BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-87/20

TIME SERIES METHODS FOR SURVEY ESTIMATION

by

William Bell
Statistical Research Division
U.S. Bureau of the Census
Washington, D.C. 20233

Steven Hillmer
School of Business
University of Kansas
Lawrence, KS 66045

Recommended by: Nash J. Monsour

Report completed: August 12, 1987

Report issued: August 13, 1987

Report revised: October 13, 1987

TIME SERIES METHODS FOR SURVEY ESTIMATION

William R. Bell

Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

Steven C. Hillmer

School of Business
University of Kansas
Lawrence, Kansas 66045

August, 1987

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.

ABSTRACT

Papers by Scott and Smith (1974) and Scott, Smith, and Jones (1977) 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 and their sampling errors is known, these results produce the linear functions of the usual estimators that have minimum mean squared error as estimators of the population values. Thus, current and past data are used in estimating the population quantity at the current time. To apply these results in practice one would identify and estimate a time series model for the time series of usual survey estimators, and estimate the covariance structure of the sampling errors over time using knowledge of the survey design. The paper reviews the theory behind this work, obtains some theoretical results on this approach, discusses some considerations involved in applying this approach, and reports on results obtained to date regarding practical application of these results to Census Bureau surveys.

ACKNOWLEDGMENTS

We wish to thank Phillip S. Kott and David F. Findley for valuable suggestions regarding the consistency results, and Abdelwahed Trabelsi for his support as an ASA/NSF/Census Research Associate. We also wish to thank all those who participated in the summer of 1986 seminar series on time series methods for survey estimation, especially Charles Alexander, Lawrence Cahoon, Ruth Detlefsen, and Jesse Pollock for making presentations. Thanks to Dennis Duke and Don Luery of Construction Statistics Division for providing estimates of sampling variances and correlations for housing starts, and to Donna Kostanich of Statistical Methods Division for providing information on variance estimation and sample changes for the Current Population Survey. Finally, a thank you to all others in the various divisions at Census who assisted us in obtaining data, patiently explained survey procedures to us, and did their best to help us understand something about survey sampling theory. Any errors and misconceptions in the paper are our responsibility, not theirs.

1. 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 considers some theoretical issues and section 4 some application considerations for the approach. Our work on this topic has been part of our continuing study investigating the application of this approach to surveys at the U. S. Bureau of the Census. In section 5 we give results for some examples we have studied to date.

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:

$$Y_t = \theta_t + e_t \qquad (2.1)$$

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. The most basic thing to keep in mind about the use of time series techniques in survey estimation is that there are two distinct sources of stochastic variation present that are conceptualized, modeled, and estimated differently. It has been our experience that many people (including us) trained in one of the two specialties have some difficulty keeping straight the source of variation typically dealt with in the other specialty.

2.1 Basic Results

Suppose that estimates Y_t are available at a set of time points labelled $t = 1,\ldots,T$. 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

$$E(\underset{\sim}{Y}) = E(\underset{\sim}{\theta}) = \underset{\sim}{\mu} = (\mu_1,\ldots,\mu_T)'$$

$$\Sigma_Y = \Sigma_\theta + \Sigma_e. \tag{2.2}$$

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 $t = 1,\ldots,T$ is given by

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

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

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

If normality is assumed (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. 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 these issues further 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), Jones (1980), and Binder and Dick (1986) review and discuss the MVLU approach. While both the MVLU and time series approaches can 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 correlation of the e_t's. We can see the distinction in terms of our results for the simple case (2.1) where only one direct estimate, Y_t, of θ_t is available, by letting $\mathrm{Var}(\theta_t) \to \infty$. This translates the idea that the θ_t's are fixed into a non-informative prior on them. Then (2.5) becomes $\underset{\sim}{\hat{\theta}} = \underset{\sim}{Y}$ and we are only using Y_t to estimate θ_t. These remarks also apply to composite estimation (Rao and Graham 1964, Wolter 1979), which can be viewed as an approximation to MVLU.

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

The time series approach affords opportunities for improving estimation in periodic surveys through the recognition of the two sources of variability and the use of the results in section 2.1. There are other problems in repeated surveys 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.

Rao, Srinath, and Quenneville (1986) have applied time series techniques to the problem of preliminary estimation in repeated surveys where the preliminary estimate is later followed by a final estimate using additional data. Time series modeling and forecasting was cited earlier as one example where sampling variability is typically ignored. Recognizing and incorporating sampling variation may lead to improved time series and econometric models and forecasts, especially in cases where sampling variation is relatively large, and the sampling errors are strongly correlated over time. Seasonal adjustment is another example where sampling variation is typically ignored; exceptions are the papers by Hausman and Watson (1985) and Wolter and Monsour (1981). In seasonal adjustment a time series is decomposed into seasonal (S_t), trend (T_t), and irregular (I_t) components. Once sampling variation is recognized, an appropriate decomposition is

$$Y_t = S_t + T_t + I_t + e_t$$

where e_t is sampling error. One would probably wish to remove both S_t and e_t

in adjusting, and so estimate $T_t + I_t$. Time series trend estimation obviously gives us another example where recognition of sampling variation could improve results. Oddly, detection of statistically significant change over time in estimates from repeated surveys is a problem closely related to trend estimation where sampling variation is recognized and time series variation typically ignored. Smith (1978) has pointed out the difficulties in this. Finally, benchmarking, the reconciling of results from a periodic survey with the results from another survey (possibly periodic of different period) or census estimating the same population charcteristics, can benefit from a treatment recognizing both time series and sampling variation (Hillmer and Trabelsi (1986)).

3. Theoretical Consideration of the Time Series Approach

In this section we consider some theoretical and philosophical issues that come up in connection with the time series approach. The next section discusses application considerations. We need to make one caveat. The decomposition (2.1) does not allow for nonsampling errors, nor does the time series approach explicity treat them. This is not to say that nonsampling errors are not important. Whether nonsampling error is generally more or less of a problem for the time series approach than for the classical approach is an open question, but it would be wise to consider the possible effects of known or suspected nonsampling errors on the time series estimators when actually applying them in particular situations.

3.1 Why assume θ_t a stochastic time series?

This issue has been discussed by SSJ and at length by Smith (1978). They observe that (1) users of data from repeated surveys treat the data Y_t as a stochastic time series in modeling and would do the same with θ_t if it were available (as noted above sampling error is typically ignored in analyzing time series data), and (2) classical results for estimation in repeated surveys (MVLU) assume a time series structure for the individual units in the population, while maintaining the anomalous position that θ_t, which is a function of these individual units (such as the total), is a sequence of fixed, unrelated quantities. In observing a result analogous to that noted above where we considered $\Sigma_\theta \to \infty$, Smith (1978, p. 208) observes that:

> This indicates just how strong is the assumption that θ_t is an unknown constant. It implies that θ_t cannot be predicted in any way from knowledge of the previous values θ_{t-1}, θ_{t-2}, etc. Surely in most repeated surveys the parameter values would change only moderately with time, and hence knowledge of θ_{t-1} would be very useful in predicting θ_t. To

ignore this information seems very wasteful.

We can push this philosophical argument even further. Suppose for a repeated survey a census is in fact done every time period, so that e_t drops out of (2.1) and we have $Y_t = \theta_t$. If we assume θ_t, and hence now Y_t, is a sequence of fixed, unrelated quantities, then data through any time point t are irrelevant to the future behavior of the series. If this were in fact the case, then there would be little point in doing the survey in the first place. The data would be out of date as soon as they were published. The presence of sampling error does not invalidate this argument. Thus, we conclude that it is more reasonable to assume θ_t is a stochastic time series than a sequence of fixed, unrelated quantities. The real questions are then whether or not we can estimate the time series structure of θ_t and e_t well enough to make beneficial use of this in survey estimation, and how worthwhile these benefits will be.

3.2 Model-based versus design-based approach

A subject of some controversy in survey inference is whether to use a model-based or design-based approach. (See Hansen, Madow, and Tepping 1983 and accompanying discussion.) The time series approach is not model-based in the same sense as the classical model-based approach to survey inference. The classical model-based approach to inference involves a model for the individual units, say y_{it}, of the population. The time series approach does not require a model for the y_{it}, it merely needs a time series model for the aggregate function of these, θ_t, as well as estimates of the variance and correlation over time of the sampling errors, which, it seems, could be either model-based or design based. Still, the results will be influenced by the model used for θ_t. While the choice of model for θ_t and its estimation may

well be of some importance, and there may be some concern about this, it should be kept in mind that the alternative of assuming θ_t unrelated over time is rather extreme, as discussed in the previous section.

3.3 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 that it is valid under fairly general conditions.

We let y_{it} be the value of the characteristic of interest for the i^{th} unit in the population at time t, and let $\Omega_t = \{y_{it}\ i=1,\ldots,N_t\}$ be the collection of all N_t of these units. We consider time points $t=1,\ldots,T$ and let $\underset{\sim}{\Omega} = (\Omega_1,\ldots,\Omega_T)'$. The y_{it} are random variables, as is $\theta_t = \theta_t(\Omega_t)$, which is a function of the y_{it}. The sample at time t, s_t (denoting the indices, not the values, of the units selected), has probability of selection $p(s_t|\underset{\sim}{\Omega})$. The

estimator Y_t of θ_t is a function of the values y_{it} for the units sampled, thus a function of both Ω_t and s_t, i.e. $Y_t = Y_t(\Omega_t, s_t)$. We could let Y_t depend on the sample at times other than t, but we ignore that here for simplicity.

We need to make one assumption about the sample selection process, that $p(s_t|\underset{\sim}{\Omega}) = p(s_t|\Omega_t)$ for all t. This is not as strong an assumption as assuming that the sample design is noninformative, which means s_t and $\underset{\sim}{\Omega}$ are independent, implying $p(s_t|\underset{\sim}{\Omega}) = p(s_t) = p(s_t|\Omega_t)$. Our assumption allows the sample selection process at time t ($p(s_t|\underset{\sim}{\Omega})$) to depend on the population values at time t (Ω_t), but assumes the population values at time points other than t (Ω_j for $j \neq t$) offer no additional information on s_t beyond that in Ω_t. This assumption might even be generalized.

The following Lemma leads to our result.

Lemma: (1) $E(Y_t|\Omega_t) = \sum_{s_t} Y_t\, p(s_t|\Omega_t) = \sum_{s_t} Y_t(\Omega_t, s_t) p(s_t|\Omega_t)$ where the sum is over all possible samples s_t. (2) $E(Y_t|\underset{\sim}{\Omega}) = E(Y_t|\Omega_t)$.

Proof: (1) is essentially obvious since given $\Omega_t = \omega_t$, $Y_t = Y_t(\Omega_t, s_t)$ takes the value $Y_t(\omega_t, s_t)$ with probability $p(s_t|\omega_t)$. A proof using general conditional expectation results can also be given. To see (2) we have $E(Y_t|\underset{\sim}{\Omega}) = \sum_{s_t} Y_t\, p(s_t|\underset{\sim}{\Omega}) = \sum_{s_t} Y_t\, p(s_t|\Omega_t) = E(Y_t|\Omega_t)$.

We will be concerned with design unbiased estimators Y_t of θ_t. This can be written $E(Y_t|\Omega_t) = \sum_{s_t} Y_t\, p(s_t|\Omega_t) = \theta_t$. The usual definition of design unbiasedness is $\sum_{s_t} Y_t\, p(s_t) = \theta_t$, but this assumes the y_{it} and so Ω_t are fixed. Our definition coincides with the usual one in this case, and also does if the sample design is noninformative. We have the following.

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

Proof: Consider $Cov(\theta_t, e_j)$ for two arbitrary time points t and j. Y_j design unbiased means $\theta_j = E(Y_j|\Omega_j) = E(Y_j|\underset{\sim}{\Omega})$ by the Lemma. Then $E(e_j|\underset{\sim}{\Omega}) = E(Y_j - \theta_j|\underset{\sim}{\Omega}) = E(Y_j|\underset{\sim}{\Omega}) - \theta_j = \theta_j - \theta_j = 0$ implying $E[E(e_j|\underset{\sim}{\Omega})] = E(e_j) = 0$. Also $E(\theta_t \cdot e_j|\underset{\sim}{\Omega}) = \theta_t E(e_j|\underset{\sim}{\Omega}) = \theta_t \cdot 0 = 0$ implying $E(\theta_t \cdot e_j) = 0$. Thus $Cov(\theta_t, e_j) = E(\theta_t \cdot e_j) - E(\theta_t)E(e_j) = 0$.

Comment: If $E(e_j|\underset{\sim}{\Omega})$ does not depend on $\underset{\sim}{\Omega}$ then e_j is said to be "mean independent" of $\underset{\sim}{\Omega}$, which is known to be a stronger condition than e_j and $\underset{\sim}{\Omega}$ uncorrelated, though not as strong as stochastic independence (unless we have normality). This shows that actually we only need $E(e_t|\underset{\sim}{\Omega}) = E(Y_t|\underset{\sim}{\Omega}) - \theta_t$ to not depend on $\underset{\sim}{\Omega}$ to get θ_t, e_t uncorrelated. This would cover the cases where Y_t has a constant additive bias (not dependent on Ω_t) as an estimate of θ_t, or, using approximate Result 3.2 which follows, a constant percentage (multiplicative) bias.

