TIME SERIES METHODS FOR SURVEY ESTIMATION

William Bell, Bureau of the Census and Steven Hillmer, University of Kansas[1]
William Bell, Bureau of the Census, Washington, D.C. 20233

I. Introduction

Papers by Scott and Smith (1974) and Scott, Smith, and Jones (1977), hereafter SSJ, suggested the use of signal extraction results from time series analysis to improve estimates in periodic surveys. If the covariance structure of the usual survey estimators (Y_t) and their sampling errors (e_t) for a set of time points is known, these results produce the linear functions of the available Y_t's that have minimum mean squared error as estimators of the population values being estimated (say θ_t) for θ_t a stochastic time series. To apply these results in practice one estimates a time series model for the observed series Y_t and estimates the covariance structure of e_t over time using knowledge of the survey design.

Section 2 of this paper gives a brief overview of the basic results and framework for this approach. Section 3 gives some theoretical results and section 4 some application considerations for the approach. In section 5 we illustrate the approach with an example.

2. Basic Ideas of the Approach

The basic idea in using time series techniques in survey estimation that distinguishes it from the classical approach is the recognition of two sources of variability. Classical survey estimation deals with the variability due to sampling -- having not observed all the units in the population. Time series analysis deals with variability arising from the fact that a time series is not perfectly predictable (often linearly) from past data. Consider the decomposition:

$$Y_t = \theta_t + e_t \tag{2.1}$$

where Y_t is a survey estimate at time t, θ_t is the population quantity of interest at time t, and e_t is the sampling error. The sampling variability of e_t is the focus of the classical survey sampling approach, which regards the θ_t's as fixed. From a time series perspective all three of Y_t, θ_t, and e_t can exhibit time series variation, as long as they are random and not perfectly predictable from past data. Standard time series analysis would treat Y_t directly and ignore the decomposition (2.1); thus the sampling variation of e_t is not treated explicity, it is only handled indirectly in the aggregate Y_t. In fact, time series analysts typically behave as if the sampling variation is not present and the true values θ_t are actually observed.

2.1 Basic Results

Suppose that estimates Y_t are available at a set of time points labelled t = 1,...,T. Let $\underset{\sim}{Y} = (Y_1,\ldots,Y_T)'$ and similarly define $\underset{\sim}{\theta}$ and $\underset{\sim}{e}$ so we have $\underset{\sim}{Y} = \underset{\sim}{\theta} + \underset{\sim}{e}$. It would be usual to assume the estimates Y_t are unbiased and that θ_t and e_t are uncorrelated so that

$$\begin{aligned} E(\underset{\sim}{Y}) = E(\underset{\sim}{\theta}) &= \underset{\sim}{\mu} = (\mu_1,\ldots,\mu_T)' \\ \Sigma_Y &= \Sigma_\theta + \Sigma_e . \end{aligned} \tag{2.2}$$

Here $\underset{\sim}{\mu}$ and Σ_θ refer to the time series structure of θ_t, which is not subject to sampling variation. In this case it is well known that the minimum mean squared error linear predictor of θ_t for t = 1,...,T is given by

$$\begin{aligned} \hat{\underset{\sim}{\theta}} &= \underset{\sim}{\mu} + \mathrm{Cov}(\underset{\sim}{\theta},\underset{\sim}{Y})\mathrm{Var}(\underset{\sim}{Y})^{-1}(\underset{\sim}{Y}-\underset{\sim}{\mu}) \\ &= \underset{\sim}{\mu} + \Sigma_\theta \Sigma_Y^{-1}(\underset{\sim}{Y}-\underset{\sim}{\mu}) \end{aligned} \tag{2.3}$$

Using (2.2) this can be reexpressed as

$$\hat{\underset{\sim}{\theta}} = \underset{\sim}{\mu} + (I - \Sigma_e \Sigma_Y^{-1})(\underset{\sim}{Y} - \underset{\sim}{\mu}) \tag{2.4}$$

$$= \underset{\sim}{\mu} + (I + \Sigma_e \Sigma_\theta^{-1})^{-1}(\underset{\sim}{Y} - \underset{\sim}{\mu}) \tag{2.5}$$

Another standard result is that the variance of the error of this estimate is

$$\mathrm{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) = \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e \tag{2.6}$$

Under normality (2.3) - (2.5) give $E(\underset{\sim}{\theta}|\underset{\sim}{Y})$, the conditional expectation of $\underset{\sim}{\theta}$ given $\underset{\sim}{Y}$, and (2.6) gives $\mathrm{Var}(\underset{\sim}{\theta}|\underset{\sim}{Y})$, the conditional variance. R. G. Jones (1980) gives the results (2.4) - (2.6) assuming $\underset{\sim}{\mu} = \underset{\sim}{0}$ (or equivalently assuming means have been subtracted). Scott and Smith (1974) and SSJ give equivalent results using classical time series signal extraction techniques which we shall consider later.

Notice that (2.3) - (2.6) require knowledge

of $\underset{\sim}{\mu}$ and any two of Σ_Y, Σ_θ, and Σ_e (the third can be obtained from (2.2)). In practice these will not be known exactly and will need to be estimated. The basic assumption underlying the application of the preceeding results, which we shall call the time series approach to survey estimation, is that $\underset{\sim}{\mu}$ and Σ_Y can be estimated from the time series data on Y_t (typically using some sort of time series model) and Σ_e can be estimated using survey microdata and knowledge of the survey design. We will discuss aspects of this in section 4.

