An Application of Time Series Methods to Labor Force Estimation Using CPS Data

Richard Tiller, U.S. Bureau of Labor Statistics
441 G Street NW, Washington, DC 20212

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 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: Kalman filter, correlated sampling error

1. INTRODUCTION

The application of signal extraction techniques to time series produced by sample surveys has attracted the interest of researchers in statistical agencies as a potentially cost effective way of improving the reliability of these estimates, see, e.g., Bell and Hillmer (1987), Binder and Dick (1989), Pfeffermann (1989), and Tiller (1989). This approach originated in the work of Scott and Smith (1974) and Scott, Smith and Jones (1977) which demonstrated that the efficiency of survey estimators could be improved upon by recognizing that the population series is itself stochastic.

This paper applies this basic approach to statewide labor force data from the Current Population Survey (CPS). The CPS is a nationwide monthly sample of about 59,000 households designed to produce estimates of the labor force status of the population. While low variance estimates of key labor force variables are produced for the nation as a whole, at the state level these same variables are subject to much higher variance. In section 2 of this paper, a signal plus noise model of the observed CPS data is formulated. Section 3 casts this model in state-space form and uses the Kalman filter (KF) to derive an optimal estimator of the underlying signal. Section 4 discusses the method used to estimate the unknown parameters. Section 5 applies the model to labor force data from a CPS state sample.

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 value (signal), $\theta(t)$, and the sampling error (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 1987).

2.1 The Signal

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

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

These components are described in more detail below.

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,

$$Mx(t) = X(t)\,\beta(t) \tag{2.3}$$

where X(t) is a 1 x k vector of the variables with a k x 1 coefficient vector, $\beta(t)$. The coefficients may be treated as either fixed or stochastic. In the latter case, $\beta(t)$ is modeled as a random walk where $v_\beta(t)$ is a vector of mutually independent random shifts,

$$\beta(t) = \beta(t-1) + v_\beta(t) \tag{2.4}$$

$$E\,(v_\beta(t)\, v'_\beta(t)) = \mathrm{Diag}\,(\sigma^2_{\beta_1}, \cdots, \sigma^2_{\beta_k}).$$

2. Trend Component:

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

$$T(t) = T(t-1) + R(t-1) + v_T(t) \tag{2.5}$$

$$R(t) = R(t-1) + v_R(t).$$

The two white noise disturbances, $v_T(t)$ and $v_R(t)$, are mutually independent with 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.

3. Seasonal Component:

The seasonal component is the sum of six trigonometric terms associated with the 12 month frequency and its 5 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, $v_{S_j}(t)$, assumed to have a common variance, σ^2_S,

$$S_j(t) = \cos \omega_j S_j(t-1) + \sin \omega_j S^*_j(t-1) + v_{S_j}(t) \quad (2.6.b)$$

$$S^*_j(t) = -\sin \omega_j S_j(t-1) + \cos \omega_j S^*_j(t-1) + v^*_{S_j}(t) \quad (2.6.c)$$

$$\omega_j = \pi j / 6.$$

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.

4. Irregular Component:

The irregular is a residual not explained by the regression or time series components. It is assumed to follow a stationary AR process,

$$\alpha_I(L)\, I(t) = v_I(t),\ E(v^2_I(t)) = \sigma^2_I \quad (2.6.d)$$

where,

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

2.2 Noise

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

One of the most important features of the CPS is the use of a 4-8-4 rotating panel (Bureau of the Census 1978). Since this system provides large overlaps between samples one month and one year apart, we can expect e(t) to be strongly autocorrelated. Moreover, when a cluster of housing units permanently drops out of a rotation group, it is replaced by nearby units, resulting in correlations between nonidentical households in the same rotation group (Train, Cahoon and Makens 1978).

Finally, the dynamics of the sample error will also be affected by the composite estimator. This is a weighted average of the current sample estimate and an estimate of change that occurred in the 6 rotation groups common to both months (Bureau of the Census 1978).

