SIGNAL EXTRACTION EXCLUDING CERTAINTY CASES FOR ESTIMATION OF MONTHLY RETAIL TRADE TIME SERIES

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

1. INTRODUCTION

Papers by Scott and Smith (1974), Scott, Smith, and Jones (1977), and R. G. Jones (1980), suggested the use of signal extraction results from time series analysis to improve estimates in periodic surveys. Given models for the true unobserved time series (population quantities) and the sampling errors, these results produce estimates of the population quantities that have minimum mean squared error among estimates that are linear functions of the observed time series of survey estimates.

This paper focuses on modeling and signal extraction to improve estimates for repeated surveys that involve certainty cases. The series we use come from the Census Bureau's Retail Trade Survey (RTS), a short discussion of which can be found in Section 2. Section 3 discusses the component modeling of the time series from the RTS. Section 4 covers signal extraction results for some series from the RTS. Then in section 5, we discuss applying signal extraction techniques to time series with certainty cases excluded, and the potential benefits of doing this. In Section 6 we apply these latter procedures to the retail trade time series.

2. THE RETAIL TRADE SURVEY (RTS)

The Census Bureau's Retail Trade Survey (RTS) produces monthly estimates of sales for detailed kinds of retail businesses at the U.S. and regional levels, and for less detailed kinds of retail businesses for some states and metropolitan areas. For this paper, data at the U.S. level are used.

The RTS has a panel of large businesses that are selected into the sample with certainty and report sales every month. In addition, three rotating list panels of smaller businesses are selected into the sample by stratified simple random sampling. (There is also an area sample to cover businesses not within the list frame. Because of its generally small contribution to total sales, we shall not consider the area sample separately here.) Each rotating list panel reports current month and previous month sales at intervals of three months. From these reports Horvitz-Thompson (HT) estimates of current and previous months sales are constructed. From the HT estimates, composite estimates are constructed as described in Wolter (1979). Sampling variances are estimated by the random group method (Wolter 1985, ch. 2) using sixteen random groups. Further information on the survey is given in Isaki, et al. (1976), Wolter, et al. (1976), and Wolter (1979). Bell and Wilcox (1990) give a summary discussion focusing on aspects of the survey relevant to time series properties of the estimates.

There are several complicating factors in the survey. First, the sample is redesigned and independently redrawn about every five years, with new samples having been introduced in September, 1977, and January of 1982 and 1987. Signal extraction and sampling error modeling for the RTS are addressed in Bell and Hillmer (1990a) and Bell and Wilcox (1990). When a new sample is introduced, there is a three month transition period where the composite estimates are not used, but this problem will not be addressed in this paper. Finally, the monthly estimates are benchmarked to annual totals estimated from an annual survey and from the economic census taken every five years. To avoid this complication, we use data that are not benchmarked. The reader should be aware, however, that for this reason, the data used here do not agree with published results.

The series used in this paper are as follows:
1) Retail Sales of Household Applicance Establishments
2) Retail Sales of Men's and Boy's Clothing Establishments
3) Retail Sales of Radio and TV Establishments

Graphs of the series can be found in Figure 1.

3. COMPONENT MODELING

Let Y_t denote the time series of the usual (composite) survey estimates, S_t denote the population quantities (signal) being estimated, and N_t denote the sampling error in Y_t as an estimate of S_t. The basic decomposition is

$$Y_t = S_t + N_t \quad . \tag{3.1}$$

For the purposes of this paper, we will use the log-additive decomposition,

$$\ln(Y_t) = \ln(S_t) + \ln(U_t) \tag{3.2}$$

where $U_t = 1 + N_t/S_t$. Reasons for using $\ln(Y_t)$ will become clear in the next section. Also, see Bell and Hillmer (1990a).