2.2 Contrast with minimum variance unbiased and composite estimation

It is important to understand the distinction between the time series approach to estimation and the approach known as Minimum Variance Linear Unbiased Estimation (MVLU). Smith (1978), R. G. Jones (1980), and Binder and Dick (1986) review and discuss the MVLU approach. While both the MVLU and time series approaches use data from time points other than t in estimating θ_t, they differ in that MVLU regards the θ_t's as fixed and still only treats one source of variation, that due to sampling. It was developed for cases (such as many rotating panel surveys) where more than one direct estimate of θ_t is available for each t and the e_t's are correlated over time due to overlap in the survey design. The use of Y_j for $j \neq t$ in estimating θ_t then comes from generalized least squares results and the e_t's correlation. These remarks also apply to composite estimation (Rao and Graham 1964, Wolter 1979), which can be thought of as an approximation to MVLU.

2.3 The Time Series Approach as a Unifying Framework for Related Problems

There are other problems in repeated surveys besides estimation where typically only one of the two sources of variability is recognized. The general framework provided offers chances for improved results in these other problems, as well as potentially unifying them as subproblems under one general approach. Problems where typically time series variation is recognized and sampling variation is ignored include time series and econometric modeling and forecasting, seasonal adjustment (exceptions are Hausman and Watson (1985) and Wolter and Monsour (1981)), and trend estimation. Problems where typically sampling variation is recognized and time series variation is ignored include detection of statistically significant changes over time (see Smith (1978)), preliminary estimation in repeated surveys (an exception is Rao, Srinath, and Quenneville (1986)), and benchmarking (an exception is Hillmer and Trabelsi (1986)).

3. Theoretical Results for the Time Series Approach

In this section we give some theoretical results for the time series approach. Proofs are omitted. These and a more detailed exposition of the results and underlying assumptions are contained in a technical paper available from the authors.

3.1 Uncorrelatedness of θ_t and e_t

Standard time series signal extraction results (to be given in section 4.3 and corresponding to (2.3) - (2.6) given earlier) typically make the following three assumptions:

(1) θ_t, or a suitable difference of it, is stationary.

(2) e_t is stationary

(3) θ_t and e_t are uncorrelated with each other at all leads and lags.

For our purposes here a time series is stationary if its mean, variance, and lagged covariances do not depend on time. Assumptions (1) and (2) are probably reasonable in many situations, and ways of dealing with certain types of nonstationarity will be discussed in section 4. Here we focus on the assumption that θ_t and e_t are uncorrelated time series, meaning $\mathrm{Cov}(\theta_t, e_j) = 0$ for all time points t and j (equivalent to independence under normality). Previous papers on the time series approach to survey estimation have merely assumed this, but since θ_t and e_t depend on the same population units it is not obvious that this assumption is valid. Fortunately, we can establish the following result, which is valid under very mild conditions.

Result 3.1: Y_t design unbiased for all $t \Rightarrow$ θ_t, e_t uncorrelated time series.

In many cases we will want to take logarithms of Y_t to help induce stationarity of θ_t and the sampling errors. In such cases we write (2.1) as

$$(3.1) \quad Y_t = \theta_t + e_t = \theta_t(1 + \tilde{u}_t) = \theta_t u_t$$

where $\tilde{u}_t = e_t/\theta_t$ and $u_t = 1 + \tilde{u}_t$. Taking logs

$$(3.2) \quad \ln(Y_t) = \ln(\theta_t) + \ln(u_t)$$

We assume that $E(\ln(u_t)|\Omega_t) \approx \ln(E(u_t|\Omega_t))$, where Ω_t is the collection of values for the population units at time t. Also then $E(\ln(u_t)) \approx \ln(E(u_t))$. We can then show:

Result 3.2: Y_t design unbiased for all $t \Rightarrow \ln(\theta_t)$, $\ln(u_t)$ approximately uncorrelated time series.

We can alternatively obtain this result using the approximation $\text{Corr}(\ln(\theta_t),\ln(u_j)) \approx \text{Corr}(\theta_t,u_j)$.

3.2 Consistency of Time Series Estimates

Following Fuller and Isaki (1981) we let Y_t^ℓ (from the ℓ^{th} sample at time t) be a sequence of estimators of the characteristic θ_t^ℓ of the ℓ^{th} population at time t (Ω_t^ℓ) where the populations and samples for $\ell = 1,2,\ldots$ are nested. (See their paper for details.) Define $\underset{\sim}{Y}^\ell$, $\underset{\sim}{\theta}^\ell$, $\underset{\sim}{e}^\ell$, $\Sigma_{\underset{\sim}{y}}^\ell$, $\Sigma_{\underset{\sim}{\theta}}^\ell$, $\Sigma_{\underset{\sim}{e}}^\ell$, $\hat{\underset{\sim}{\theta}}^\ell$, and $\hat{\theta}_t^\ell$ in the obvious fashion. We have

Result 3.3: $Y_t^\ell \to \theta_t^\ell$ in mean square as $\ell \to \infty$ for t=1,...,T implies $\hat{\theta}_t^\ell \to \theta_t^\ell$ in mean square as $\ell \to \infty$ for t=1,...,T.

Convergence in probability is a more familiar concept in survey sampling. We have

Result 3.4: If $Y_t^\ell \to \theta_t^\ell$ in probability for t=1,...,T and there exist random variables ζ_t with finite variance such that $|Y_t^\ell - \theta_t^\ell| \le \zeta_t$ (almost surely) uniformly in ℓ, then $\hat{\theta}_t^\ell \to \theta_t^\ell$ in probability for t=1,...,T.