Another important feature of the CPS is its changing variance due to sample redesigns, changes in sample sizes, and changes in the population values. To capture the autocorrelated and heteroscedastic structure of e(t), we may express it in multiplicative form (see Bell and Hillmer 1989) as

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

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

$$e^*(t) = \phi_e(L)\, \alpha^{-1}_e(L)\, v_e(t) \quad (2.9.b)$$

$$\gamma(t) = \sqrt{\sigma^2_e(t) / \sigma^2_{e^*}} \quad (2.9.c)$$

where,

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

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

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

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

$$g(L) = \phi_e(L)\, \alpha^{-1}_e(L).$$

3. STATE-SPACE SYSTEM AND THE KALMAN FILTER

For estimation and signal extraction, the component signal and noise models are put in state-space form. The signal and noise are the state variables, Z(t), whose evolution over time is described by the transition equation,

$$Z(t) = F\, Z(t-1) + G\, v(t) \quad (3.1)$$

with covariance matrix,

$$E(v(t)\, v'(t)) = Q$$

and the state variables are transformed into the observed sample series, y(t), by the observation equation

$$y(t) = H(t)\, Z(t) \quad (3.2)$$

where the system matrices F, G, Q and H(t) are m x m, m x ℓ, ℓ x ℓ, and 1 x m.

As shown in the complete paper in equations 3.3 - 3.7, each component model has its own state-space form which can be combined to form the overall model.

$$Z(t)_{(mx1)} = \left[Z_\beta(t), Z_T(t), Z_S(t), Z_I(t), Z_{e^*}(t) \right]' \quad (3.8)$$

$$F_{(mxm)} = \text{block diagonal } (F_\beta, F_T, F_S, F_I, F_{e^*})$$

$$G_{(m x \ell)} = \begin{bmatrix} G' & | & 0 \\ & | & \\ - - & | & - - \\ & | & \\ 0 & | & G_{e^*} \\ & | & (n x 1) \end{bmatrix},$$

$$G' = \text{block diagonal } (G_\beta, G_T, G_S, G_I)$$

$$H(t)_{(1xm)} = [\, H_\beta(t), H_T, H_S, H_I, \gamma(t)\, H_{e^*} \,]$$

$$Q_{(\ell x \ell)} = \text{block diagonal } (Q_\beta, Q_T, Q_S, Q_I, Q_{e^*}).$$

The signal and noise components are given by

$$\theta(t) = H_\theta(t)\, Z(t),\ e(t) = H_e(t)\, Z(t) \quad (3.9)$$

where,

$$H_\theta(t) = [\, X(t), H_T, H_S, H_I, 0_{m-n} \,]$$

$$H_e(t) = [\, 0_{m-n}, \gamma(t), 0_{n-1} \,].$$

The KF produces an estimate of the signal which is

optimal with respect to the model assumptions and from the point of view of survey sampling is also a design-consistent estimator. Assume the system matrices for a particular model of the observations, F, G, H and Q are given and that the disturbance, v_t, and initial state vector, Z_0, are multivariate normal and independent of each other. It follows that the distributions of Z(t) and y(t) given sample values up to t-1 are themselves normal,

$$(Z(t)/y(t\text{-}1)) \sim N\,(Z(t/t\text{-}1),\ P(t/t\text{-}1)) \qquad (3.10.a)$$

$$(y(t)/y(t\text{-}1)) \sim N\,(y(t/t\text{-}1),\ f(t/t\text{-}1)). \qquad (3.10.b)$$

