard errors

5. Further Research

Work continues on developing sampling error variance and covariance estimation derived directly from sample unit data (Lent 1991) to provide additional information on the autocorrelation effects of the CPS sample overlaps. The development of GVFs for variance estimation based directly on state level sample designs is also a possibility that will be explored.

Dempster and Hwang (1991) have developed a variance component model for estimating the sample autocorrelation structure from the CPS rotation groups. Alternative ways of fitting ARMA models to these correlations will be tested. Also, additional work is planned to assess model-based estimates of variance including accounting for uncertainty in estimated parameters and testing sensitivity to alternative model specifications. Finally, the basic model structure can be further expanded to explicitly account for the effects of outliers and other types of data irregularities as well as for more general types of intervention effects.

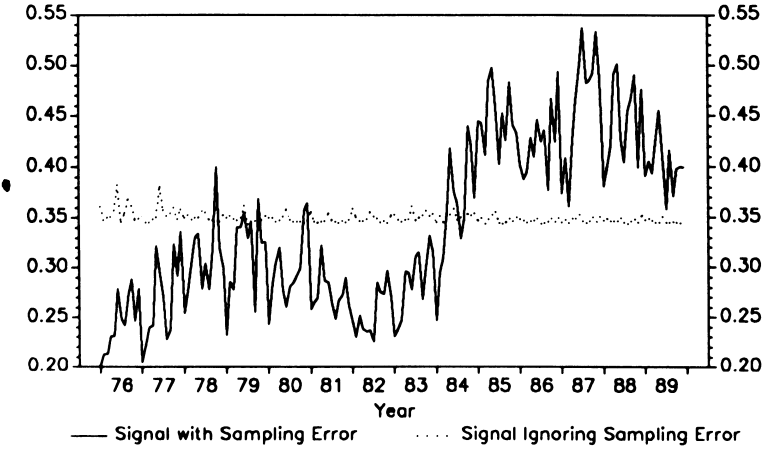

Fig. 10. Weight on CPS in signal update

6. Conclusions

A signal-plus-noise model was formulated and fit to a state CPS unemployment rate series. The signal was represented by a structural time series model with explanatory variables. The noise, or sampling error, was modeled by an ARMA process with changing variance. The estimator derived from this model is design consistent. A simplifed form of this basic model was implemented in 1989 for 39 states and the District of Columbia.

To assess the practical importance of explicitly accounting for sampling error, a model of the signal was fit to the CPS as if it were the true series. Compared to the complete signal-plus-noise model, this reduced model is not design consistent and this resulted in an inability to reflect major changes in sample reliability. The time series model, including the sampling error component, achieved on average a 50% reduction in variance over the survey estimator. However, the exact magnitude of this gain must be treated with caution. Specifically, more work is needed on developing state-specific sampling error variances and autocorrelations, testing sensitivity to alternative specifications of the signal, and quantifying the uncertainty in the estimated model parameters.

7. References

Bell, W.R. and Hillmer, S.C. (1987). Time Series Methods for Survey Estimation. U.S. Bureau of the Census Statistical Research Division Report Series, #CENSUS/SRD/RR-87/20.

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). Modelling and Estimation for Repeated Surveys. Survey Methodology, 15, 29–45.

Bureau of the Census (1978). The Current Population Survey: Design and Methodology. Technical Paper 40, Washington, D.C.: Author.

Dempster, A.P. and Hwang, J.S. (1991). A Sampling Error Model of Statewide Labor Force Estimates from the CPS. Paper prepared for the U.S. Bureau of Labor Statistics, Washington, DC.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, vol. I. New York: John Wiley.

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

Hausman, J. and Watson, M. (1985). Errors in Variables and Seasonal Adjustment Procedures. Journal of the American Statistical Association, 80, 531–540.

IMSL (1987). Math/Library User's Manual, ver. 1.0. Houston: Author.

Kalman, R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering, Transactions ASME, Series D, 82, 35–45.

Lent, J. (1991). Variance Estimation for Current Population Small Area Labor Force Estimates. In Proceedings of the Survey Research Methods Section, American Statistical Association, forthcoming.

Maybeck, P.S. (1979). Stochastic Models, Estimation, and Control, vol. 2. Orlando: Academic Press.

Pfeffermann, D. (1989). Estimation and Seasonal Adjustment of Population Means Using Data from Repeated Surveys. In Proceedings of the Survey Research Methods Section, American Statistical Association, 1–10.

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. (1989). A Kalman Filter Approach to Labor Force Estimation Using Survey Data. In Proceedings of the Survey Research Methods Section, American Statistical Association, 16–25.

Tiller, R. (1990). An Application of Time Series Methods to Labor Force Estimation Using CPS Data. In Proceedings of the Survey Research Methods Section, American Statistical Association, 284–289.

Train, G., Cahoon, L., and Makens, P. (1978). The Current Population Survey Variances, Inter-Relationships, and Design Effects. In Proceedings of the Survey Research Methods Section, American Statistical Association, 443–448.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Received November 1990
Revised April 1992