What these consistency results show is that if the errors in the original estimates Y_t of θ_t are small (Σ_e is small) then the errors $\hat{\theta}_t - \theta_t$ will be small as well. This is because $\hat{\underset{\sim}{\theta}} - \underset{\sim}{Y}$ becomes small as Σ_e becomes small, thus when there is little error in the original estimates Y_t the time series approach will not change them much. Binder and Dick (1986) have noted this phenomenon, and also pointed out that in this case it does not matter what time series model is used. That is, the convergence depends only on $\Sigma_e^\ell \to 0$ and not on $\underset{\sim}{\mu}$ or Σ_θ. Thus, the consistency results extend to allowing $\underset{\sim}{\mu}$, Σ_θ, and also Σ_e^ℓ to be replaced by estimates $\hat{\underset{\sim}{\mu}}^\ell$, $\hat{\Sigma}_\theta^\ell$, and $\hat{\Sigma}_e^\ell$, as long as $\hat{\underset{\sim}{\mu}}^\ell$ and $\hat{\Sigma}_\theta^\ell$ converge to something as $\ell \to \infty$ (it doesn't matter what as long as the limit of $\hat{\Sigma}_\theta^\ell$ is positive definite), and $\hat{\Sigma}_e^\ell \to 0$ (which should generally hold when $\Sigma_e^\ell \to 0$). Estimation of model parameters is not an issue in regard to these consistency results. While it is reassuring to know that the time series estimates behave sensibly in the situation of small error in the original estimates, the gains from the time series approach (see (2.6)) will come in the opposite case -- when $\text{Var}(e_t)$ is large.

We can extend the consistency results to the case where we take logarithms and estimate $\ln(\theta_t)$ in (3.2). In this case let $\Sigma_u^\ell = \text{Var}(\ln(\underset{\sim}{u}^\ell))$ where $\underset{\sim}{u}^\ell = (u_1^\ell,\ldots,u_T^\ell)'$ is from the ℓ^{th} population. Let $\underset{\sim}{\mu}$ and Σ_θ refer to $\ln(\underset{\sim}{\theta})$, and $\Sigma_Y^\ell = \Sigma_\theta + \Sigma_u^\ell$ refer to $\ln(\underset{\sim}{Y}^\ell)$. Analogous to (2.4) our estimate is

$$(3.3)\quad \ln(\hat{\underset{\sim}{\theta}}^\ell) = \underset{\sim}{\mu} + [I - \Sigma_u^\ell(\Sigma_Y^\ell)^{-1}]\,(\ln(\underset{\sim}{Y}^\ell) - \underset{\sim}{\mu}).$$

Result 3.5: $Y_t^\ell \to \theta_t^\ell$ in mean square for t=1,...,T implies $\ln(Y_t^\ell) \to \ln(\theta_t^\ell)$ and $\ln(\hat{\theta}_t^\ell) \to \ln(\theta_t^\ell)$ in mean square for t=1,...,T.

As before we could get a convergence in probability result by imposing a boundedness condition on the $\ln(u_t^\ell)$. Having $\ln(\hat{\theta}_t)$ as an estimate of $\ln(\theta_t)$, we might wish to take $\exp[\ln(\hat{\theta}_t)]$ as an estimate of θ_t. We have the following Corollary to Result 3.5.

Corollary: $Y_t^\ell \to \theta_t^\ell$ in mean square as $\ell \to \infty$ for t=1,...,T implies (see (3.3)) $\exp[\ln(\hat{\theta}_t^\ell)] \to \theta_t^\ell$ in probability as $\ell \to \infty$ for t=1,...,T.

4. Application Considerations

Application of the time series approach to survey estimation requires (1) estimation of the sampling error covariances, $\text{Cov}(e_t,e_j)$, in Σ_e, (2) estimation of the mean ($\underset{\sim}{\mu}$) and covariance structure of θ_t or Y_t (Σ_θ or Σ_Y), generally through some sort of time series model,

and (3) computation of the estimates $\hat{\theta}_t$ from the formulas of section 2 or something else equivalent. In this section we make a few remarks on these aspects of implementation.

4.1 Estimation of Sampling Error Covariances

In principle, estimation of sampling error covariances, $Cov(e_t,e_j)$, is the same problem as estimation of sampling variances, $Var(e_t)$, which is routinely done for many periodic surveys and for which many methods are available (Wolter 1985). In practice, there may be difficulties in linking survey microdata over time to do this. SSJ refer to direct estimation of sampling error covariances using survey microdata as a primary analysis. If this cannot be done it may still be possible to estimate Σ_e using only the time series data on Y_t by making some assumptions about e_t and θ_t. SSJ refer to such procedures as a secondary analysis. They give examples of both types of analysis. However, there is a fundamental identification problem with doing a secondary analysis. Without an independent estimate of Σ_e all we really know about Σ_θ and Σ_e is that they sum to Σ_Y. Thus, for any Σ_θ and Σ_e such that $\Sigma_Y = \Sigma_\theta + \Sigma_e$ let $\Sigma_{\theta'} = \Sigma_\theta - V$ and $\Sigma_{e'} = \Sigma_e + V$ for some symmetric matrix V such that $\Sigma_{\theta'}$ and $\Sigma_{e'}$ are positive semidefinite. Then we can also write $\Sigma_Y = \Sigma_{\theta'} + \Sigma_{e'}$. Use of $\Sigma_{\theta'}$ and $\Sigma_{e'}$ will result not in the estimation of θ_t, but in the estimation of a time series θ'_t with covariance structure given by $\Sigma_{\theta'}$. Analogous results have been obtained for time series models in other contexts; Tiao and Hillmer (1978) consider the simple example of e_t uncorrelated over time, and Bell and Hillmer (1984) discuss the well-known identifiability problem in seasonal adjustment. Knowledge of the survey design may suggest assumptions about e_t that will help to narrow the range of choices for the decomposition. Still this issue should be considered for any particular example where a secondary analysis is contemplated because of the possibility of unverifiable assumptions having a profound effect on the results.