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

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

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

$$\ln(Y_t) = \ln(\theta_t) + \ln(1 + \tilde{u}_t) = \ln(\theta_t) + \ln(u_t) \qquad (3.2)$$

Now notice $E(\tilde{u}_t|\underset{\sim}{\Omega}) = E(e_t/\theta_t|\underset{\sim}{\Omega}) = E(e_t|\underset{\sim}{\Omega})/\theta_t = 0$ if Y_t is design unbiased. We then assume that $E(\ln(u_j)|\underset{\sim}{\Omega}) \approx \ln E(u_j|\underset{\sim}{\Omega}) = \ln(1) = 0$ so $E(\ln(u_j)) \approx 0$. Similarly $E[\ln(\theta_t)\cdot\ln(u_j)|\underset{\sim}{\Omega}] = \ln(\theta_t)\ E[\ln(u_j)|\underset{\sim}{\Omega}] \approx 0$ so $E[\ln(\theta_t)\cdot\ln(u_j)] \approx 0$ and thus $\mathrm{Cov}(\ln(\theta_t),\ln(u_j)) \approx 0$. Hence, we have

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

We could alternatively have obtained this result using the approximation $\mathrm{Corr}(\ln(\theta_t), \ln(u_j)) \approx \mathrm{Corr}(\theta_t, u_j)$ and noting $\tilde{u}_j$ and u_j are mean independent of $\underset{\sim}{\Omega}$.

3.4 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 $\underset{\sim}{Y}^\ell$, $\underset{\sim}{\theta}^\ell$, $\underset{\sim}{e}^\ell$, $\underset{\sim}{\Omega}^\ell$, $\underset{\sim}{u}^\ell$, $\underset{\sim}{\Sigma}_y^\ell$, $\underset{\sim}{\Sigma}_\theta^\ell$, $\underset{\sim}{\Sigma}_e^\ell$, $\hat{\underset{\sim}{\theta}}^\ell$, and $\hat{\theta}_t^\ell$ in the obvious fashion. We consider what happens to the time series estimators $\hat{\underset{\sim}{\theta}}^\ell$ when the estimators $\underset{\sim}{Y}^\ell$ are consistent, i.e. $Y_t^\ell \to \theta_t^\ell$ in some fashion as $\ell \to \infty$ for all t. For now we assume $\underset{\sim}{\mu}^\ell$, Σ_θ^ℓ, and Σ_e^ℓ, or models leading to these, are known for each ℓ. Since $\underset{\sim}{\mu}^\ell$ and $\underset{\sim}{\Sigma}_\theta^\ell$ are really superpopulation parameters for the time series, θ_t^ℓ, we wish to estimate, we shall assume these are the same for each population ℓ, that is, $\underset{\sim}{\mu}^\ell \equiv \underset{\sim}{\mu}$ and $\Sigma_\theta^\ell \equiv \Sigma_\theta$ (a positive definite matrix) for all ℓ. This is also partly for

convenience since we could get the same results assuming $\underset{\sim}{\mu}^{\ell} \to \underset{\sim}{\mu}$ and $\Sigma_{\theta}^{\ell} \to \Sigma_{\theta}$ as $\ell \to \infty$.

From (2.5) it would appear that $\underset{\sim}{Y}^{\ell} \to \underset{\sim}{\theta}^{\ell}$ would imply $\hat{\underset{\sim}{\theta}}^{\ell} \to \underset{\sim}{\theta}^{\ell}$ as long as $\Sigma_e^{\ell} \to 0$. This condition suggests we need mean square convergence of Y_t^{ℓ} to θ_t^{ℓ}. We thus consider estimators Y_t^{ℓ} of θ_t^{ℓ} such that

$$E[(Y_t^{\ell} - \theta_t^{\ell})^2] = E[(e_t^{\ell})^2] \to 0 \quad \text{as } \ell \to \infty .$$

Since $E[(e_t^{\ell})^2] = Var(e_t^{\ell}) + [E(e_t^{\ell})]^2$ this implies both $Var(e_t^{\ell}) \to 0$ and $E(e_t^{\ell}) \to 0$. Assuming $Y_t^{\ell} \to \theta_t^{\ell}$ in mean square for t=1,...,T then implies $\Sigma_e^{\ell} \to 0$. We can now establish

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

Proof: From $\underset{\sim}{Y}^{\ell} = \underset{\sim}{\theta}^{\ell} + \underset{\sim}{e}^{\ell}$ with $\Sigma_e^{\ell} \to 0$ we have $\Sigma_Y^{\ell} \to \Sigma_{\theta}$ even if $\underset{\sim}{\theta}^{\ell}$ and $\underset{\sim}{e}^{\ell}$ are correlated. (We have not assumed $\underset{\sim}{Y}^{\ell}$ design unbiased.) From (2.4) we have

$$\hat{\underset{\sim}{\theta}}^{\ell} - \underset{\sim}{\theta}^{\ell} = (\underset{\sim}{Y}^{\ell} - \underset{\sim}{\theta}^{\ell}) - \Sigma_e^{\ell}(\Sigma_Y^{\ell})^{-1}(\underset{\sim}{Y}^{\ell} - \underset{\sim}{\mu}) \qquad (3.3)$$

The first term on the right converges to $\underset{\sim}{0}$ in mean square; the second has mean $\underset{\sim}{0}$ and variance $\Sigma_e^{\ell}(\Sigma_Y^{\ell})^{-1}\Sigma_Y^{\ell}(\Sigma_Y^{\ell})^{-1}\Sigma_e^{\ell} = \Sigma_e^{\ell}(\Sigma_Y^{\ell})^{-1}\Sigma_e^{\ell} \to 0$ as $\ell \to \infty$. Since both terms converge to $\underset{\sim}{0}$ in mean square so does $\hat{\underset{\sim}{\theta}}^{\ell} - \underset{\sim}{\theta}^{\ell}$.

Convergence in probability is a more familiar concept in survey sampling. If $Y_t^{\ell} \to \theta_t^{\ell}$ in probability for t=1,...,T this does not necessarily guarantee $\Sigma_e^{\ell} \to 0$, which is mean square convergence, a stronger condition. But if there

is a random variable ζ_t with finite variance such that $|e_t^\ell| \le \zeta_t$ (almost surely) uniformly in ℓ, then $Y_t^\ell \to \theta_t^\ell$ in probability implies $Y_t^\ell \to \theta_t^\ell$ in mean square (**Chung**, 1968, p. 64). So a result on convergence in probability is obtained **by making** this assumption to get $Y_t^\ell \to \theta_t^\ell$ in mean square and using Result 3.3.

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. From (3.3) we see 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 to $\underset{\sim}{0}$ of (3.3) 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 $\mathrm{Var}(e_t)$ 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 $\Sigma_u^\ell = \mathrm{Var}(\ln(\underset{\sim}{u}^\ell))$ 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

$$\ln(\hat{\underset{\sim}{\theta}}^\ell) = \underset{\sim}{\mu} + [I - \Sigma_u^\ell(\Sigma_Y^\ell)^{-1}]\ (\ln(\underset{\sim}{Y}^\ell) - \underset{\sim}{\mu}). \tag{3.4}$$

If we are taking logarithms it is reasonable to assume Y_t^ℓ and θ_t^ℓ remain bounded away from 0, say $|Y_t^\ell| \geq \nu$ and $|\theta_t^\ell| \geq \nu$ (almost surely) for all t and ℓ for some constant $\nu > 0$.

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.

Proof: The analogue to (3.3) is

$$\ln(\hat{\underset{\sim}{\theta}}^\ell) - \ln(\underset{\sim}{\theta}^\ell) = (\ln(\underset{\sim}{Y}^\ell) - \ln(\underset{\sim}{\theta}^\ell)) - \Sigma_u^\ell(\Sigma_Y^\ell)^{-1}(\ln(\underset{\sim}{Y}^\ell) - \underset{\sim}{\mu})$$

If we can show $\Sigma_u^\ell \to 0$ we will have the result since this implies $\ln(\underset{\sim}{Y}^\ell) \to \ln(\underset{\sim}{\theta}^\ell)$ in mean square, and the second term on the right behaves exactly as that in (3.3). Notice

$$E[(\tilde{u}_t^\ell)^2] = E[(e_t^\ell)^2/(\theta_t^\ell)^2] \leq (E(e_t^\ell)^2)/\nu^2 \to 0 \text{ as } \ell \to \infty$$

thus $E[(\tilde{u}_t^{\ell})^2] = E[(u_t^{\ell} - 1)^2] \to 0$. This implies $Var(u_t^{\ell}) \to 0$ and $E(u_t^{\ell}) \to 1$. By Jensen's inequality (Chung, 1968, p. 45) since $\exp(\cdot)$ is a convex function

$$1 \leq \exp(E[\ln(u_t^{\ell})^2]) \leq E(\exp[\ln(u_t^{\ell})^2]) = E[(u_t^{\ell})^2]$$

But $E[(u_t^{\ell})^2] = Var(u_t^{\ell}) + [E(u_t^{\ell})]^2 \to 1$ so $\exp(E[\ln(u_t^{\ell})^2] \to 1$ implying $E[\ln(u_t^{\ell})^2] \to 0$. This yields $Var(\ln(u_t^{\ell})) \to 0$ as desired.

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 (3.4)) $\exp[\ln(\hat{\theta}_t^{\ell})] \to \theta_t^{\ell}$ in probability as $\ell \to \infty$ for t=1,...,T.

Proof: Since $\ln(\hat{\theta}_t^{\ell}) \to \ln(\theta_t^{\ell})$ in mean square implies convergence in probability, the result follows since $\exp(\cdot)$ is a continuous function (Chung, 1968, p. 66).

4. Application Considerations

Application of the time series approach to survey estimation requires (1) estimation of the sampling error covariances, $Cov(e_t,e_j)$, in Σ_e, (2) estimation of the mean (μ) 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.

There is a fundamental identification problem with doing a secondary analysis. Given time series data on Y_t we can get at Σ_Y by estimating a model for Y_t, and given estimates of $Cov(e_t,e_j)$ from a primary analysis we can get at Σ_θ through $\Sigma_\theta = \Sigma_Y - \Sigma_e$. 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 $\mathrm{Cov}(e_t, e_{t+k})$, which is the same for all t and depends only on k. This is an important consideration for practice. Recall that in section 3.4 it was noted that when $\mathrm{Var}(e_t)$ 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 $\mathrm{Var}(e_t)$ 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 $\mathrm{Cov}(e_t, e_{t+k})$ 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 economic surveys it may be more appropriate to assume e_t is relative covariance stationary. In this case we consider the decompositions previously given as (3.1) and (3.2):

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

$$\ln(Y_t) = \ln(\theta_t) + \ln(1 + \tilde{u}_t) = \ln(\theta_t) + \ln(u_t) \qquad [3.2]$$

It was shown in section 3.3 that if Y_t is design unbiased then $E(\tilde{u}_t|\underset{\sim}{\Omega}) = 0$, $E(\ln(u_t)) \approx 0$, and $\ln(\theta_t)$ and $\ln(u_t)$ are approximately uncorrelated time series. Now we will assume it is not $Var(e_t|\underset{\sim}{\Omega})$ but the relative variance, $R_t = Var(e_t|\underset{\sim}{\Omega})/\theta_t^2$, which remains stable over time. Considering the decompostion (3.1) we notice that

$$Var(\tilde{u}_t) = Var[E(\tilde{u}_t|\underset{\sim}{\Omega})] + E[Var(\tilde{u}_t|\underset{\sim}{\Omega})] = 0 + E(R_t)$$

since

$$Var(\tilde{u}_t|\underset{\sim}{\Omega}) = Var(e_t/\theta_t|\underset{\sim}{\Omega}) = Var(e_t|\underset{\sim}{\Omega})/\theta_t^2 = R_t.$$

We then note that

$$Var[\ln(u_t)] = Var[\ln(1 + \tilde{u}_t)] \approx Var(\tilde{u}_t) = E(R_t)$$

if $\tilde{u}_t$ is not too large. Applying a similar argument to 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 (3.4)). An alternative to this is to go ahead and estimate the time varying $\text{Var}(e_t)$ 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 starting 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 published 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 be reasonable

to assume that θ_t suitably differenced is stationary or approximately so in many cases. While θ_t and e_t could mathematically both be nonstationary in some offsetting way so that their sum Y_t was stationary (or arguing similarly for some difference of all these series), this seems unlikely in practice.

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. 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

$$\begin{aligned} Y_t &= \theta_t + e_t \\ \delta(B)\,[\theta_t - (\beta_1 X_{1t} + \cdots + \beta_k X_{kt})] &= \delta(B) z_t = w_t \end{aligned} \tag{4.1}$$

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. To give a simple illustration, a typical ARIMA model used for monthly seasonal series is the

"airline model" (Box and Jenkins 1976, ch. 9)

$$(1 - B)(1 - B^{12})\, z_t = (1 - \eta_1 B)(1 - \eta_{12} B^{12})\, a_t \qquad (4.2)$$

where η_1 and η_{12} are parameters and a_t is a sequence of iid random variables with mean 0 and $\mathrm{Var}(a_t) = \sigma^2$. A typical structural model for similar data would use the seasonal + trend + irregular decomposition mentioned in section 2.3 with simple ARIMA models for the components:

$$\begin{aligned} z_t &= S_t + T_t + I_t \\ \text{where}\quad (1 + B + \cdots + B^{11})\, S_t &= \epsilon_{1t} \\ (1 - B)^2\, T_t &= \epsilon_{2t} \end{aligned} \qquad (4.3)$$

and ϵ_{1t}, ϵ_{2t}, and I_t are independent white noise series with variances (the model parameters) σ_1^2, σ_2^2, and σ_I^2. There is a correspondence between the two types of models since ARIMA component models 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 covariance 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 ($\mathrm{Var}(e_t)$ 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, but e_t is not

correlated or only weakly correlated over time (as in nonoverlapping surveys), or has a correlation pattern similar to that of θ_t, directly modeling Y_t may be adequate. The examples given by SSJ are cases in point. On the other hand, if e_t is strongly correlated over time in a very different way from θ_t, then it will probably 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. Standard packages for estimating ARIMA models with regression terms are not set up to also handle sampling error components. Programs for estimating structural models may not handle components with the sorts of models that may be appropriate for sampling errors, and may not handle components whose covariance structure is already known (estimated separately from the time series model estimation).