3.1. MODELING THE SIGNAL COMPONENT, S_t, FOR THE RTS

Since autoregressive-integrated-moving average (ARIMA) models have often been used successfully for the analysis of the observed series, Y_t, from periodic surveys with little or no sampling error, it seems likely that they should also prove useful for modeling S_t. In addition, experience with modeling time series Y_t suggests that dealing with nonstationarity in S_t will be very important. Nonlinear transformations, differencing, and use of regression mean functions can be quite useful for dealing with the usual types of nonstationarity.

Therefore, models for S_t (assuming a logrithmic transformation) can be written in the form,

$$\delta(B)[\ln(S_t)-\mu_t] = [\theta_s(B)/\phi_s(B)]b_t \tag{3.3}$$

where $\delta(B)$, $\phi_s(B)$, and $\theta_s(B)$ are the differencing, autoregressive (AR), and moving average (MA) operators, respectively; μ_t is the regression term (e.g., trading day regression), and b_t is a white noise series (iid $N(0,\sigma_b^2)$). ARIMA models for Y_t can be used as starting points in developing models for S_t.

3.2. MODELING THE SAMPLING ERROR COMPONENT, N_t, FOR THE RTS

The first step in modeling N_t is to estimate the

sampling error covariances over time, $Cov(N_t, N_{t+k})$. In principle, this is the same problem as estimation of sampling variances (k=0), which is routinely done for periodic surveys and for which many methods are available (Wolter, 1985).

The next step is to develop a model for the correlation structure of the sampling errors. One possible model that takes into account the rotating panel nature of the survey and the form of the composite estimates is

$$(1-.75B)(1-\phi^3B^3)(1-\Phi B^{12})\ln(U_t) = (1-\eta B)c_t \quad (3.5)$$

where, again, $U_t = 1 + N_t/S_t$, $c_t \sim N(0, \sigma_c^2)$. This model is developed in Bell and Hillmer (1990a), and also in Bell and Wilcox (1990), who give the parameter values associated with the sampling error components for the three retail trade series considered here.

4. MODEL ESTIMATION AND SIGNAL EXTRACTION

Three general approaches to deriving time series results and doing computations might be called the classical approach, the matrix approach, and the Kalman filter approach. For this paper, the Kalman filter approach (see Anderson and Moore (1979) for a general discussion) was used because it is particularly convenient for handling component models with such features as changing variances over time. The problem of initializing the Kalman filter for nonstationary models is discussed in Bell and Hillmer (1990b). Model estimation proceeds by maximum likelihood assuming normality, and the Kalman filter can be used to evaluate the likelihood as suggested by R. H. Jones (1980). For a more complete discussion, see Bell and Hillmer (1989).

Taking models of the form (3.3) for $\ln(S_t)$ with models of the form (3.5) for $\ln(U_t)$ taken as fixed, the parameters of the models for $\ln(S_t)$ were estimated. Estimation of these models produced the following:

1) Household Appliance Establishments

$$(1-B)(1-B^{12})(\ln(S_t)-\Sigma\beta_i T_{it}) = (1-.43B-.12B^2)(1-.52B^{12})b_t \quad (4.1a)$$

$$\sigma_b^2 = 1{,}475 \times 10^{-6}$$

2) Men's and Boy's Clothing Establishments

$$(1-B^{12})(1-.60B-.24B^2)(\ln(S_t)-\Sigma\beta_i T_{it}-\Sigma\alpha_i h(t,t_i)) = (1-.14B^{12})b_t \quad (4.1b)$$

$$\sigma_b^2 = 856 \times 10^{-6}$$

3) Radio and TV Establishments

$$(1-B)(1-B^{12})(\ln(S_t)-\Sigma\beta_i T_{it}) = (1-.02B)(1-.70B^{12})b_t \quad (4.1c)$$

$$\sigma_b^2 = 1{,}377 \times 10^{-6}$$

$\Sigma\beta_i T_{it}$ and $\Sigma\alpha_i h(t,t_i)$ are regression functions for trading-day and Easter holiday variation, respectively. The Q statistics associated with each of the estimated models were reasonable.