If a full primary analysis can be conducted this will yield a direct estimate of Σ_e. This imposes no constraints on the covariance structure of e_t other than Σ_e be symmetric and positive definite. In many cases it may be reasonable to assume e_t is covariance stationary or (see below) relative covariance stationary. If this can be assumed this suggests pooling information over time to estimate $Cov(e_t,e_{t+k})$, which is the same for all t and depends only on k. This is an important consideration for practice. Recall that in section 3.2 it was noted that when $Var(e_t)$ is small the time series estimates will not change the original estimates much, and that the gains from use of the time series estimates will come when $Var(e_t)$ is large. Unfortunately, estimation of sampling error covariances is likely to be more difficult in the latter situation, such as when the sample size is small. If stationarity of e_t can be assumed then information about sampling covariances can be pooled over time, effectively increasing the sample size for this purpose. One simple approach is to average estimates of $Cov(e_t,e_{t+k})$ over t in some way.

In some cases it may be possible to make further assumptions about e_t yielding a model describing its covariance structure in terms of a small number of parameters. SSJ suggest some models for single- and multi-stage overlapping surveys, and note that when the pattern of overlap is such that units remain in the sample for no more than q time periods, then the covariance structure of e_t can be represented as a moving average model of order q. Miazaki (1986) used such a sampling error model in analyzing National Crime Survey data. Hausman and Watson developed an autoregressive - moving average model of order (1,15) depending on only one parameter for sampling error in the Current Population Survey.

For many surveys it may be more appropriate to assume e_t is relative covariance stationary, i.e. the relative variance $R_t = Var(e_t|\Omega_t)/\theta_t^2$ remains stable over time. Consider the decomposition (3.1). We can show that

$$Var[\ln(u_t)] \approx Var(\tilde{u}_t) = E(R_t)$$

if $\tilde{u}_t$ is not too large. Proceeding similarly with lagged covariances, we see it would be reasonable to assume $\ln(u_t)$ is stationary. If it is also reasonable to take $\ln(\theta_t)$ then we can proceed with the decomposition (3.2) as we would have with (2.1) and exponentiate results at the end (see (3.3)). An alternative to this is to go ahead and estimate the time varying $Var(e_t)$ and use the results (2.3) - (2.6) (or the Kalman filter) which do not actually require e_t to be stationary, rather than the signal extraction formulas given later which do. However, this will complicate things, and it seems likely that often when e_t is nonstationary but $\ln(u_t) \approx \tilde{u}_t$ is approximately stationary, that we will be better off using (3.2) than (2.1).

4.2 Time Series Modeling

General treatments of time series modeling are readily available elsewhere, a good start-

ing point being the book by Box and Jenkins (1976). Here we comment on a few aspects of modeling we consider especially important and a few particular to the problem of accounting for sampling error in modeling.

The first step in modeling should be to deal with nonstationarity in the data. We have already mentioned the possibility of taking logarithms of Y_t to help render both the sampling error and θ_t (approximately) covariance stationary. Other transformations of Y_t might also be considered, though we would then usually not be able to directly interpret the transformed series as the sum of a population value and sampling error. A choice between $\ln(Y_t)$ and no transformation will be enough to deal with many cases.

Simply taking logarithms is not likely to be enough to render θ_t and Y_t stationary. However, many time series Y_t have been modeled assuming that taking the first difference $(1-B)Y_t = Y_t - Y_{t-1}$ (B is the backshift operator such that $BY_t = Y_{t-1}$), or a seasonal difference such as $(1-B^{12})Y_t = Y_t - Y_{t-12}$, or both, produces a stationary series. It will thus often be reasonable to assume that θ_t suitably differenced is stationary or approximately so.

We may also want to allow Y_t and θ_t to have a mean function that varies over time -- the μ_t of section 2.1. This requires a parametric form for μ_t, such as the linear regression function $\mu_t = \beta_1 X_{1t} + \cdots + \beta_k X_{kt}$. An example of this sort of thing for time series data from economic surveys is the modeling of calendar variation (see Bell and Hillmer 1983). For seasonal data, seasonal indicator variables for the X_{it} (analogous to one-way analysis of variance) are useful if the seasonal pattern in θ_t is stable over time. Particular examples will dictate the choice of regression variables. The type of model we are thus suggesting for θ_t (or $\ln(\theta_t)$) is a regression model with correlated errors, with the correlation in the errors described by a time series model that will likely involve differencing. Notice that if we are differencing θ_t we must also difference the regression variables the same way since the regression relation is generally specified between the undifferenced θ_t and X_{it}. Thus, if we are taking $(1-B)\theta_t$ we should also take $(1-B)X_{it}$ for $i = 1,\cdots,k$.

These three techniques -- transformation, differencing, and use of regression mean functions -- appear to be sufficient in practice to render many time series approximately stationary. Some authors have chosen to use regression on polynomials of time rather than differencing to help induce stationarity. R. G. Jones (1980), and Rao, Srinath, and Quenneville (1986) have suggested this in connection with the use of the time series approach to survey estimation. We recommend <u>against</u> the use of polynomial regression on time. It is known that using polynomial regression on time when differencing is needed has potentially dire consequences for regression results and time series analysis, while unnecessary differencing has far less serious effects. (See Nelson and Kang 1984 and the references given there.) In fact, if a model with a polynomial function of time is really appropriate, analysis of the differenced data can discover this (Abraham and Box 1978). Or since differencing, like taking derivatives, annihilates polynomials, use of certain models (noninvertible moving average) for differenced data can produce results equivalent to polynomial regression (Harvey 1981). The moral of this is that polynomial regression on time can lead to trouble while differencing probably will not. While the literature has not considered these issues in the particular context of the time series approach to survey estimation, it seems far safer to difference than to hope polynomial regression on time is appropriate or that it will not have bad effects.