A final problem worth considering is that of model specification, or what Box and Jenkins (1976) refer to as identification. Their approach involves looking at autocorrelations and partial autocorrelations of the data. Since our data is on Y_t this can be used if we are directly modeling Y_t, but it will not readily identify a model for θ_t. Assuming that we have estimates of sampling error covariances, $\text{Cov}(e_t, e_{t+k}) = \gamma_e(k)$, a logical approach is to compute estimates of covariances for Y_t, $\gamma_Y(k)$, and then estimate covariances for θ_t as $\gamma_\theta(k) = \gamma_Y(k) - \gamma_e(k)$. Autocorrelations and partial autocorrelations can then be computed for θ_t. This can also be done for various differences of Y_t, e_t, and θ_t. Another approach is to just specify a model for θ_t of reasonable and simple form, or specify and fit several models and use a model selection criterion to choose between them. The latter approach is typically used with structural models. In this connection Kitagawa and Gersch (1984) suggest the use of Akaike's AIC criterion for model

discrimination.

4.3 Signal Extraction Computations

Here we consider alternative approaches to doing the signal extraction computations. The basic results were 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

$$\hat{\underset{\sim}{\theta}} = \Sigma_\theta \, \Sigma_Y^{-1} \, \underset{\sim}{Y} \qquad \mathrm{Var}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) = \Sigma_e - \Sigma_e \, \Sigma_Y^{-1} \, \Sigma_e \tag{4.4}$$

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

$$\hat{\theta}_t = \gamma_\theta(B)/\gamma_Y(B) \; Y_t \qquad \gamma_{\hat{\theta}-\theta}(B) = \gamma_e(B) - \gamma_e(B)^2/\gamma_Y(B) \tag{4.5}$$

where $\gamma_Y(B)$ is the covariance generating function of Y_t, defined by

$$\gamma_Y(B) = \sum_{-\infty}^{\infty} \gamma_Y(k) \, B^k$$

$$\gamma_Y(k) = \mathrm{Cov}(Y_t, Y_{t+k}) = (2\pi)^{-1} \int_{-\pi}^{\pi} e^{i\lambda k} \, \gamma(e^{-i\lambda}) \, d\lambda \tag{4.6}$$

and similarly for $\gamma_\theta(B)$, etc. Comparing (4.4) and (4.5) 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 can be simplified. If Y_t follows the ARMA model $\phi(B) \; Y_t = \eta(B) \; a_t$, then $\gamma_Y(B) = \eta(B)\eta(F)\sigma_a^2/\phi(B)\phi(F)$ where $F = B^{-1}$ is the forward shift operator. The last relation in (4.6) gives one way of computing covariances from the generating function, and could be used to compute $\mathrm{Var}(\hat{\theta}_t - \theta_t)$. Alternatively, $\gamma_{\hat{\theta}-\theta}(B)$ could simply be expanded 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 (4.5) 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 (4.4). $\mathrm{Var}(\hat{\theta}_t - \theta_t)$ 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.

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. (The variance results cited for the classical approach are only approximate in some cases, with the approximation error decreasing with the length of the series.) Thus, choice of approach depends on computational considerations, not on the results that will be obtained. Jones (1980), and Rao, Srinath, and Quenneville (1986) refer to the use of (4.4), or of (2.3) - (2.6), as stochastic least squares. This approach is rarely used in time series analysis because (1) given models for Y_t and θ_t there is a fair amount of work involved in solving for the elements of Σ_Y and Σ_θ, (2) since fairly long time series (e.g. T = 100) are often used in

practice the T×T matrix Σ_Y will be large, and (3) the matrix Σ_Y can be ill-conditioned if the correlation in Y_t is strong. These problems are worst if Y_t is nonstationary requiring differencing, in which case direct application of the stochastic least squares results would be virtually impossible. This approach would be more practical if the correlation in Y_t were mild and T were small. We stated results in this paper in this form to simplify the results and presentation, not because we advocate generally doing the computations this way.

Bell (1984) extended the classical signal extraction results (4.5) 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. If Y_t follows the ARIMA model $\phi(B)(1-B)Y_t = \eta(B)a_t$, then we use $\eta(B)\eta(F)\sigma_a^2/(1-B)(1-F)\phi(B)\phi(F)$ for $\gamma_Y(B)$ in (4.5) and similarly for $\gamma_\theta(B)$. If e_t does not require differencing then Y_t and θ_t will both require the same differencing operator and this will cancel in (4.5). 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.

The approach to signal extraction that is best may well depend on the problem at hand. The stochastic least squares results are the most general, but are difficult computationally unless T is small. 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

approximations 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. Examples

We now present some examples to illustrate the application of the time series approach to survey estimation to Census Bureau surveys. We have attempted to treat the examples realistically in the sense of recognizing and attempting to deal with aspects of the surveys that seem to be important in regard to applying the time series approach. This does not mean that we claim to have solved all the practical problems involved, even for the specific examples. The efforts here should be regarded as a first attempt. Also, some simplifications were made when these did not distract from the essentials of the problem.

5.1 Single Family Housing Starts

Housing starts are estimated monthly in the Survey of Construction (SOC) for various types of structures and geographic areas. Here we consider single family housing starts for the total U.S. and a breakdown into four geographic regions -- Northeast (NE), Midwest (MW), South (SO), and West (WE). We use time series of the usual survey estimates from January 1964 through December 1986 (276 observations). The following information on the SOC, and the variance and correlation estimates we use later, were provided to us by Jesse Pollock, Don Luery, and Dennis Duke of Construction Statistics Division (CSD) at Census.

The SOC uses a stratified three-stage cluster sample. First, Primary Sampling Units (PSUs) from the Current Population Survey (CPS) are stratified by demographic characteristics and by builiding permit activity in a recent year (1982 for the current sample). Some PSUs are selected with certainty, others are selected one PSU per stratum with probability proportional to the size of the 16 and over population. Within PSUs, places covered by building

permit systems and land areas not covered by building permit systems are identified. For the latter an area sample is used to estimate new construction. For the former building permit offices are sampled, then permits in these offices are sampled and the person listed on the permit is contacted to determine if building has started. The sampled permits are tracked monthly essentially until the unit has started. The sample is redrawn approximately every 5 years, and the PSUs are redrawn every 10 years in connection with CPS redesign (the last time in November 1984). An algorithm is used to maximize PSU overlap from old to new samples.

Estimation of housing starts in building permit places is done by computing start rates -- ratios of the number of units started in a month to the number of building permits issued in that month -- and applying these to estimates from the larger Building Permit Survey of the number of permits issued. Within each given type of structure and geographic area these figures are aggregated over the past 60 months (allowing for units started whose permits were issued as long as 5 years ago) to estimate total starts in building permit places in the current month. These figures are then added to estimates of starts in non-permit areas (which make up about 10% of starts over the total U.S.). Variances of sampling errors are estimated by collapsing pairs of strata and using the Keyfitz two-per-stratum method. Covariances of sampling errors at two time points can be obtained in an analogous fashion.

While the files for processing the SOC are large, they are not as large as for some other Census Bureau surveys. In fact, after each calendar year CSD constructs a file on tape with the preceeding year's data. They are also able to link data files across successive years to estimate sampling error lagged covariances and hence correlations. This was done for data from

January 1982 through December 1986, providing 60 variance estimates and 60-k estimates of lag k correlations for k = 1,...,23. Sampling error variances depend on the level of the series, but relative variances are more stable over time, suggesting that we consider whether the sampling errors may be assumed to be relative covariance stationary. However, the redrawing of the sample every 5 years may present a problem in this regard. Note that the last sample revision in November 1984 falls in the middle of the period for which we have sampling error variance and correlation estimates.

To investigate relative covariance stationarity of the sampling errors, log relative variances and correlation estimates for lags 1,...,6 and 12 were plotted over time for all four regions and the total U.S. Figure 1a and 1b show the log relative variance and lag-1 correlation for the Northeast region, with the point at the start of the new sample circled in 1a and the one lag-1 correlation crossing the transition from old to new sample circled in 1b. While there is considerable fluctuation over time in both plots, there are no apparent trends over time or obvious differences between the old and new samples. This behavior is typical of most of the variance and correlation plots at all lags. Averaging estimates separately for the old and new samples did not produce appreciably different results, and even time series analyses of correlation estimates for lags 1-3 suggested these estimates were not themselves autocorrelated. All this suggests assuming the relative variance and correlations remain constant over time, and averaging the individual estimates in some way to improve the estimates. The exception to this is the South region, which showed apparent differences in the old and new samples. (The total U.S. results reflected some of this effect of the South results.) This may be due to a change implemented in the new sample for handling non-permit places, which are much more prevalent in the South than in the

other three regions. At this point we are not sure what, if anything, to do about this. For purposes of illustration we shall go ahead and treat the South the same as the other regions.

Correlation estimates were averaged over time leaving out those correlations where one month was in the old and one month in the new sample, which are not comparable to the others. Relative variance estimates were produced in the spirit of maximum likelihood estimation for the lognormal distribution -- taking the average of the logs of the 60 relative variance estimates, adding one half of the sample variance of these 60 log estimates to this, and exponentiating the result. These results did not differ much from simply averaging the relative variances. (The one large outlier in February 1986 for the Northeast region, the 50th point in Figure 1a, was omitted from the calculation.) The resulting relative variance and correlation estimates for lags 1-12 are presented in Table 1. Notice that the estimates show very little correlation in the sampling errors. A possible explanation for this is that there may be some negative correlation in the errors of the estimates for permit places (if start rates are high in one month there are fewer unstarted units with permits in future months) and stronger positive correlation in the non-permit areas (if a non-permit area shows more or less than average construction activity in a month it is likely to do the same in neighboring months, and it will still be in the sample unless the sample has been redrawn). Since the majority of construction activity is in permit areas, this about washes out, leaving little or no correlation in the sampling errors. (This is not true in the new sample for the South, which has the most non-permit activity.) In what follows we shall assume the sampling errors are relative covariance stationary, with the lag-1 correlations for the Midwest and West regions the only nonzero correlations in the sampling errors.

The lack of correlation in the sampling errors simplifies some matters greatly. We might otherwise be concerned about the effects on the data of the sample revisions every 5 years, since lag-k correlations between months in two different samples could be quite different from those for months in the same sample. But there is no reason to expect correlations across sample revisions to be larger than correlations within the same sample, and since these are all small it seems we can ignore the sample revisions (again, except for the South). Also, since the sampling errors contribute almost uncorrelated noise to the time series Y_t of the usual survey estimates, we proceeded by modeling $\ln(Y_t)$ directly, rather than using a component model that explicitly allows for the sampling error. We have fairly long time series (276 observations) available for this. Although there may be some concern about our assumption of relative covariance stationarity of the sampling errors holding for the entire length of this time series, as long as we can assume the sampling error correlation is mild throughout we can use these long time series for modeling $\ln(Y_t)$. We could restrict the signal extraction to more recent data with little change in results if we were concerned about our assumptions on the sampling error (such as constant relative variance) holding into the distant past.

The time series of the usual survey estimates (Y_t) of single family housing starts for the four regions and total U.S. are plotted in Figures 2a-e (only the last 10 years are shown for clarity). Strong seasonality is evident for all five series. Examination of autocorrelation functions suggested the need for both regular and seasonal differencing, and suggested multiplicative seasonal moving average models all of the form

$$(1-B)(1-B^{12})\ \ln(Y_t) = (1 - \eta_1 B - \eta_2 B^2)(1 - \eta_{12} B^{12})\ a_t \qquad (5.1)$$

with η_2 nonzero only for the Northeast. Parameter estimates for the models

are given in Table 2. The estimate of η_{12} was 1 for the South, West, and total U.S, indicating cancellation of the operator $(1-B^{12})$ on both sides of (5.1), which is equivalent to using a model with a fixed seasonal (Abraham and Box 1978). These models were in fact then refit with seasonal indicator variables and used in this form, though we do not present the estimates of the seasonal mean parameters here. Examination of residuals from these models and residual autocorrelations suggested no major inadequacies in the models, though did suggest that one might consider using a different (higher) residual variance (σ_a^2) in the winter months. CSD adjusts all 5 time series for trading-day variation; we could incorporate trading-day terms in (5.1) as in Bell and Hillmer (1983). While we would generally do this at least for the South and total U.S., for which we found trading day was strongly significant, it was marginal to insignificant for the other regions so in the interest of simplifying the presentation we shall ignore it here.