These estimated models, (4.1a,b,c) with the corresponding sampling error models of form (3.5), were used to produce signal extraction estimates of $\ln(S_t)$, which were then exponentiated to produce estimates of S_t (as in Bell and Hillmer 1990a). Graphs of the signal extraction estimates are visually indistinguishable from graphs of the composite estimates and, therefore, are not included in this paper.

5. SIGNAL EXTRACTION EXCLUDING CERTAINTY CASES

In the RTS (as in some other surveys) some sample units are selected with probability 1 ("certainty cases"). The certainty cases in the RTS are the larger businesses that contribute more, often much more, to total sales than the individual noncertainty cases. Stratified sampling with a certainty stratum can produce much better estimates than, say, simple random sampling. In the RTS, certainty status for a company is determined by comparing sales reported in the most recent economic census to "certainty cutoffs" that vary by kind of business; this is discussed in Isaki et al. (1976) and in the Monthly Retail Trade Reports.

When there are certainty cases, the usual survey estimates treat these separately, adding their known total to a sample-weighted estimate of the total for the noncertainty cases. The same can also be done for time series signal extraction estimation. Considering first the additive decomposition (3.1), let $S_t = S_{1t} + S_{2t}$ where S_{1t} is the (known) total for the certainty cases, and S_{2t} is the (unknown, to be estimated) total for the noncertainty cases. Then

$$Y_t = S_{1t} + S_{2t} + N_t$$

$$\Rightarrow X_t = Y_t - S_{1t} = S_{2t} + N_t \quad (5.1)$$

and since we know $X_t = Y_t - S_{1t}$, we can perform modeling and signal extraction on it to estimate S_{2t}. Call this estimate $\hat{S}_{2t}$, and then take $\hat{S}_t = S_{1t} + \hat{S}_{2t}$, with error $S_{2t} - \hat{S}_{2t}$. Of course, we could simply model and do signal extraction on Y_t to estimate S_t; call this estimate $\check{S}_t$. From results in Bell and Hillmer (1990a), the variance matrices of the signal extraction errors under these two approaches are (let $Z_t = \delta(B)X_t$, $W_t = \delta(B)Y_t$, $\underline{S} = (S_1, \ldots, S_n)'$, etc.),

$$Var(\underline{S} - \hat{\underline{S}}) = Var(\underline{S}_2 - \hat{\underline{S}}_2) = \Sigma_N - \Sigma_N\Delta'\Sigma_Z^{-1}\Delta\Sigma_N \quad (5.2)$$

$$Var(\underline{S} - \check{\underline{S}}) = \Sigma_N - \Sigma_N\Delta'\Sigma_W^{-1}\Delta\Sigma_N \quad (5.3)$$

(5.2) differs from (5.3) in that Σ_Z^{-1} replaces Σ_W^{-1}. Since $Z_t = W_t - \delta(B)S_{1t} = \delta(B)[Y_t - S_{1t}]$, if there is a substantial certainty component, then we might expect $\Sigma_Z < \Sigma_W$ implying $\Sigma_Z^{-1} > \Sigma_W^{-1}$ and $Var(\underline{S} - \hat{\underline{S}}) < Var(\underline{S} - \check{\underline{S}})$. (Here $\Sigma_Z < \Sigma_W$ means $\Sigma_W - \Sigma_Z$ is a positive definite matrix.) Thus, removing certainty cases may lead to improvement in the signal extraction estimates. The improvement realized in practice will obviously depend on how large the certainty component is, and on how successful we are at modeling X_t relative to modeling Y_t.