Their means and variances are given by the prediction equations of the KF,

$$Z(t/t\text{-}1) = F\,Z(t\text{-}1/t\text{-}1) \qquad (3.11.a)$$

$$P(t/t\text{-}1) = F\,P(t\text{-}1/t\text{-}1)\,F' + G\,Q\,G' \qquad (3.11.b)$$

$$y(t/t\text{-}1) = H(t)\,Z(t/t\text{-}1) \qquad (3.11.c)$$

$$f(t/t\text{-}1) = H(t)\,P(t/t\text{-}1)\,H(t)'. \qquad (3.11.d)$$

Upon observing y(t), the posterior means and variances of the conditional normal distributions are given by the KF update equations,

$$Z(t/t) = Z(t/t\text{-}1) + K(t)\,\tilde{y}(t) \qquad (3.12.a)$$

$$P(t/t) = (1 - K(t)\,H(t)')\,P(t\text{-}1/t\text{-}1) \qquad (3.12.b)$$

$$y(t/t) = y(t) \qquad (3.12.c)$$

$$f(t/t) = 0 \qquad (3.12.d)$$

where,

$$K(t) = P(t/t\text{-}1)\,H(t)'\,/\,f(t/t\text{-}1)$$

$$\tilde{y}(t) = y(t) - H(t)\,Z(t/t\text{-}1).$$

Given the model assumptions, each component of the estimator Z(t/t) has minimum mean square error.

The posterior means and variances of the signal and the noise are given by,

$$\theta(t/t) = \theta(t/t\text{-}1) + h(t)\,\tilde{y}(t) \qquad (3.13.a)$$

$$e(t/t) = e(t/t\text{-}1) + (1 - h(t))\,\tilde{y}(t) \qquad (3.13.b)$$

$$\text{Var}\,\theta(t/t) = H_\theta(t)\,P(t/t)\,H'_\theta(t) \qquad (3.13.c)$$

$$\text{Var}\,e(t/t) = H_e(t)\,P(t/t)\,H'_e(t). \qquad (3.13.d)$$

The estimator of the signal, $\theta(t/t)$, being a linear combination of minimum mean square error components, has itself, the minimum mean square error property.

The weight h(t) varies between 0 and 1, the closer it is to 1, the more $\theta(t/t)$ is shrunk towards the sample estimate y(t). The equation below indicates what governs this shrinkage:

$$h(t) = \frac{\text{Var}[\theta(t)/\theta(t\text{-}1)] + A}{\sigma_e^2(t) + B + \text{Var}[\theta(t)/\theta(t\text{-}1)] + A} \qquad (3.14)$$

where,

$$\text{Var}\,[\theta(t)/\theta(t\text{-}1)] =$$

$$\sum_{i=1}^{k} X_i^2(t)\,\sigma^2_{v_{\beta_i}} + \sigma^2_{v_T} + \sigma^2_{v_R} + \sum_{j=1}^{6} \sigma^2_{S_j} + \sigma^2_{v_I}$$

$$A = \text{Var}\,[H_\theta F\,Z(t\text{-}1/t\text{-}1)] = H_\theta\,F\,P(t\text{-}1/t\text{-}1)\,F'\,H'_\theta$$

$$\sigma_e^2(t) = \gamma^2(t)\,\sigma^2_{v_{e*}}$$

$$B = \text{Var}\,[H_e F\,Z(t\text{-}1/t\text{-}1)] = H_e\,F\,P(t\text{-}1/t\text{-}1)\,F'\,H'_e$$

The amount by which $\theta(t/t)$ is adjusted toward y(t) depends upon the size of the model-based variance of the signal, Var $[\theta(t)/\theta(t\text{-}1)]$, relative to the sampling error variance, $\sigma_e^2(t)$. As $\sigma_e^2(t) \longrightarrow 0$, $h(t) \longrightarrow 1$.

Therefore, $\theta(t/t)$ is design consistent.

The KF produces the minimum MSE estimator of the state vector, Z(t), based on all sample data through time t, in a recursive manner. Smoothing produces minimum MSE estimates for each point in time. The basic type of smoothing (fixed interval) used in this paper is described by Maybeck (1979).