Signal extraction estimates of the "true" level of housing starts θ_t were produced for each region and the total U.S., using the classical signal extraction results given in (4.5) to estimate $\ln(\theta_t)$ from $\ln(Y_t)$ and then exponentiating. The results are the dotted lines in Figures 2a-e. However, these are often difficult to distinguish from the original estimates because of the strong seasonality and the fact that the changes from the original estimates are often small. In Figures 3a-e we show the ratio of the signal extraction to the original estimates for the last 10 years of data, which are actually the estimates of $u_t^{-1} = \exp(-\ln(u_t))$. (The computations were actually done this way, estimating $\ln(u_t)$ from $\ln(Y_t)$ and using these results to estimate $\ln(\theta_t) = \ln(Y_t) - \ln(u_t)$.) We see the smallest adjustments are made to the South and U.S. estimates, larger adjustments to the Midwest and West, and the largest adjustments, often exceeding 5 and sometimes exceeding 10

percent, to the Northeast. We used results of Hillmer (1985) to estimate relative variances of the error in the signal extraction estimates. These are shown in Figure 4a-e, along with the (constant) usual sampling relative variance for comparison. Notice that the variances are higher right at the beginning and end of the series where we have little data before or after the time point for which we are estimating. Still the signal extraction relative variances are nearly constant over most of the length of the series. This constant is in fact $\text{Var}(\ln \hat{\theta}_t - \ln \theta_t) \equiv \nu_0 \approx \text{Rel Var}(\hat{\theta}_t - \theta_t)$ given by (4.5) and (4.6). Table 3 shows the usual relative sampling variances, $\gamma_u(0)$, ν_0, and their ratio, and analogous results in terms of coefficients of variation (CVs). We see the largest improvements in variance are for the Northeast and West, on the order of 30-35%, with improvements for the Midwest, South, and U.S. on the order of 15%. The significance of these improvements should also be judged in terms of how much error there was in the original estimates to begin with. The reductions in variance for the South and U.S. may not be worthwhile in light of how small the variances were to begin with, and the previously mentioned difficulties in applying the results to the South and U.S. The results for the Northeast and West seem much more worthwhile.

5.2 Teenage Unemployment

We now 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.

Sampling Error Model

The sampling error autocorrelations can be estimated from the detailed survey results in the same manner that the variances of the sampling errors are estimated. Train, Cahoon and Makens (1978) report the average autocorrelations for the teenage unemployed sampling errors based upon the disaggregated survey results between December 1974 and December 1975. These autocorrelations are reproduced in Table 4a. The autocorrelatons from the model

$$(1-\phi B)e_t = (1-\eta B)c_t \qquad (5.2)$$

with $\phi = .6$ and $\eta = .3$ are reported in Table 4b. 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 (5.2) 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 that may affect this correlation structure 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 4a, 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

$$\hat{\text{Var}}(e_t) = -.0000153\ Y_t^2 + 1.971\ Y_t \qquad (5.3)$$

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 (5.3) were provided by Donna Kostanich of the Statistical Methods Division. They were developed in 1977, about the middle

indicating that the correlation structure of the sampling errors can be approximated by the model

$$(1-.6B)e_t = (1-.3B)c_t \tag{5.6}$$

with the variance of e_t being given by equation (5.3). Since the composite estimates are design unbiased, it is reasonable to assume that the stochastic process $\{\theta_t\}$ is uncorrelated with the process $\{e_t\}$. It remains to specify the general form of the model for θ_t. In doing this we can be guided by the correlation structure of the observed data, Y_t, bearing in mind that the correlations of Y_t and its differences are determined by the structures of both θ_t and e_t.

The ACF of Y_t fails to die out, suggesting the need to first difference the data. The ACF of the first difference of Y_t exhibits a persistent periodic pattern suggesting the need for an additional seasonal difference of θ_t to achieve stationarity. The autocorrelations of $(1-B)(1-B^{12})Y_t$ are reported in Table 5. The most prominent features are the autocorrelations at lags 1 and 12 suggesting the model $(1-B)(1-B^{12})Y_t = (1-\alpha_1 B)(1-\alpha_{12}B^{12})a_t$ may be appropriate. Based upon these considerations, we tentatively assumed that a reasonable model to approximately describe the correlation structure of θ_t is

$$(1-B)(1-B^{12})\theta_t = (1-\eta_1 B)(1-\eta_{12}B^{12})b_t. \tag{5.7}$$

This choice is also based upon our experience that models of this form have proven to be appropriate in describing the correlation structure of many seasonal time series.

Time Series Model Estimation and Checking

Thus, we take as a model for the observed Y_t;

$$\begin{aligned} &(1)\quad Y_t = \theta_t + e_t \\ &(2)\quad (1-B)(1-B^{12})\theta_t = (1-\eta_1 B)(1-\eta_{12}B^{12})b_t \\ &(3)\quad (1-.6B)e_t = (1-.3B)c_t \end{aligned} \tag{5.8}$$

of our time series, and so are reasonable for use with our data. Slightly different coefficients may be more appropriate for more recent data. Thus, the estimated variance of the error is a nonlinear function of the estimated level and is not constant over time. This creates the need to consider parameter estimation and signal extraction methods that allow for a changing sampling error variance. One way to deal with this problem is by use of a Kalman filter algorithm to evaluate the likelihood for its maximization for parameter estimation, and to use a Kalman smoother to compute the signal extraction estimates.

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

$$\underset{\sim}{X}_{t+1} = F\underset{\sim}{X}_t + G\underset{\sim}{v}_t \tag{5.4}$$

$$Y_t = H_t\underset{\sim}{X}_t \tag{5.5}$$

(Note that in our problem there is no added error in equation (5.5)). Then given observations $Y_1,\ldots,Y_n$ one can use the Kalman filter algorithm to evaluate the likelihood function (see Jones, 1985) and use a standard nonlinear optimization routine to find the parameters that maximize the likelihood function. In our particular case, the matrix H_t in the measurement equation (5.5) 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.

Time Series Model Identification

In this example we are assuming that the observed values, Y_t, are the sum of a true value θ_t and a sampling error e_t. We have external information

with $\{\theta_t\}$ uncorrelated with $\{e_t\}$ and $\text{Var}(e_t)$ given in (5.3). 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. This particular model has the unusual features that one of the component models has known parameter values and a known variance that is changing over time. Because of these features, as discussed earlier, the likelihood function was 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 reason to question the model. The autocorrelations of the residuals reported in Table 6 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.

Signal Extraction Estimates

A Kalman fixed interval smoother was used to compute the signal extraction estimates and their variances, using the model (5.8) with the estimated parameters, and equation (5.3) for $\text{Var}(e_t)$. The signal extraction estimates are plotted along with the usual composite estimates for the last 100 observations in Figure 5a. The seasonal difference $(1-B^{12})$ of the signal extraction estimates and the seasonal difference of the composite estimates are plotted in Figure 5b. It is apparent from these figures that there is a difference in these two estimates and that the signal extraction estimates are smoother than the composite estimate.

The standard errors of both the last 100 signal extraction estimates and

the last 100 composite estimates are plotted in Figure 6a. 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 6b 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 6a and 6b 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.

This example shows some of the complexities involved with applying signal extraction ideas when the variance of the sampling errors is known to be changing over time. The example also illustrates the potential improvements that are possible by incorporating time series methods in estimating the monthly values for this important time series.

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.

Chung, K. L. (1968), A Course in Probability Theory, New York: Harcourt, Brace and World, Inc.

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.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," Journal of the American Statistical Association, (with discussion), 78, 776-807.

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.

SINGLE FAMILY HOUSING STARTS
Four Regions and Total U.S.

Table 1
Averaged Relative Sampling Variance and Lag Correlations

		Lag						
Region	**Rel** Var	1	2	3	4	5	6	12
NE	.007617	.038	.066	-.067	.008	-.019	.063	.089
MW	.002829	-.136	.011	.014	.031	.021	.035	.047
SO	.001366	-.027	.031	.107	.073	.080	.112	.116
WE	.002534	-.208	-.043	-.021	-.032	.004	.010	-.002
US	.000706	-.038	.017	.031	.027	.029	.080	.084

Table 2
Parameter Estimates for Time Series Model (5.1)

Region	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_{12}$	$\hat{\sigma}_a^2$
NE	.48 (.06)	.15 (.06)	.79 (.04)	.03267
MW	.44 (.06)	-	.89 (.03)	.03904
SO	.38 (.06)	-	1.00 -	.01096
WE	.29 (.06)	-	1.00 -	.01739
US	.21 (.06)	-	1.00 -	.00837

Table 3
Original and Signal Extraction Estimates
Relative Variances and Coefficients of Variation

	$\gamma_u(0)$	ν_0	$\nu_0/\gamma_u(0)$	CV = $\sqrt{\gamma_u(0)}$	$\sqrt{\nu_0}$	$\sqrt{\nu_0}/CV$
NE	.007617	.005020	.66	.087	.071	.81
MW	.002829	.002470	.87	.053	.050	.93
SO	.001366	.001118	.82	.037	.033	.90
WE	.002534	.001744	.69	.050	.042	.83
US	.000706	.000607	.86	.027	.025	.93

Table 4a
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 4b
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

Table 5
Teenage Unemployment -- Autocorrelations of $(1-B)(1-B^{12})\ Y_t$

Lag	1	2	3	4	5	6	7	8	9	10	11	12
Correlation	-.34	.03	-.02	.04	-.03	-.01	-.08	.24	-.10	.09	.04	-.32

Lag	13	14	15	16	17	18	19	20	21	22	23	24
Correlation	.16	-.07	.07	-.17	.04	.08	.14	-.20	.06	.02	.07	-.17

Table 6
Teenage Unemployment -- Autocorrelations of Standardized Residuals from Model (5.8)

Lag	1	2	3	4	5	6	7	8	9	10	11	12
Correlation	-.02	.04	.04	.09	.02	.11	.12	.16	-.01	.08	-.11	.08
Std. Error	.09	.09	.09	.09	.09	.09	.09	.09	.09	.09	.09	.09

Lag	13	14	15	16	17	18	19	20	21	22	23	24
Correlation	-.05	-.03	.09	-.13	-.02	.09	.13	-.11	.00	-.03	-.02	-.10
Std. Error	.09	.09	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10

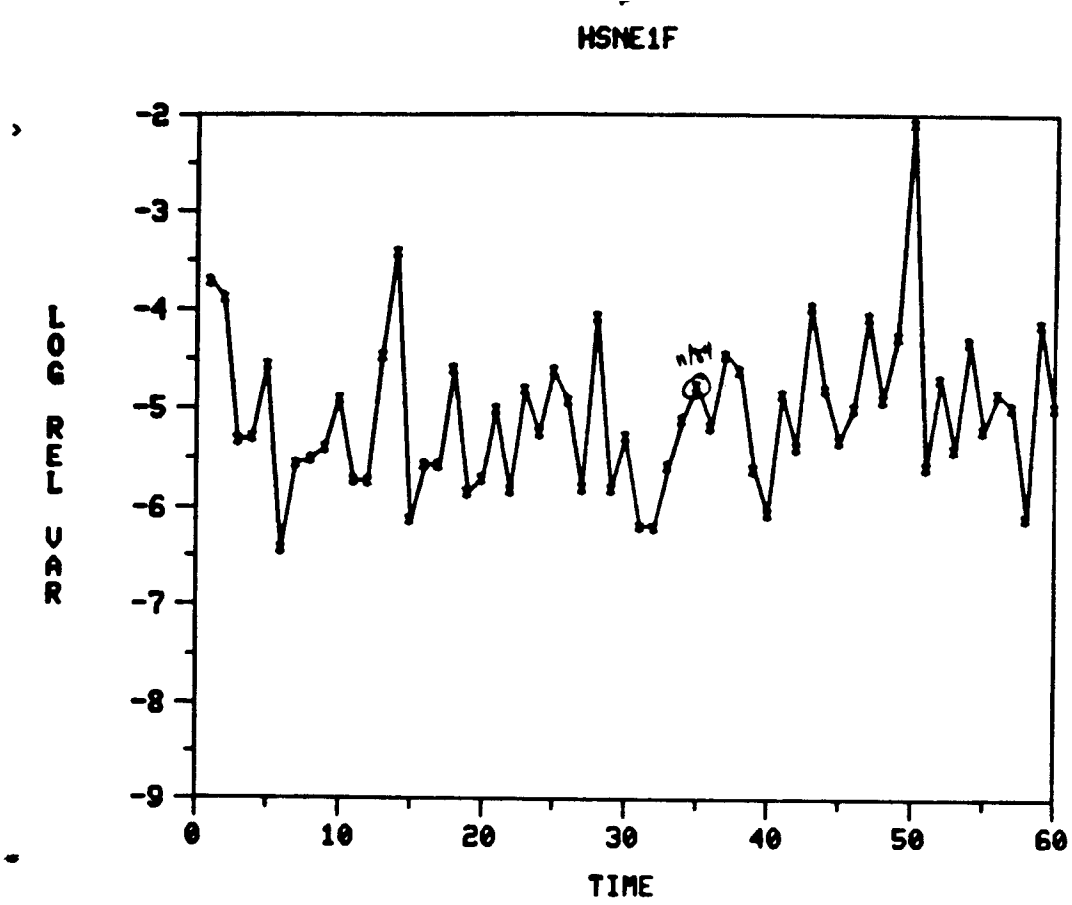

Figure 1a.