We could go further and use S_{1t} in a multivariate or transfer function model including S_{2t} (hence, S_{1t} would be in the model for X_t). The resulting signal extraction estimate, $\bar{S}_2$ say, would be an estimate of $E[\underline{S}_2|\underline{Y}, \underline{S}_1] = E[\underline{S}_2|\underline{X}, \underline{S}_1]$, and $\bar{\underline{S}} = \underline{S}_1 + \bar{\underline{S}}_2$ would be an estimate of $E[\underline{S}|\underline{Y}, \underline{S}_1]$. The error variance is then

$$\mathrm{Var}(\underline{S} - \bar{\underline{S}}) = \mathrm{Var}[\underline{S}_2|\underline{Y},\underline{S}_1] = \mathrm{Var}[\underline{S}|\underline{Y},\underline{S}_1] \leq \mathrm{Var}(\underline{S}|\underline{Y})$$

$$= \mathrm{Var}(\underline{S} - \check{\underline{S}}).$$

Thus, theoretically, $\bar{\underline{S}}$ is known to be better than $\check{\underline{S}}$, whereas $\hat{\underline{S}}$ may not always be theoretically better than $\check{\underline{S}}$ since $\Sigma_Z < \Sigma_W$ may not hold. However, $\bar{\underline{S}}$ is more difficult to obtain, requiring multivariate modeling of S_{1t} and X_t with sampling error N_t. In this paper, we restrict attention to $\hat{\underline{S}}$, actually using the multiplicative analogue we now describe.

A similar approach can be used when, as in the RTS, it is more appropriate to treat the sampling error as multiplicative. In this case we replace (5.1) by

$$Y_t = S_{1t} \cdot A_t \cdot U_t \qquad (5.4)$$

where $A_t = S_t/S_{1t}$. Then, redefining X_t:

$$X_t \equiv Y_t/S_{1t} = A_t \cdot U_t$$

$$\Rightarrow \ln(X_t) = \ln(A_t) + \ln(U_t) \qquad (5.5)$$

We then proceed with modeling and signal extraction for (5.5), producing $\ln(\hat{A}_t)$ and hence $\hat{A}_t = \exp[\ln(\hat{A}_t)]$, and then $\hat{S}_t = S_{1t} \cdot \hat{A}_t$. It can be shown that (now $Z_t = \delta(B)\ln(X_t)$)

$$\mathrm{Var}[\ln(\underline{S}) - \ln(\hat{\underline{S}})] = \mathrm{Var}[\ln(\underline{A}) - \ln(\hat{\underline{A}})]$$

$$= \Sigma_U - \Sigma_U\Delta'\Sigma_Z^{-1}\Delta\Sigma_U \qquad (5.6)$$

and that (5.6) approximately gives relative variances and covariances of $S_t - \hat{S}_t$. As before, (5.6) can be compared to $\Sigma_U - \Sigma_U\Delta'\Sigma_W^{-1}\Delta\Sigma_U$ (where now $W_t = \delta(B)\ln(Y_t)$), to see if removing certainty cases leads to improved estimates.

(5.1) and (5.5) have an important feature, namely, that for either the additive or multiplicative decompositions, the sampling error in Y_t is the same as that in X_t. Thus, the removal of certainty cases does not affect the sampling error models used. We only need model the new observed series X_t($\equiv Y_t - S_{1t}$ or $\equiv Y_t/S_{1t}$) with the same sampling error component as before. In the next section we do this for some RTS time series using the sampling error models of section 3.2.

6. MODELING AND SIGNAL EXTRACTION IN THE RTS EXCLUDING CERTAINTY CASES

As mentioned above, it is more appropriate to treat the sampling error as multiplicative for the RTS. Graphs of the RTS series with certainty cases removed can be found in Figures 2a,b,c. The mean percentage of certainty cases was 14.4 percent, 30.9 percent, and 31.5 percent for household appliances, men's and boy' clothing, and radio and TV establishments, respectively. Of note is the erratic behavior of the series. This may lead one to believe there will be difficulties in modeling the series.