4. ESTIMATION OF THE SIGNAL PARAMETERS

Given knowledge of the parameters of the noise model, we can estimate the unknown signal parameters by maximum likelihood. Assuming the disturbance vector v(t) is multivariate normal, the innovation form of the likelihood function is computed via the KF. The parameter space must then be searched to locate the maximum value of the likelihood. The quasi-Newton approach, discussed by Dennis and Schnabel (1983, pp. 111-29) and implemented in the IMSL subroutine, DUMINF (IMSL 1987), was used to maximize the likelihood.

5.0 A STATE UNEMPLOYMENT RATE EXAMPLE

In this section, we describe an application of the signal plus noise model to an unemployment rate series collected from a state CPS sample, covering the period from January 1976 to December 1989.

5.1 Modeling the Signal

In modeling the unemployment rate series at the state level, three explanatory variables are available for inclusion. These are:

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 rationale for including the above variables is discussed by Tiller (1989). Stochastic trend and seasonal components are added to account for residual variation in the signal.

5.2 Modeling the Sampling Error

To model e(t), we must first develop design-based estimates of its variance and autocovariances. Once this is done, the autocovariance structure is approximated by an ARMA

model whose coefficients are used as the parameters of the noise component of the state-space model.

1. Variance Estimates:

To assess the reliability of national CPS statistics, the Census Bureau (1968) uses the method of generalized variance functions (GVF). This approach fits variance curves to groups of statistics for which variance estimates have been directly computed. For state level statistics, variance estimates were not directly computed. The parameters of the GVF 's were developed indirectly, as discussed by the author in the complete paper.

2. Autocovariance Estimates

In principle, autocovariances can be directly computed using the same design-based techniques as for variances. In fact, this has never been done for state level CPS statistics and only rarely done at the national level. This study draws upon autocovariances, specific to a state, developed from preliminary work by Art Dempster and Steve Miller. Their approach exploits the availability of state level time series data for the eight CPS rotation groups. 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.

Given the state autocorrelations, the next step is to develop their ARMA approximations. An ARMA (1, 12) model was fit to the autocorrelations. For more details, consult the author's complete paper.

5.3 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 the table 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 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 irregular term to account for residual autocorrelation.

Part B of the table presents the results of diagnostic testing performed on the standardized innovations generated from the Kalman filter. Conditional on the parameters, these innovations should behave as normally distributed white noise variables. For a discussion of the individual tests, see Harvey (1989). Examination of the test results give 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 innovations to be both autocorrelated and heteroscedastic. In fact, the table indicates the presence of heteroscedasticity and non-normality in the innovations. 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 1 compares the smoothed signal from the model that accounts for sampling error with the CPS. The signal is considerably smoother than the CPS. Elimination of the sampling error from the CPS by signal extraction removes about 46 percent of the variance of month-to-month change.

Figure 2a plots the GVF standard errors for the CPS (black line) and the standard errors for the smoothed signal accounting for sampling error (grey line) and ignoring sampling error (dashed line). The CPS standard error shows a considerable amount of variation, rising to a peak of about .7 percentage points in the recession years of the early 1980s and dropping to around .4 percentage points in recent years. While a declining unemployment rate accounted for part of this drop, the most important factor was a 62 percent 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 percent less and has shown much less variability. However, there has been a clear upward trend in the ratio of the signal to the CPS standard error (figure 2b) primarily due to the sample expansion.

The direct impact of sample expansion on the model estimates may be illustrated by the behavior of the weight given an individual CPS observation in the Kalman filter update of the signal estimate (see equation 3.14). Again, focusing on the model that includes sampling error, figure 2c shows that these weights increased about 40 percent 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 2a, the standard error of the smoothed signal (dashed line) is almost constant except at the end points. It lies below the signal estimated from the model accounting for sample error prior to

sample expansion and above afterwards. Turning to figure 2c we see that estimating the signal from a model ignoring sampling error (dashed 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, there are reasons to believe these gains are overstated. The model-based variances do not account for uncertainty in the estimated signal and noise parameters and the model of the signal is only an approximation, and hence is subject to misspecification bias.