Let $z_t = \theta_t - \mu_t$ where, e.g., $\mu_t = \beta_1 X_{1t} + \cdots + \beta_k X_{kt}$. At this point the model we are suggesting is $Y_t = \theta_t + e_t$ with

$$(4.1)\quad \delta(B)[\theta_t - (\beta_1 X_{1t} + \cdots + \beta_k X_{kt})] = \delta(B) z_t = w_t$$

where $\delta(B)$ is a differencing operator such as $(1-B)$ or $(1-B)(1-B^{12})$ and w_t is a stationary series. We can use an analogous model if we are taking logarithms of the data. We still need a model for w_t, or equivalently a model for z_t incorporating differencing. Two types of models popular in the time series literature are the autoregressive - integrated - moving average (ARIMA) models discussed by Box and Jenkins (1976), and the structural (or unobserved components, or state-space) models considered by Harvey and Todd (1983) and Kitagawa and Gersch (1984), among others. We refer the reader to these references for complete treatments of these models. There is a correspondence between the two types of models since structural models with ARIMA components imply some ARIMA model for the sum of the components z_t. For low order nonseasonal models this correspondence implies that in many cases both modeling approaches can yield the same model for z_t. Both approaches have their proponents, but even for seasonal series the jury is still out as to how much difference there really is between the models, let alone which is to be preferred.

An important feature of modeling the time series Y_t is the presence of a component, the sampling error e_t, that we know something about. There are two ways to get at the covar-

iance structure of θ_t. We can directly model Y_t, not explicitly accounting for e_t, and derive the covariance structure for θ_t by subtraction. Or we can specify a model for θ_t and fit a model to Y_t corresponding to this model for θ_t and the assumed known covariance structure or model for e_t. If there is little sampling variation present (Var(e_t) small) then it will make little difference which approach is used, but this is also the situation where the time series approach will not make much difference either. If there is substantial sampling variation directly modeling Y_t may be adequate in some cases, but in general it may be important to use a model for Y_t explicitly incorporating separate models for θ_t and e_t. We feel more experience with this type of modeling is needed before firm recommendations can be given. New computer software may also be needed.

4.3 Signal Extraction Computations

Here we consider alternative approaches to computing the basic results given earlier as (2.3) - (2.6). We can obviously apply these by subtracting the means μ_t from the data Y_t to start, using the results assuming means equal to zero, and then adding μ_t back to $\hat{\theta}_t$ at the end. In this section we shall thus assume means equal to zero for simplicity. In this case (2.3) and (2.6) become

$$(4.2)\quad \hat{\underset{\sim}{\theta}} = \Sigma_\theta \Sigma_Y^{-1} \underset{\sim}{Y} \quad \text{and} \quad \mathrm{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) = \Sigma_e - \Sigma_e \Sigma_Y^{-1} \Sigma_e$$

Scott and Smith (1974) and SSJ used classical time series signal extraction results given, e.g., by Whittle (1963). Assuming a doubly infinite sequence Y_t is available, and that Y_t, θ_t, and e_t are all stationary, these results for our problem become

$$(4.3)\quad \hat{\theta}_t = \gamma_\theta(B)/\gamma_Y(B)\, Y_t \quad \text{and}$$

$$\gamma_{\hat{\theta}-\theta}(B) = \gamma_e(B) - \gamma_e(B)^2/\gamma_Y(B)$$

where $\gamma_Y(B)$ is the covariance generating function of Y_t, defined by $\gamma_Y(B) = \sum_{-\infty}^{\infty} \gamma_Y(k)\, B^k$, where $\gamma_Y(k) = \mathrm{Cov}(Y_t, Y_{t+k})$, and similarly for $\gamma_\theta(B)$, etc. Comparing (4.2) and (4.3) we see that covariance generating functions are the analogues of covariance matrices for use with infinite time series instead of random vectors. Given models for Y_t, θ_t, and e_t the results simplify. For the ARMA model $\phi(B)Y_t = \eta(B)a_t$, $\gamma_Y(B) = \eta(B)\eta(F)\sigma_a^2/\phi(B)\phi(F)$ where $F = B^{-1}$ is the forward shift operator. We can expand $\gamma_{\hat{\theta}-\theta}(B)$ to pick out $\gamma_{\hat{\theta}-\theta}(0) = \mathrm{Var}(\hat{\theta}_t - \theta_t)$.

These results are useful for computing the estimate of θ_t and the variance of the error in the estimate when we have a reasonably long time series of observations on Y and t is somewhere in the middle of the series. For t near the endpoints 1 or T alternative formulas given by SSJ and Whittle (1963) can be used. Another option is to use the model for Y_t to forecast and backcast the series, append the forecasts and backcasts to the end and beginning of the data $Y_1, \ldots, Y_T$, and apply the symmetric filter in (4.3) to get $\hat{\theta}_t$. Bell (1980) established that this procedure converges pointwise (as the number of forecasts and backcasts extend into the infinite future and past) to the results for $\hat{\theta}_t$ given by (4.2). $\mathrm{Var}(\hat{\theta}_t - \theta_t)$ can then be obtained using results of Pierce (1979) or Hillmer (1985).

A third approach to doing the computations is to put the model for $Y_t = \theta_t + e_t$ into state space form and use the Kalman filter/smoother (Anderson and Moore 1979). This recursively computes the $\hat{\theta}_t$ and $\mathrm{Var}(\hat{\theta}_t - \theta_t)$ for $t = 1, \ldots, T$; covariances of the estimation errors can also be obtained.

Bell (1984) extended the classical signal extraction results (4.3) under certain assumptions to the case of nonstationary series requiring differencing. Essentially the results remain the same with the differencing operators carried along in the covariance generating function as autoregressive operators. The Kalman filter/smoother does not require stationarity, but does require assumptions about initial conditions that have often been made rather arbitrarily, especially in the nonstationary case. This problem has been addressed by the modified Kalman filter of Kohn and Ansley (1986,1987). Bell and Hillmer (1987) show how to obtain results equivalent to those of Kohn and Ansley with the ordinary Kalman filter.