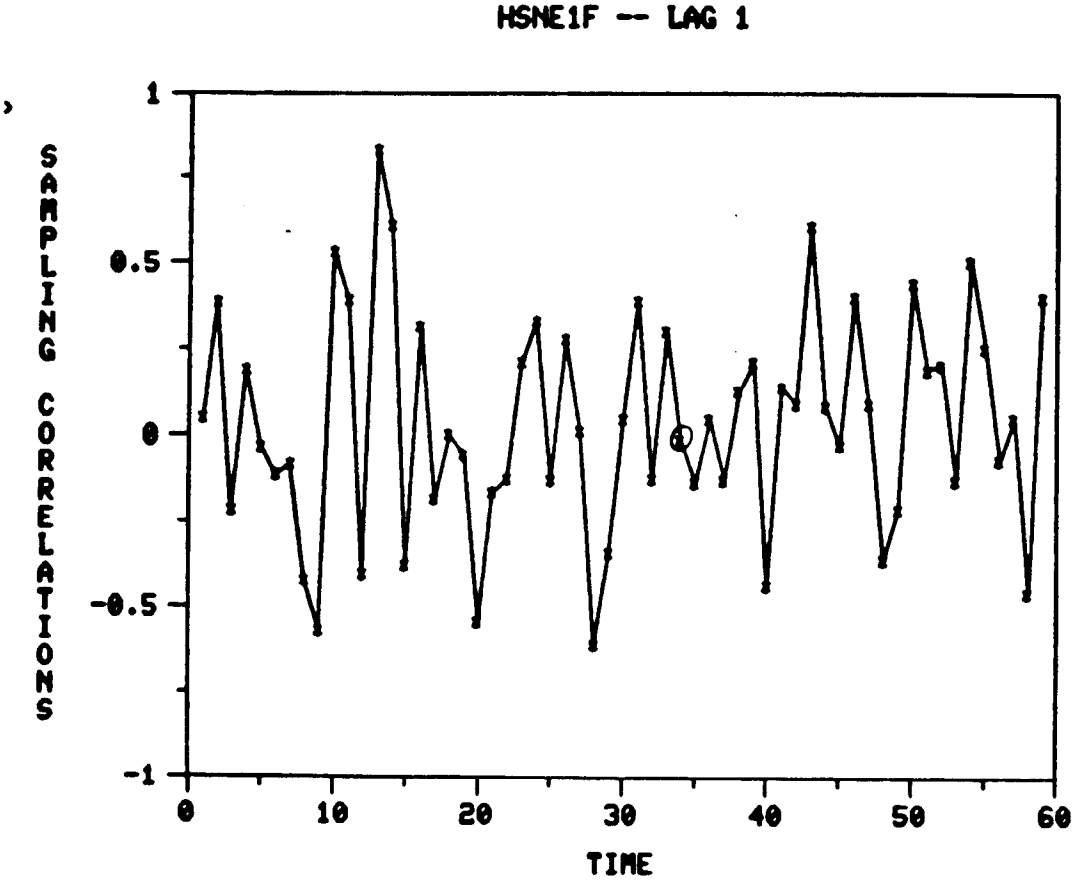

Figure 1b.

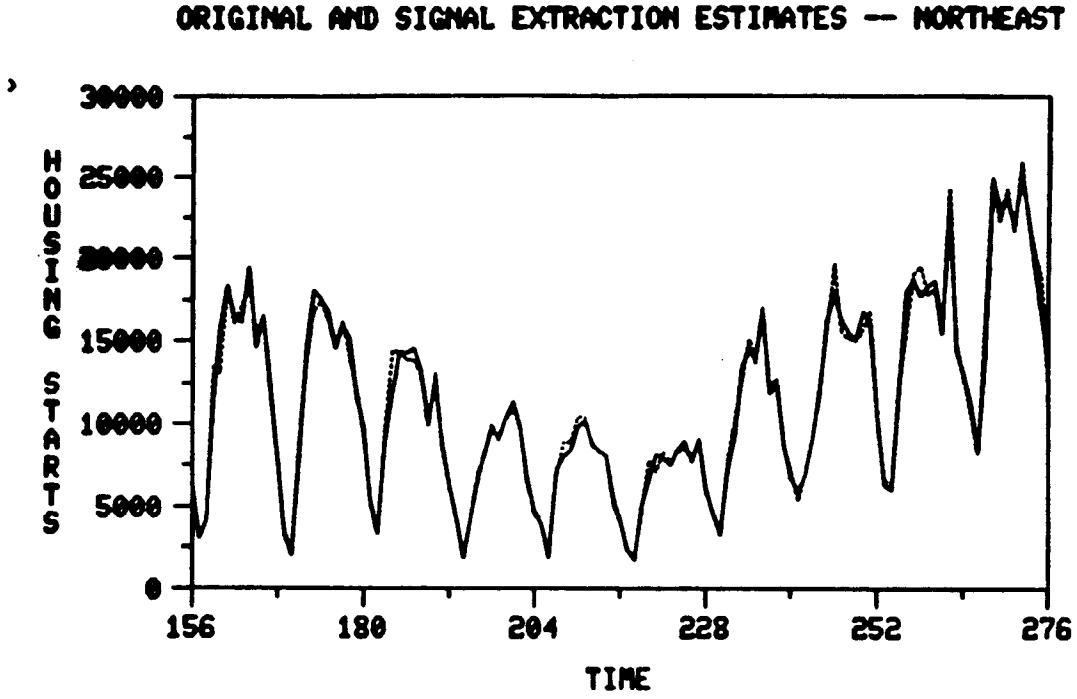

Figure 2a.

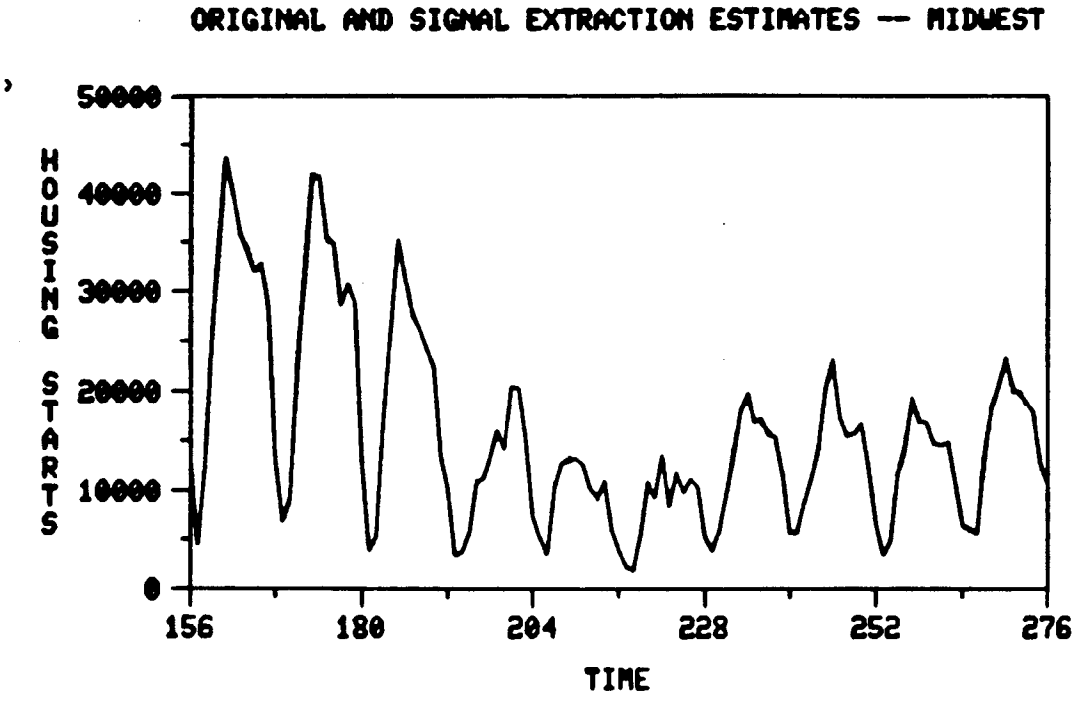

Figure 2b.

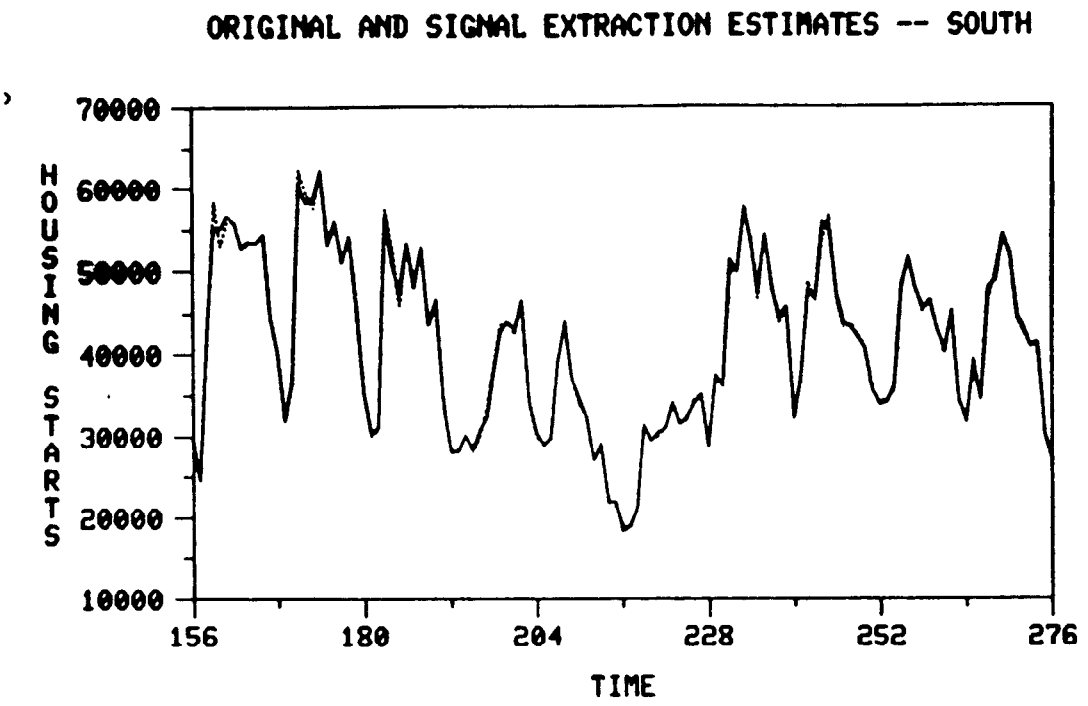

Figure 2c.

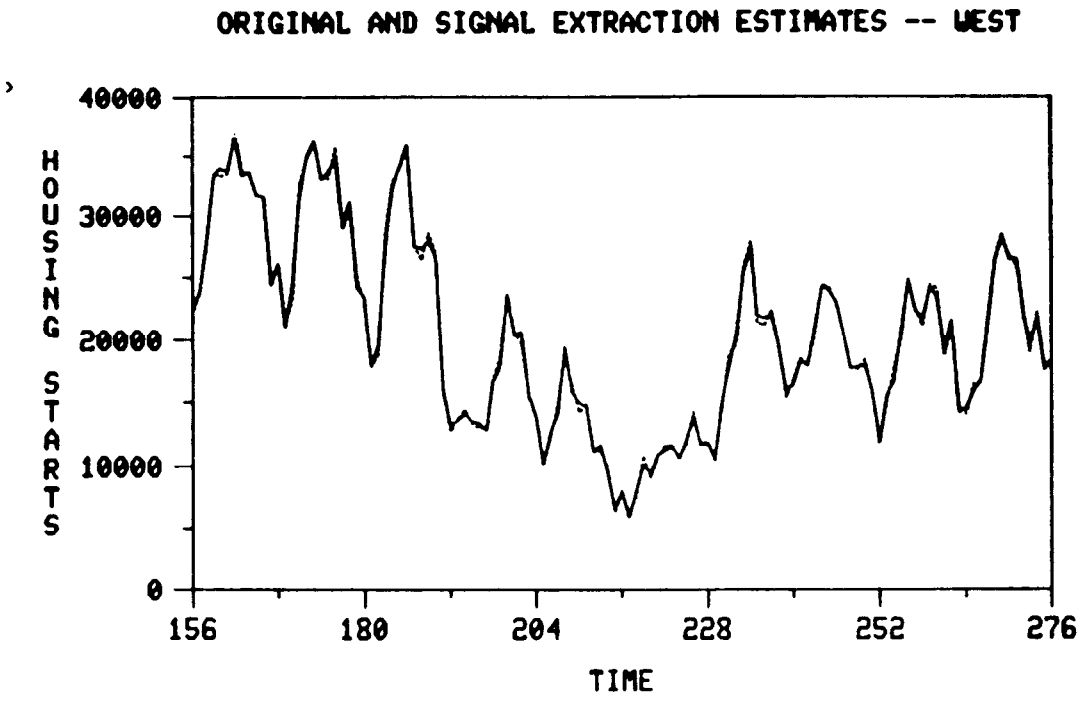

Figure 2d.