SUMMARY

A signal plus noise model was formulated and fit to a state CPS unemployment rate series. While it appears that the model-based approach produced substantial gains over the CPS sample estimator, a considerable amount of additional work is necessary before any firm conclusions can be drawn.

Acknowledgements

The author thanks Art Dempster and Steve Miller for providing error covariance estimates and 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.

REFERENCES

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

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.

Dennis, J.E., and Schnabel, R.B. (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Englewood Cliffs, NJ: Prentice-Hall.

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

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

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," paper presented at the annual meeting of the American Statistical Association.

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," paper presented at the annual meeting of the American Statistical Association.

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*, pp. 443-448.

Table

Parameter Estimates and Test Diagnostics

A. Parameter Estimates

	Ignoring Sampling Error	With Sampling Error
Regression Coefficients (t-values)		
UI claims rate	.592(6.8)	.610(7.1)
EP	-.314(-6.1)	-.286(-6.2)
Entrant rate	1.207(11.2)	.987(8.9)
Time Series Components		
Trend		
Level ($\sigma^2_{v_T}$)	.020	.013
Seasonal ($\sigma^2_{v_S}$)	$.412 \times 10^{-3}$	$.337 \times 10^{-3}$
Irregular		
Variance ($\sigma^2_{v_I}$)	.224	0
Coefficient ($\alpha_{I,1}$)	.358	--
Likelihood	-143	-111

B. Diagnostics

Test Statistics	Ignoring Sampling Error	With Sampling Error
Ljung-Box [-12]	8.51	9.07
Ljung-Box [-24]	18.40	14.33
Heteroscedasticity w/Time	3.55*	1.17
Bera-Jarque Normality Test	10.01*	2.59
Skewness	-.14	.32
Excess kurtosis	1.20	.15
Post-Sample Pred. Var. Test	.27	.45
Post-Sample Bias Test	.04	.08

*significant at the 5% level

Figure 1
Unemployment Rate, CPS and Signal
CPS = Black line
Signal= Grey line

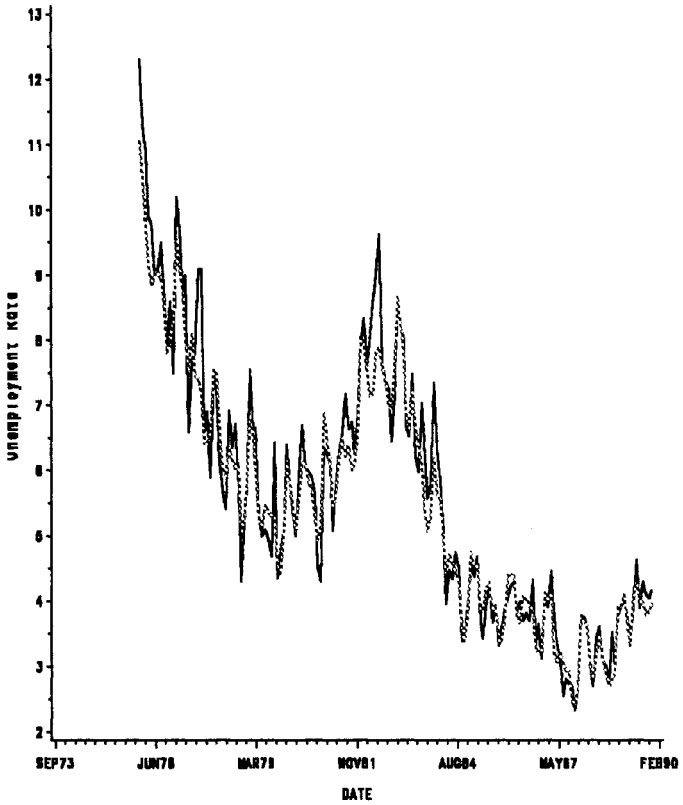

Figure 2a
Standard Errors, CPS and Signal
CPS Unemployment Rate = Black line
Signal with sampling error = Grey line
Signal ignoring sampling error = Dashed line