It is important to remember that the three approaches discussed for doing the signal extraction computations will, if all are using the same models and assumptions, produce the same results (with the exception for the classical results noted below). Thus, choice of approach depends on computational considerations, not on the results that will be obtained. The stochastic least squares results (4.2) (or (2.3) - (2.6)) are the most general, but are difficult computationally unless T is small. They also would be effectively impossible to apply directly in the nonstationary case. The classical results cannot be used in certain important cases, such as when the variance of the sampling errors changes over time. Also, they sometimes provide only approxima-

tions to the finite sample results, though these approximations are usually quite good as long as T is reasonably large. When the classical results are applicable they are computationally efficient, sometimes very easy to use, and they help give insight into what is going on through the filter weights of $\gamma_\theta(B)/\gamma_Y(B)$. The Kalman filter/smoother can be used as long as the problem can be put in state space form, which is sufficient for quite general problems, including the case of changing variances. It will accurately compute the exact finite sample results. For these reasons the Kalman filter/smoother may be preferred for a general purpose computer program.

5. Example -- Teenage Unemployment (CPS)

We analyze the time series of the total number of teenage unemployed, which is collected as part of the Current Population Survey (CPS) by the Census Bureau. The CPS is a monthly survey composed of eight rotating panels. Each panel is included in the survey for four months, left out of the survey for the next eight months, and then included in the survey for four final months. This rotation procedure produces a 75% overlap in the sample from month to month and a 50% overlap from year to year. We might expect correlation in the sampling errors for months with samples that overlap due to the rotation scheme. We might also expect that sampling errors for months with no sample overlap would be uncorrelated. However, when a sample unit leaves the survey it is usually replaced by a neighboring unit from the same geographic area, which may induce correlation at months with no sample overlap. The correlation in the sampling errors will also be affected by the composite estimation procedure used to derive the published estimates. The composite estimates used are an average of the ratio estimate for the current month, and the sum of last month's composite estimate and an estimate of the change between the current month and preceding month. Hausman and Watson (1985) derive a model for the sampling error in the CPS that depends on a single unknown parameter. Unfortunately, their derivation ignores the practice of replacing sample units with neighboring units. It may be difficult to modify the Hausman-Watson model to account for this practice.

Train, Cahoon, and Makens (1978) report the average autocorrelations for the teenage unemployed sampling errors based upon the survey microdata between December 1974 and December 1975. These autocorrelations are reproduced in Table 1a. The autocorrelation function for the model

$$(5.1) \qquad (1-\phi B)e_t = (1-\eta B)c_t$$

with $\phi = .6$ and $\eta = .3$ is reported in Table 1b. It appears that this model well approximates the estimated autocorrelation structure of the teenage unemployed sampling errors. It should be noted that agreement between the two sets of autocorrelations at the higher lags is less important than at the lower lags because there was more data available to estimate the lower lag autocorrelations, presumably making them more reliable. In our subsequent analysis we will use model (5.1) to describe the autocorrelations of the sampling errors.

There have been many changes to CPS over the years, and for our purposes it is important to be aware of those changes that will possibly affect the correlation structure of the sampling errors. Two major changes are (i) the redesign based on the 1970 Census starting in January of 1972, and (ii) the redesign based on the 1980 Census starting in January of 1984. In order to get a reasonably long time series that is consistent with the autocorrelations reported in Table 1a, we use the teenage unemployed data from January 1972 through December of 1983 in our analysis. Once the model has been estimated it could be used to produce signal extraction estimates for more recent data (assuming, of course, that the model still applies), such as data from January 1984 through the current time.

In order to compute the signal extraction estimates, we need estimates of the variances of the sampling errors. The Census Bureau uses the method of generalized variance functions (Wolter 1985, Chapter 5) for these variance estimates. If Y_t is the composite estimate of the number in thousands of teenage unemployed at time t, then the estimate of the variance of the sampling error e_t is given by

$$(5.2) \qquad \hat{Var}(e_t) = -.0000153\, Y_t^2 + 1.971\, Y_t$$

The use of generalized variance functions in CPS is discussed in Technical Paper 40 (U.S. Department of Commerce, Bureau of the Census 1968). The particular coefficients in (5.2) were provided by Donna Kostanich of the Statistical Methods Division. They were developed in 1977, about the middle of our time series, and so are reasonable for use with our data. Slightly different coefficients may be more appropriate for more recent data. The relation between the estimated variance of the sampling error and the estimated level cannot be transformed away. We shall use the Kalman filter to deal with this problem.

If $Y_t = \theta_t + e_t$ where each of the components follow ARIMA type models, it is straightforward (see, e.g., Gersch and Kitagawa, 1983) to write these in state space form

$$(5.3) \qquad \underset{\sim}{X}_{t+1} = F\underset{\sim}{X}_t + G\underset{\sim}{v}_t$$

$$(5.4) \qquad Y_t = H\underset{\sim}{X}_t$$

(Note that in our problem there is no added error in equation (5.4)). Then given observations $Y_1,\ldots,Y_n$ one can use the Kalman filter algorithm to evaluate the likelihood function (see R. H. Jones, 1980) and use a standard nonlinear optimization routine to find the param-

eters that maximize the likelihood function. In our particular case, the matrix H_t in the measurement equation (5.4) will not be time invariant because one element of H_t will be the standard error of e_t, which depends on Y_t. Once the parameters have been estimated, the Kalman filter and a fixed interval Kalman smoother (see Anderson and Moore, 1979) can be used to compute the signal extraction estimates and their variances.

Since the composite estimates are design unbiased, by Result 3.1 θ_t is uncorrelated with e_t. Having specified a model for e_t, we now must specify the model for θ_t. In doing this we consider the correlation structure of the observed data, Y_t. The ACF of Y_t fails to die out, suggesting the need to first difference. The ACF of the first difference of Y_t exhibits a persistent periodic pattern suggesting the need for an additional seasonal difference to achieve stationarity. Since the model (5.1) for e_t does not include differencing, the difference operators are attributable to the model for θ_t. The ACF of $(1-B)(1-B^{12})Y_t$ has prominent negative values at lags 1 and 12. While this reflects combined effects of θ_t and e_t, so that the model for θ_t is not immediately apparent, a model for θ_t not inconsistent with the results for Y_t is

$$(5.5)\quad (1-B)(1-B^{12})\theta_t = (1-\eta_1 B)(1-\eta_{12}B^{12})b_t.$$

The uncertain nature of model identification here makes it especially important to check the model for Y_t after estimation.