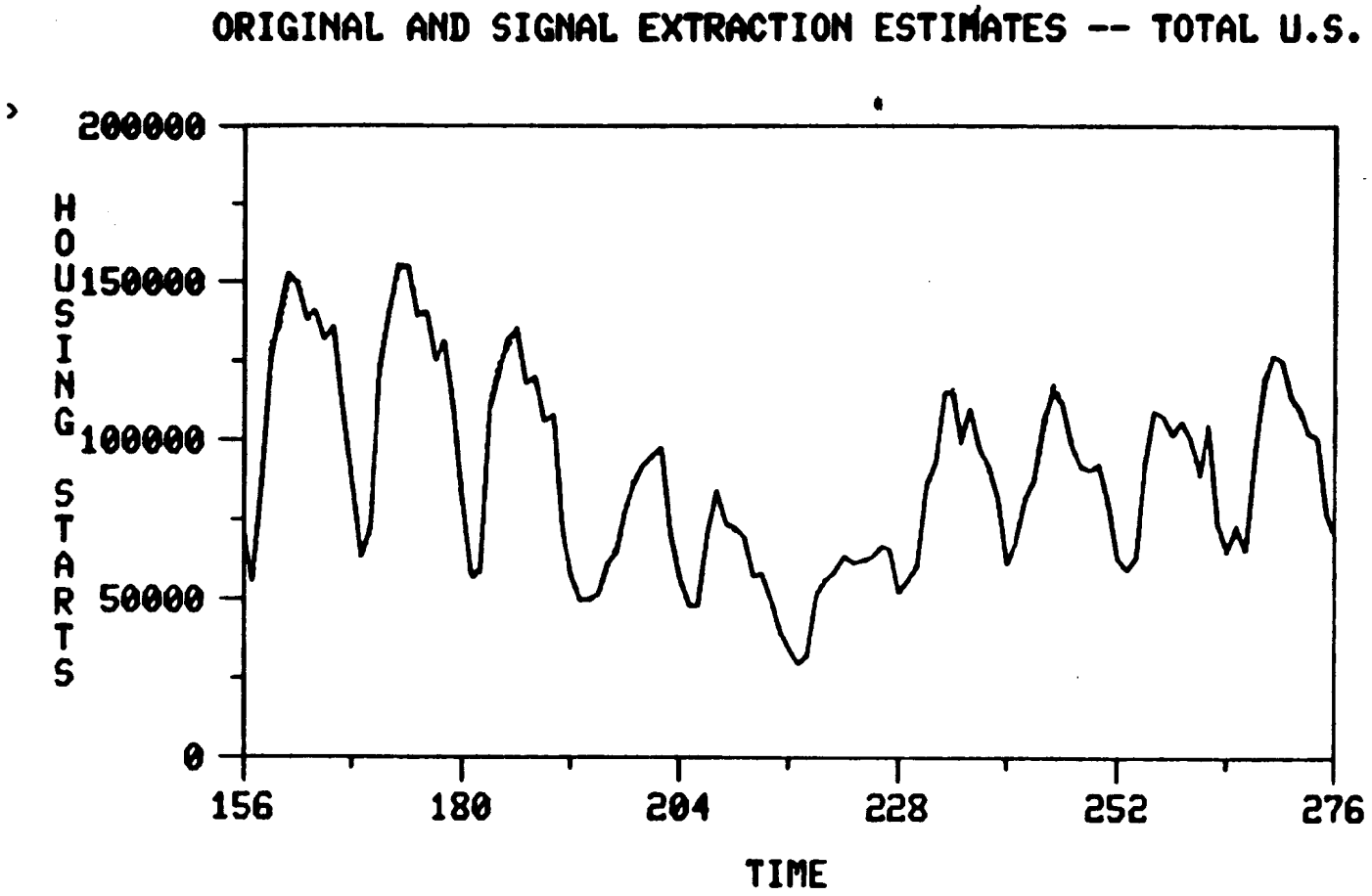
ORIGINAL AND SIGNAL EXTRACTION ESTIMATES -- TOTAL U.S.
HOUSING STARTS
200000
150000
100000
50000
0
156
180
204
228
252
276
TIME

Figure 2e.

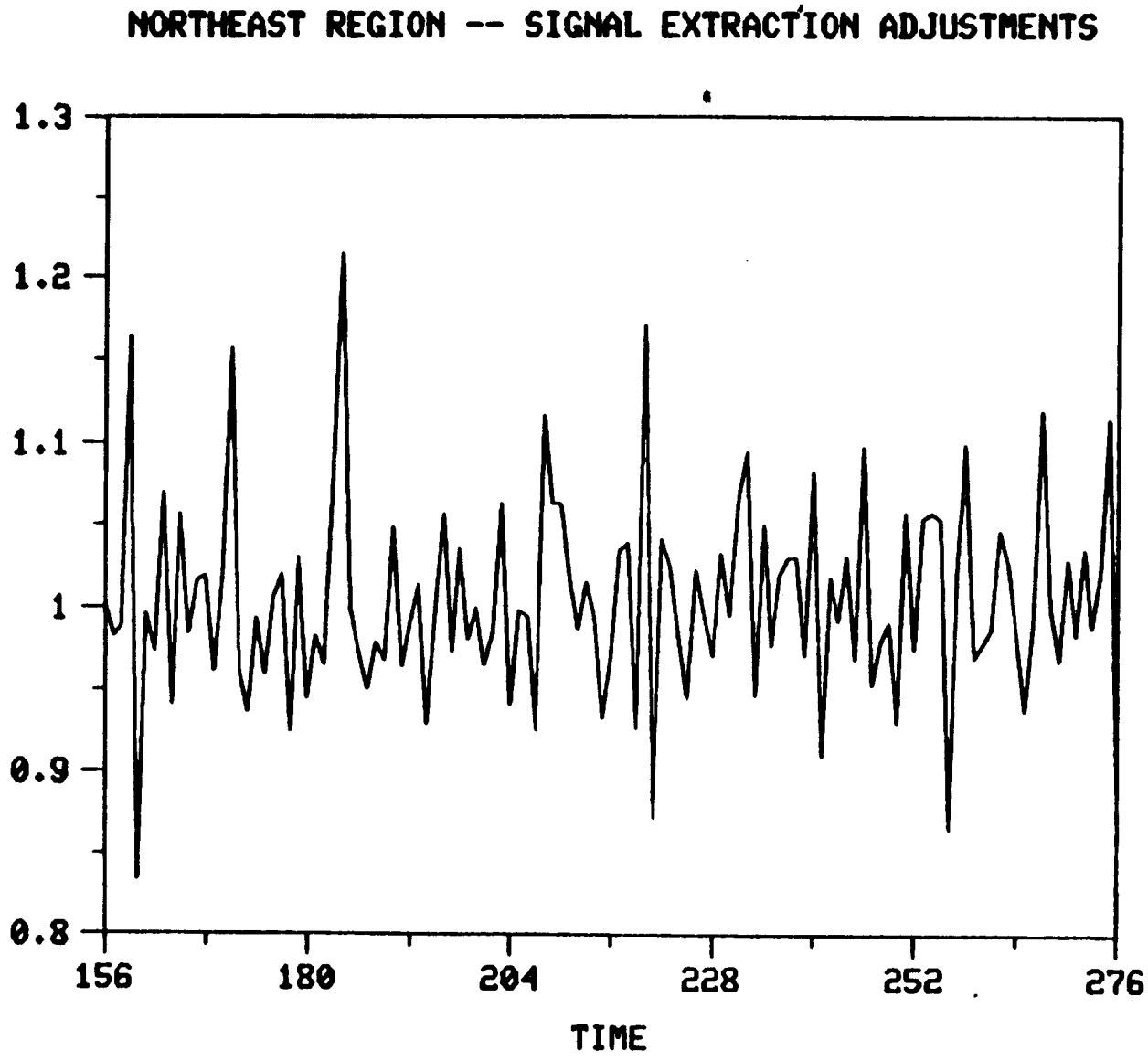
NORTHEAST REGION -- SIGNAL EXTRACTION ADJUSTMENTS
1.3
1.2
1.1
1
0.9
0.8
156
180
204
228
252
276
TIME

Figure 3a.

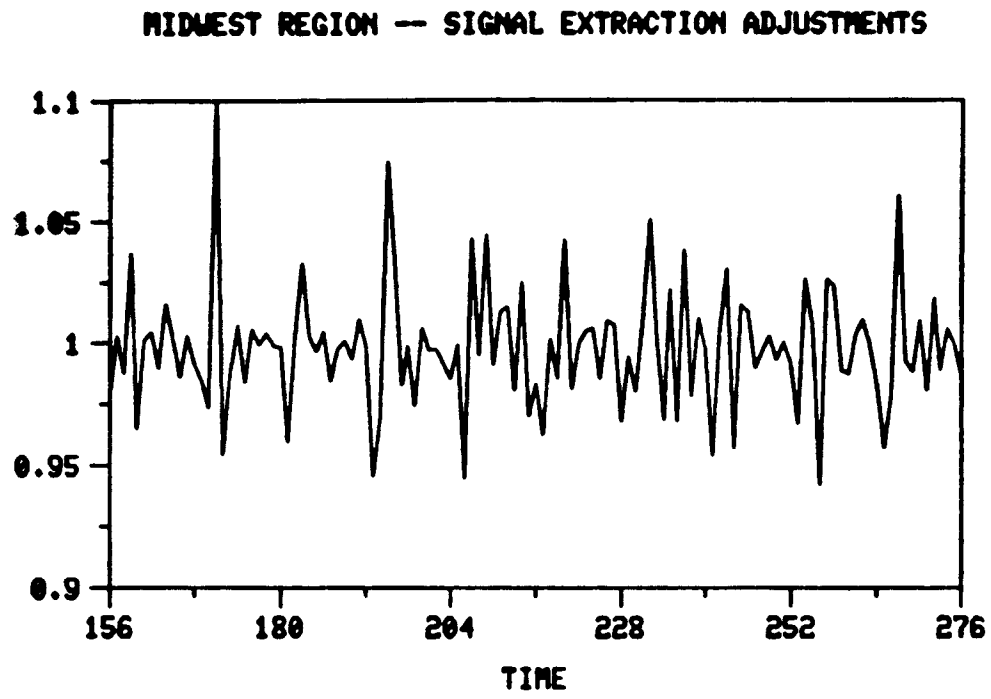

Figure 3b.

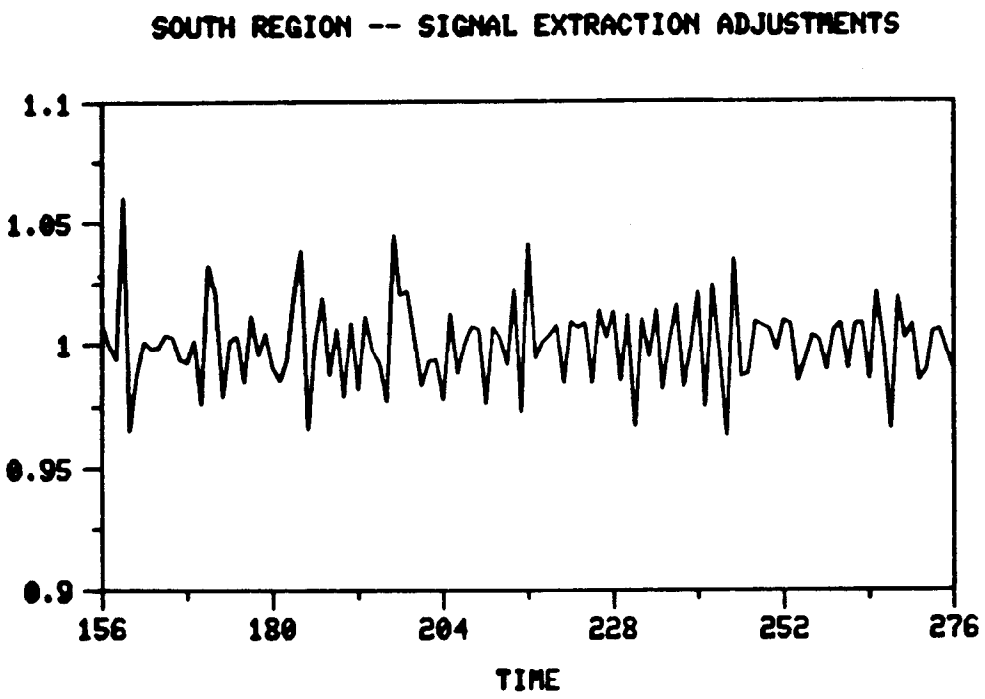

Figure 3c.

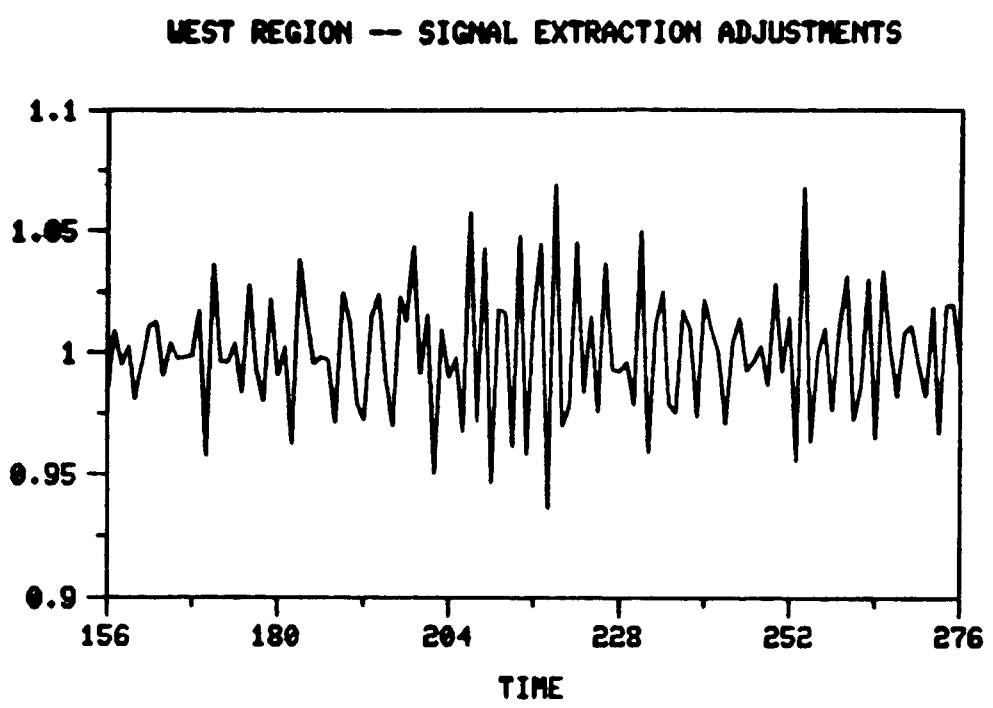

Figure 3d.