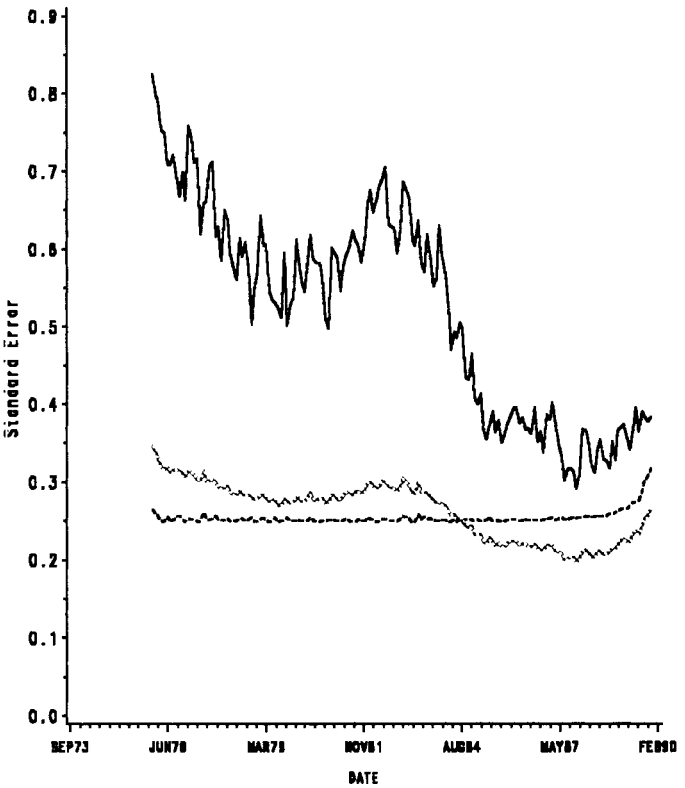

Figure 2b
Ratio of Standard Errors, Signal to CPS
(Model includes sampling error)

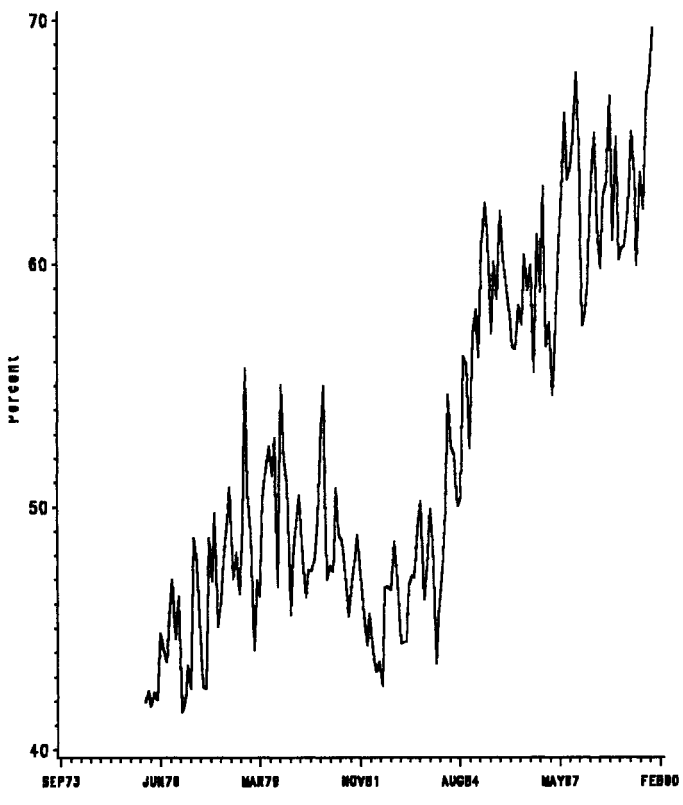

Figure 2c
Weight on CPS in Signal Update
Including sampling error = Solid line
Ignoring sampling error = Dashed line