Our model for $Y_t = \theta_t + e_t$ is specified by (5.1), (5.2), (5.5), and the uncorrelatedness of θ_t and e_t. We need to estimate the parameters η_1, η_{12}, and σ_b^2 from the observed teenage unemployment data, $\{Y_t\}$. This was done by numerically maximizing the likelihood function under the assumption of Gaussian errors, with the likelihood evaluated using a Kalman filter algorithm. The maximum likelihood estimates are

$$\hat{\eta}_1 = .26 \qquad \hat{\eta}_{12} = .78 \qquad \hat{\sigma}_b^2 = 3931$$

A time series plot of the residuals from the model revealed no major problems with the model. The autocorrelations of the residuals are all smaller than two times their standard errors, and the Ljung-Box Q statistic (Ljung and Box, 1978) computed for 24 lags is 24.6, well below the .05 critical value of 33.9 for a chi-squared distribution with 22 degrees of freedom. Thus, examination of the residuals gives no reason to question the validity of the model.

A Kalman fixed interval smoother was used to compute the signal extraction estimates and their variances, using the model with the estimated parameters, and equation (5.2) for $\mathrm{Var}(e_t)$. The signal extraction estimates are plotted along with the usual composite estimates for the last 100 observations in Figure 1a. The seasonal difference $(1-B^{12})$ of the signal extraction estimates and the seasonal difference of the composite estimates are plotted in Figure 1b. It is apparent from these graphs that the signal extraction estimates are smoother than the composite estimates.

The standard errors of both the last 100 signal extraction estimates and the last 100 composite estimates are plotted in Figure 2a. They both vary over time, with the standard errors of the signal extraction estimates being uniformly smaller than the standard errors of the composite estimates. Figure 2b shows the ratios of the signal extraction to the composite standard errors. As a rough measure of the average improvement, the geometric mean of these ratios is .79, reflecting about a 21% reduction in the standard error, or a 38% reduction in the variance due to signal extraction. From Figures 2a and 2b it is also apparent that the difference in standard errors is smaller near the end of the data. This behavior is to be expected since at the end of the series the signal extraction estimates cannot make use of future data.

REFERENCES

Abraham, B. and Box, G.E.P. (1978), "Deterministic and Forecast-Adaptive Time-Dependent Models," Applied Statistics, 27, 120-130.

Anderson, B.D.O. and Moore, J. B. (1979), Optimal Filtering, Englewood Cliffs: Prentice-Hall.

Bell, W. R. (1980), "Multivariate Time Series: Smoothing and Backward Models," Ph.D. thesis, University of Wisconsin-Madison.

Bell, W. R. and Hillmer, S. C. (1983), "Modeling Time Series with Calendar Variation," Journal of the American Statistical Association, 78, 526-534.

_____(1984), "Issues Involved with the Seasonal Adjustment of Economic Time Series," (with discussion), Journal of Business and Economic Statistics, 2, 291-320.

_____(1987), "Initializing the Kalman Filter in the Non-stationary Case: With Application to Signal Extraction," paper prepared for the 1987 meeting of the American Statistical Association, San Francisco.

Binder, D. A. and Dick, J. P. (1986), "Modelling and Estimation for Repeated Surveys" Statistics Canada Technical Report.

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

Fuller, W. A. and Isaki, C. T. (1981), "Survey Design Under Superpopulation Models," in Current Topics in Survey Sampling ed. D. Krewski, R. Platek, and J.N.K. Rao, New York: Academic Press, 199-226.

Gersch, W. and Kitagawa, G. (1983), "The Prediction of Time Series With Trends and Seasonalities," Journal of Business and Economic Statistics, 1, 253-264.

Harvey, A. C. (1981), "Finite Sample Prediction and Overdifferencing," Journal of Time Series Analysis, 2, 221-232.

Harvey, A. C. and Todd, P.H.J. (1983), "Forecasting Economic Time Series With Structural and Box-Jenkins Models: A Case Study," (with discussion), Journal of Business and Economic Statistics, 1, 299-315.

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

Hillmer, S. C. (1985), "Measures of Variability for Model-Based Seasonal Adjustment Procedures," Journal of Business and Economic Statistics, 3, 60-68.

Hillmer, S. C. and Trabelsi, A. (1986), "Benchmarking of Economic Time Series," University of Kansas School of Business Working Paper.

Jones, R. G. (1980), "Best Linear Unbiased Estimators for Repeated Surveys," Journal of the Royal Statistical Society, Series B, 42, 221-226.

Jones, R. H. (1980), "Maximum Likelihood Fitting of ARMA Models to Time Series With Missing Observations," Technometrics, 22, 389-395.

Kitagawa, G. and Gersch, W. (1984), "A Smoothness Priors-State Space Modeling of Time Series With Trend and Seasonality," Journal of the American Statistical Association, 79, 378-389.

Kohn, R. and Ansley, C. F. (1986), "Estimation, Prediction, and Interpolation for ARIMA Models With Missing Data," Journal of the American Statistical Association, 81, 751-761.

Ljung, G. M. and Box, G. E. P. (1978), "On a Measure of Lack of Fit in Time Series Models," Biometrika, 65, 297-304.

Miazaki, E. S. (1985), "Estimation for Time Series Subject to the Error of Rotation Sampling," unpublished Ph.D. thesis, Department of Statistics, Iowa State University.