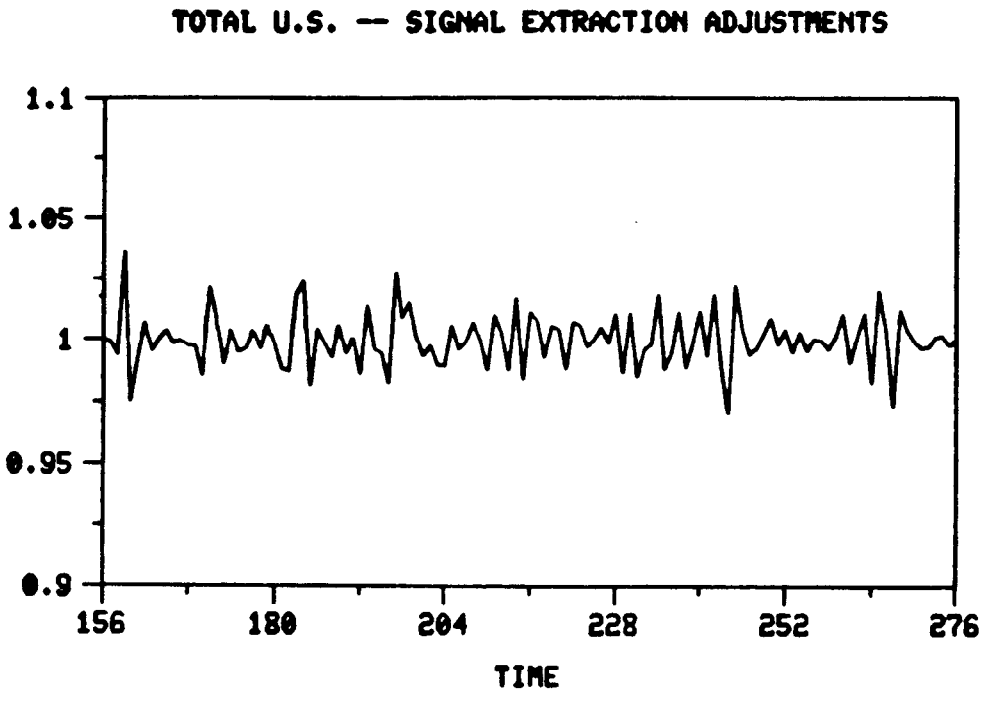

Figure 3e.

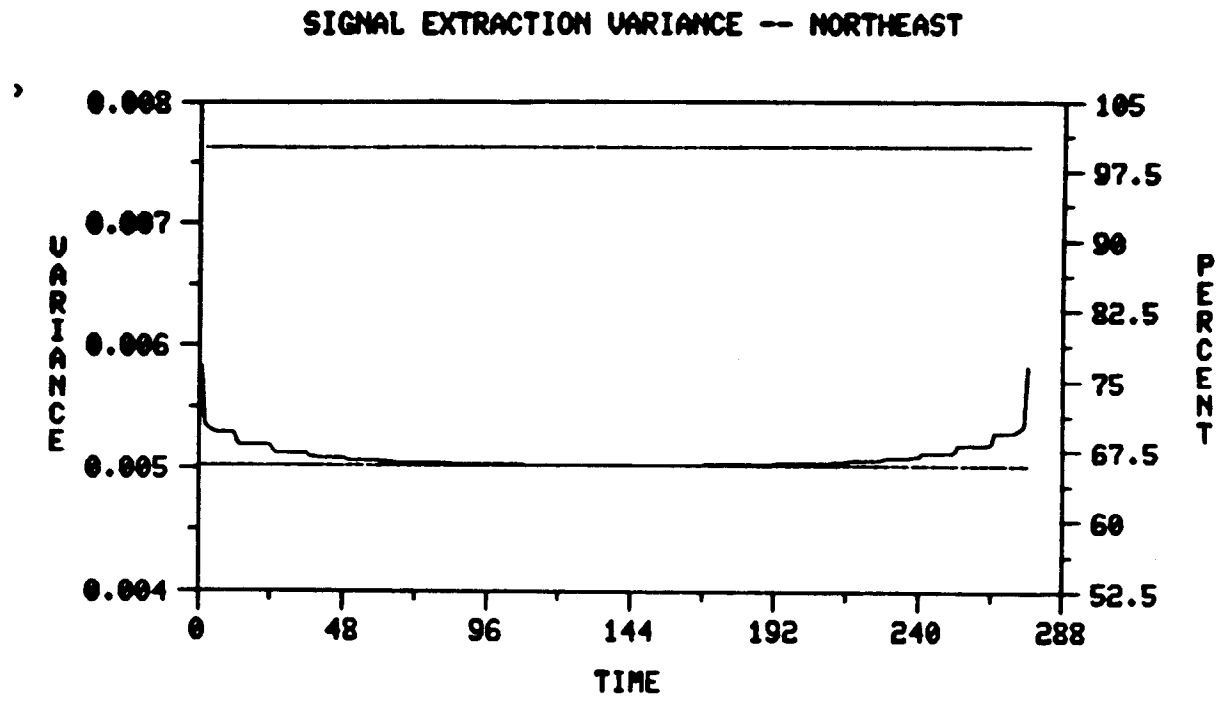

Figure 4a.

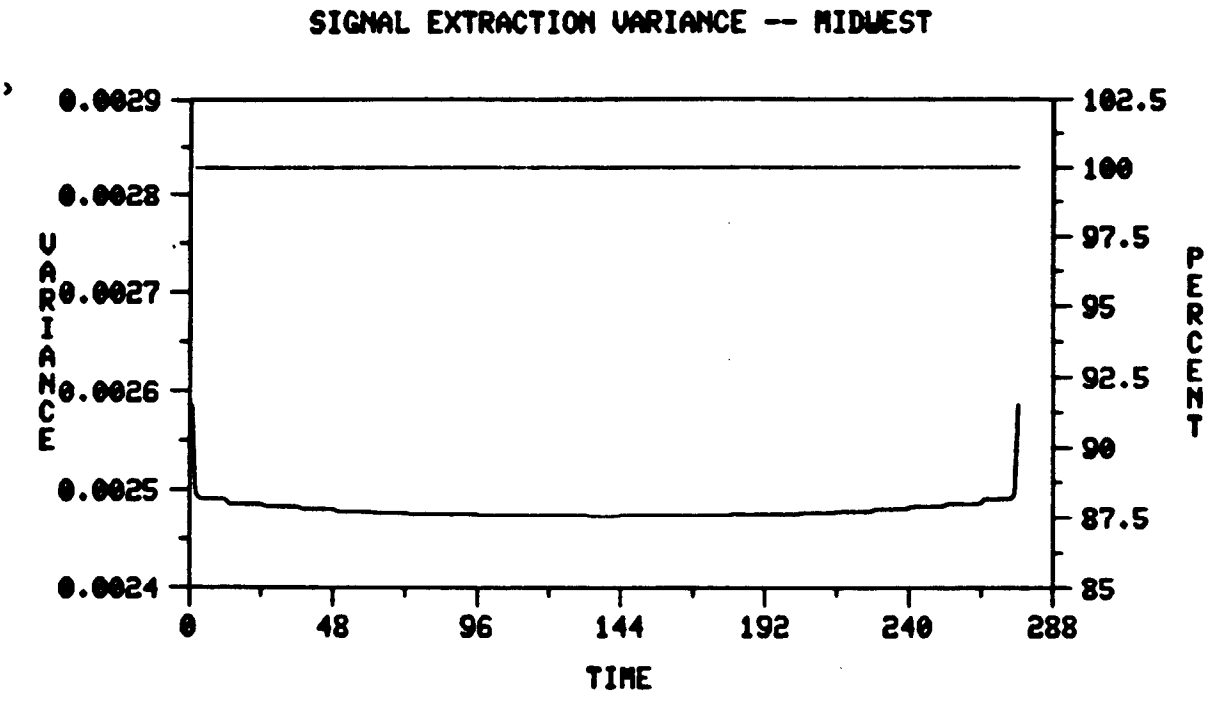

Figure 4b.

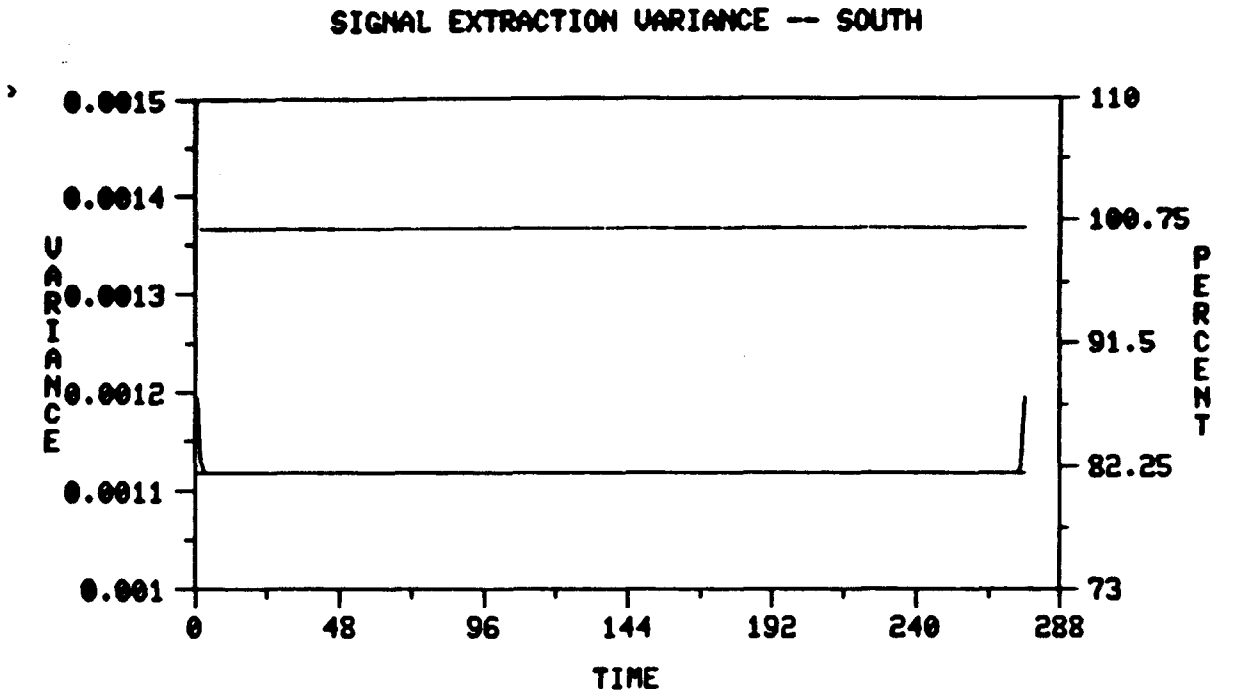

Figure 4c.

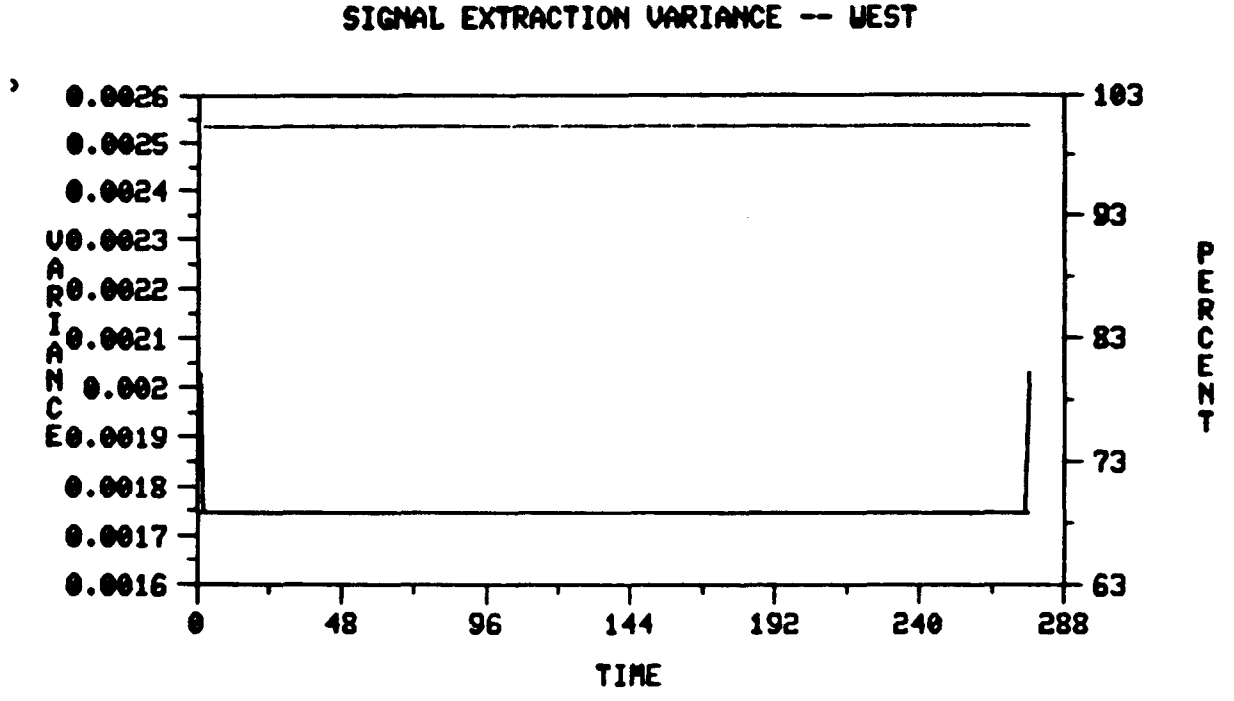

Figure 4d.