Nelson, C. R. and Kang, H. (1984), "Pitfalls in the Use of Time as an Explanatory Variable in Regression," Journal of Business and Economic Statistics, 2, 73-82.

Pierce, D. A. (1979), "Signal Extraction Error in Nonstationary Time Series," Annals of Statistics, 7, 1303-1320.

Rao, J.N.K. and Graham, J. E. (1964), "Rotation Designs for Sampling on Repeated Occasions," Journal of the American Statistical Association, 59, 492-509.

Rao, J.N.K., Srinath, K. P., and Quenneville, B. (1986), "Optimal Estimation of Level and Change Using Current Preliminary Data," paper presented at the International Symposium on Panel Surveys, Washington, D.C., November, 1986.

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.

Smith, T.M.F. (1978), "Principles and Problems in the Analysis of Repeated Surveys," Survey Sampling and Measurement, ed. N. K. Namboodiri, New York: Academic Press, 201-216.

Tiao, G. C. and Hillmer, S. C. (1978), "Some Consideration of Decomposition of a Time Series," Biometrika, 65, 497-502.

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

U. S. Department of Commerce, Bureau of the Census (1968), "The Current Population Survey: Design and Methodology" by Robert H. Hanson, Technical Paper No. 40, Washington, D. C., U.S.Government Printing Office.

Whittle, P. (1963), Prediction and Regulation by Linear Least-Square Methods, Princeton: Van Nostrand.

Wolter, K. M. (1979), "Composite Estimation in Finite Populations," Journal of the American Statistical Association, 74, 604-613.

_____(1985), Introduction to Variance Estimation, New York: Springer-Verlag.

Wolter, K. M. and Monsour, N. J. (1981), "On the Problem of Variance Estimation for a Deseasonalized Series," in Current Topics in Survey Sampling, ed. D. Krewski, R. Platek, and J.N.K. Rao, New York: Academic Press, 199-226.

FOOTNOTES

[1]This paper reports the general results of research undertaken by Census Bureau staff and staff of the University of Kansas. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau or of the University of Kansas.

The paper is based in part upon work supported by the National Science Foundation under grant SES 84-01460, "On-Site Research to Improve the Government-Generated Social Science Data Base." The research was partially conducted at the U.S. Bureau of the Census while the second author was a participant in the American Statistical Association/Census Bureau Research Program, which is supported by the Census Bureau and through the NSF grant. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Table 1a
Teenage Unemployment Sampling Error Autocorrelations

Lag	1	2	3	4	5	6	7	8	9	10	11	12
Correlation	.35	.24	.14	.08	.03	.01	.02	.01	.02	.06	.01	.08

Table 1b
Autocorrelations for an ARMA (1,1) Model with $\phi = .6$ and $\eta = .3$

Lag	1	2	3	4	5	6	7	8	9	10	11	12
Correlation	.34	.20	.12	.07	.04	.03	.02	.01	.01	.00	.00	.00

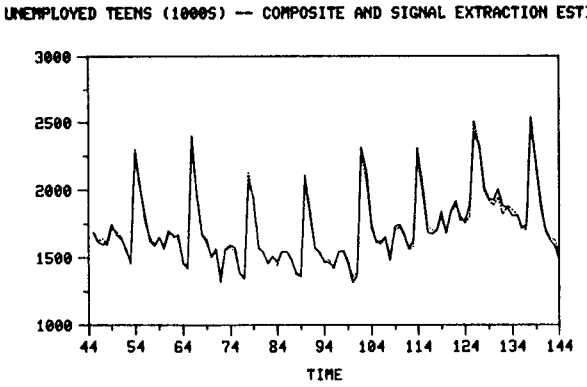

Figure 1a.

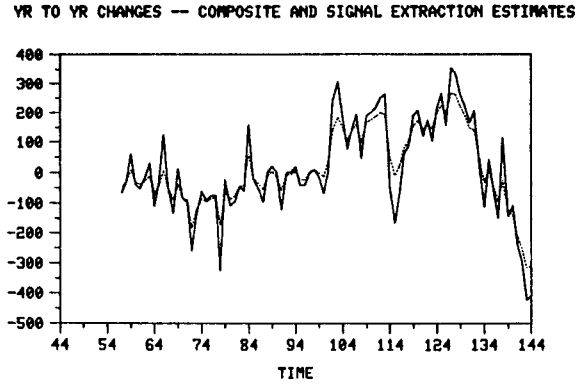

Figure 1b.

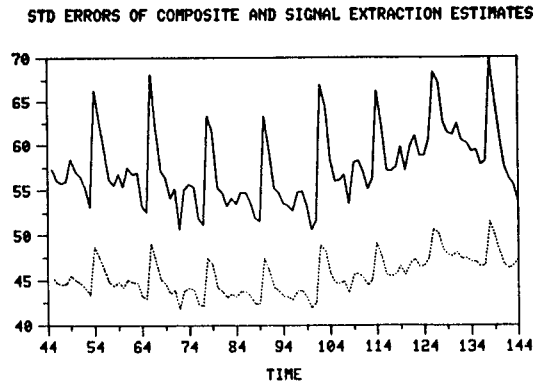

Figure 2a.

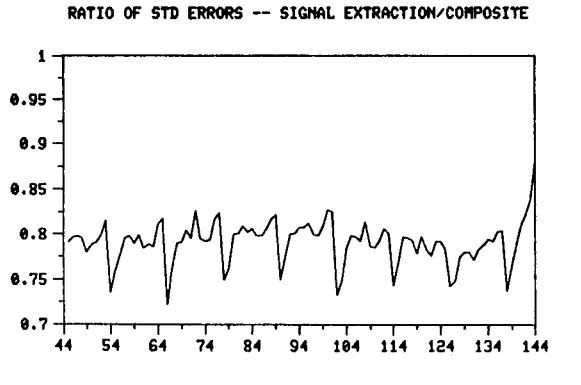

Figure 2b.