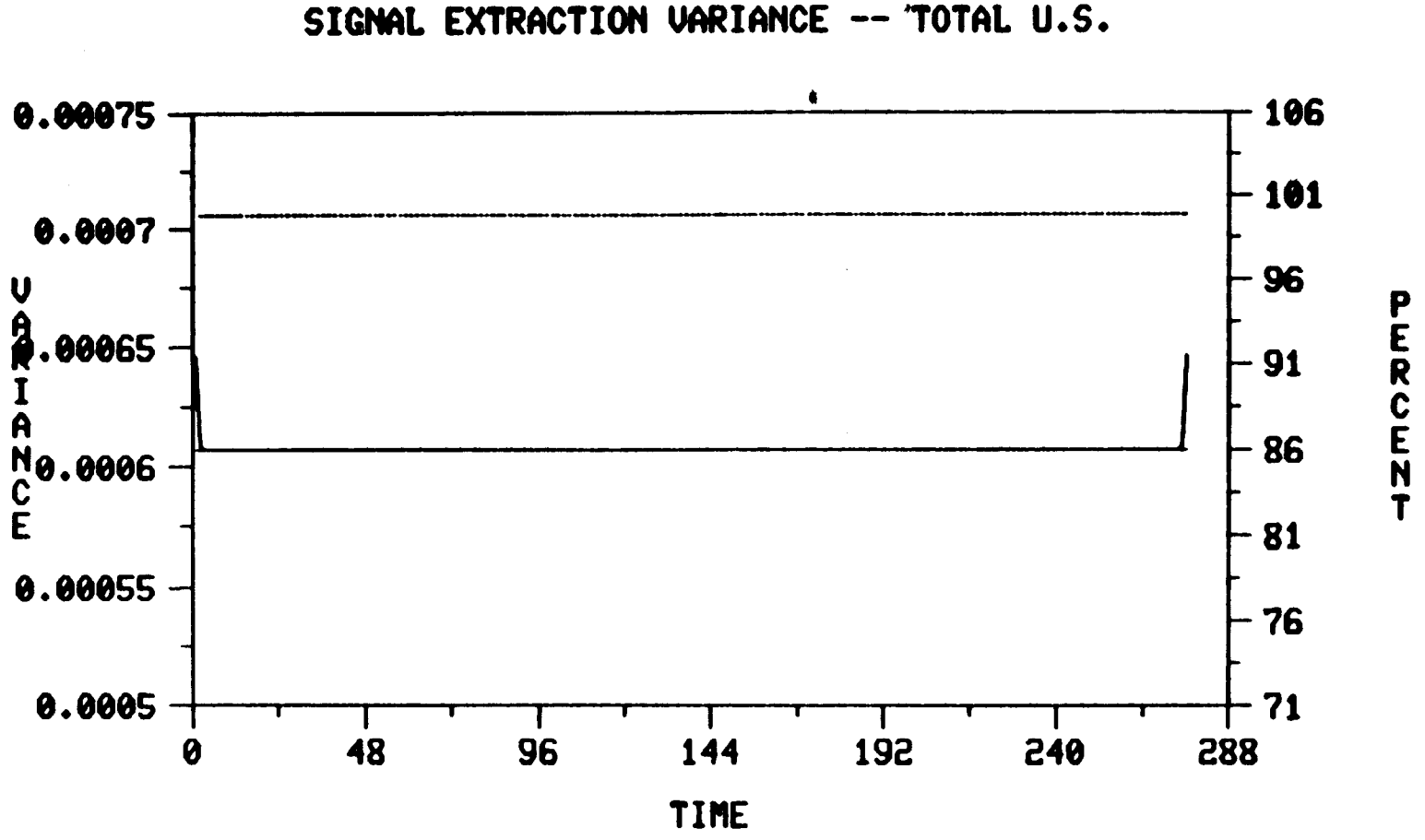
SIGNAL EXTRACTION VARIANCE -- TOTAL U.S.
0.00075
0.0007
0.00065
0.0006
0.00055
0.0005
VARIANCE
106
101
96
91
86
81
76
71
PERCENT
0
48
96
144
192
240
288
TIME

Figure 4e.

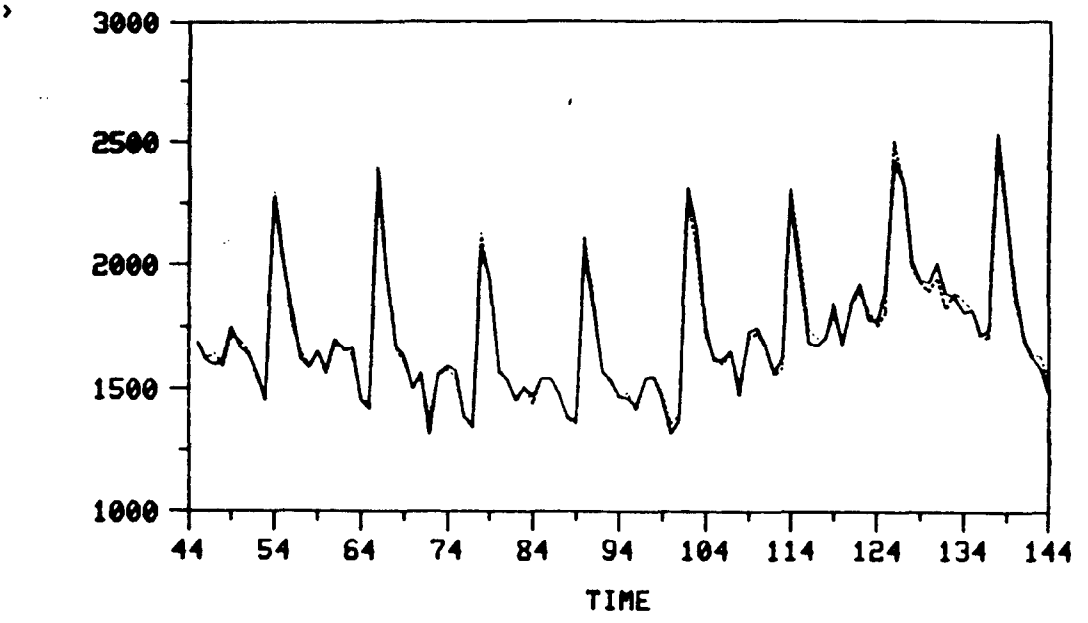

Figure 5a.

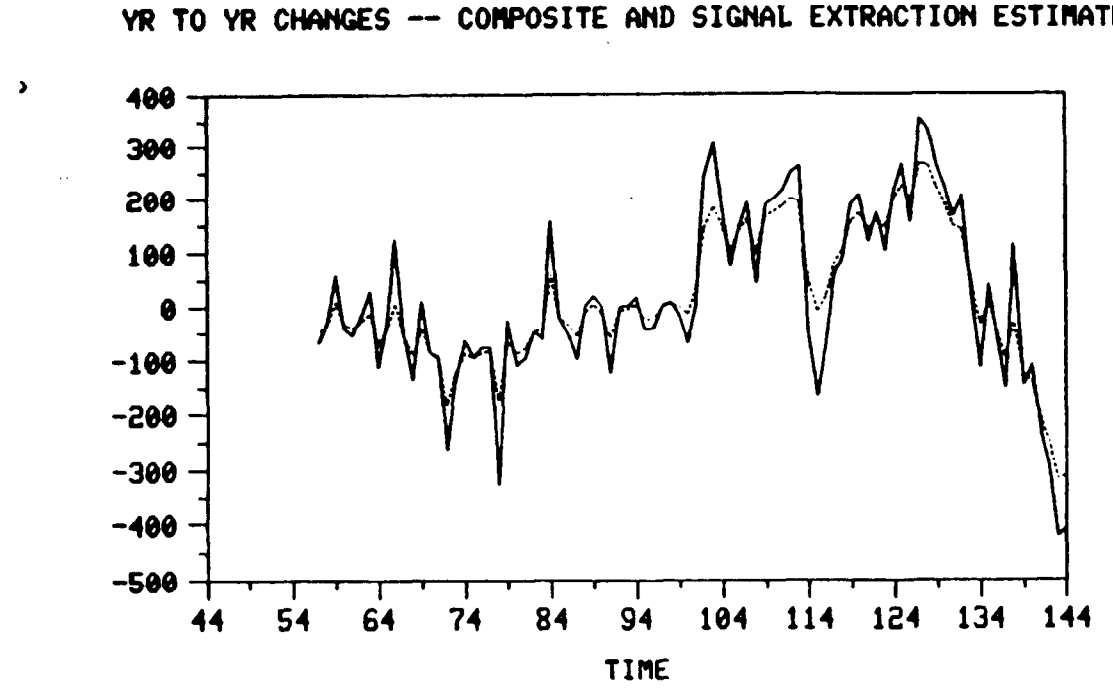

Figure 5b.

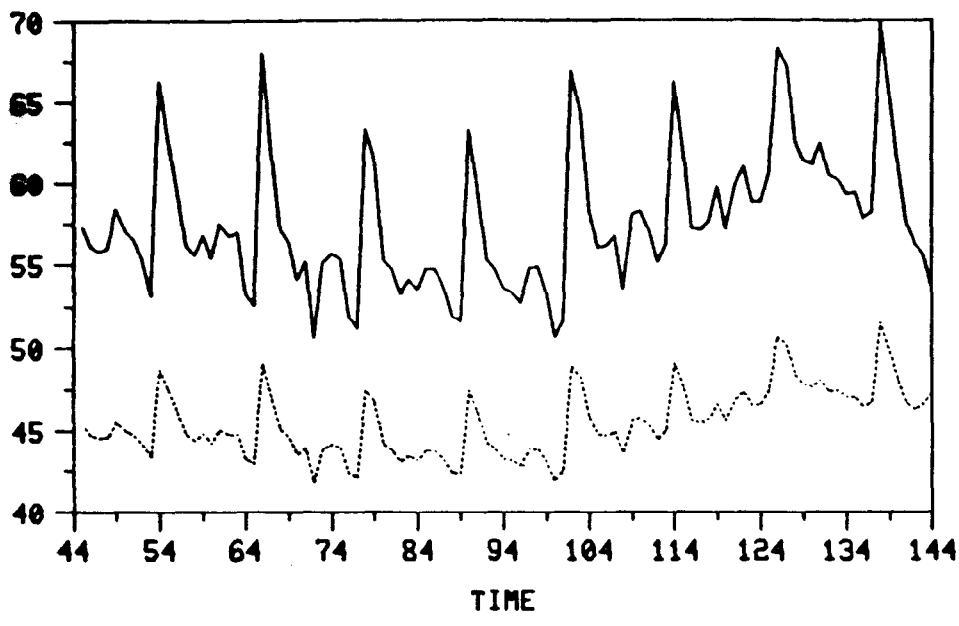

Figure 6a.

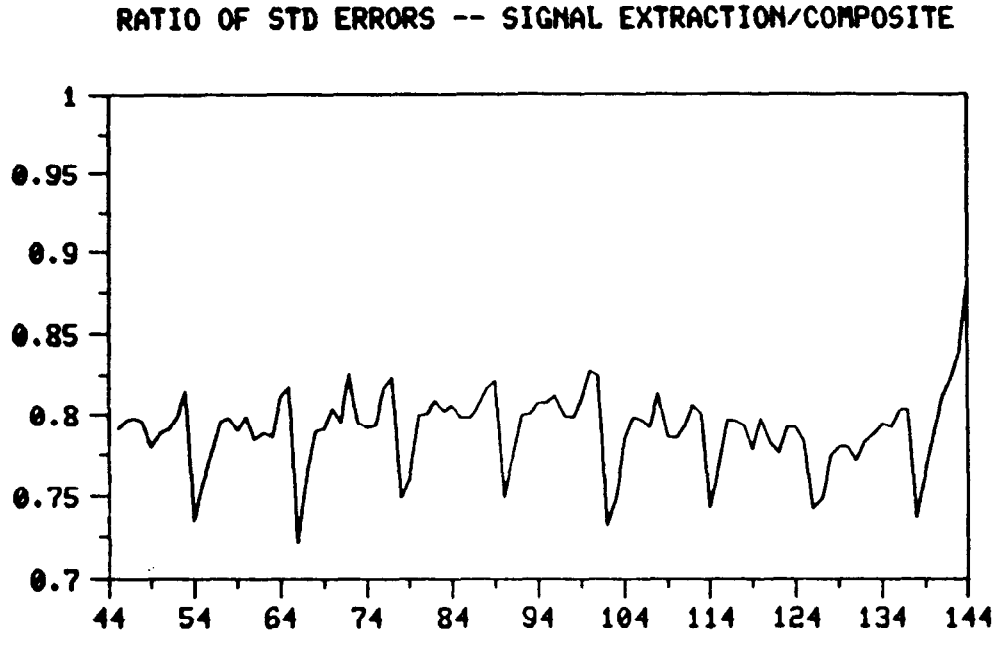

Figure 6b.