Actually, data on total sales of certainty cases in the RTS was not readily available. However, data on sales of "Group II" cases were available by kind of business. The "Group II" panel only includes those businesses with eleven or more retail establishments, which generally contains most, though not all of the certainty cases. While it would be preferrable to remove (divide out) sales for all certainty cases, it is nevertheless valid to remove only the Group II sales. For convenience in the examples here we shall still use the term "certainty cases", although, strictly speaking, "Group II firms" would be correct.

6.1 MODELING THE COMPONENTS FOR RTS WITH CERTAINTY CASES EXCLUDED

With respect to the series discussed earlier in this paper, the general form of the model for A_t is:

$$\delta(B)[\ln(A_t) - \mu_t] = [\theta_A(B)/\phi_A(B)]b_t \qquad (6.1)$$

where $\delta(B)$, $\theta_A(B)$, and $\phi_A(B)$ are the differencing, autoregressive, and moving average operators, respectectively; μ is the regression term, and b_t is a white noise series (iid $N(0,\sigma_b^2)$). As noted earlier, the sampling error models will be the same as those developed for the case where certainty cases are not removed.

6.2 MODEL ESTIMATION AND SIGNAL EXTRACTION FOR RTS WITH CERTAINTY CASES EXCLUDED

In actuality, a lot of the seasonality in the original series was eliminated upon dividing the original series by the series of certainty cases. However, for the sake of comparison, we used similar models to those of the case where certainty cases were not removed. This is to insure that the differences which may occur between the models with certainty cases and without certainty cases are not due to modeling differences.

Taking models of the form (6.1) for $\ln(A_t)$ with models of the form (3.5) for $\ln(U_t)$, the parameters of the models for $\ln(A_t)$ were estimated as mentioned in section 4. Estimation of the models produced the following:

1) Household Appliance Establishments

$$(1-B)(1-B^{12})(\ln(A_t)-\Sigma\beta_i T_{it}) = (1-.32B-.24B^2)(1-.67B^{12})b_t \qquad (6.2a)$$

$$\sigma_b^2 = 2{,}900 \times 10^{-6}$$

2) Men's and Boy's Clothing Establishments

$$(1-B^{12})(1-.33B+.09B^2)(\ln(A_t)-\Sigma\beta_i T_i t-\Sigma\alpha_i h(t,t_i)) = (1-.36B^{12})b_t \qquad (6.2b)$$

$$\sigma_b^2 = 2{,}932 \times 10^{-6}$$

3) Radio and TV Establishments

$$(1-B)(1-B^{12})(\ln(V_t)-\Sigma\beta_i T_{it}) = (1-.19B)(1-.76B^{12})b_t \qquad (6.2c)$$

$$\sigma_b^2 = 1{,}948 \times 10^{-6}$$

The Q statistics associated with each of the estimated models were reasonable.

Once again, the estimated models, (6.2a,b,c) with models of the form (3.5), were used to produce signal extraction estimates of $\ln(A_t)$, which were added to $\ln(S_{1t})$ and exponentiated to produce estimates of S_t. Again, graphs of the signal extraction estimates are visually indistinguishable from the graphs of the composite estimates and, therefore, are not included in this paper.

Figure 3 presents graphs of the signal extraction error variances. For Men's and Boy's Clothing Establishments, the variances behave as expected. That is, the variance of the signal extraction error is lower when certainty cases are removed. However, the graphs for the other two series present just the opposite claim. Hence, it is not clear whether, in practice, exclusion of certainty cases will reduce the variance of the signal extraction error and thus improve the signal extraction estimates.

7. CONCLUSIONS

In this paper, we have examined the sensitivity of signal extraction results for time series from the Census Bureau's Retail Trade Survey (RTS) to removing certainty cases from the series. Using available information on sampling error autocorrelations and (relative) variances, we constructed time series models for sampling errors in the RTS. Then, using Box-Jenkins type ARIMA models for the signal and sampling error components, we estimated models and did signal extraction for some time series from the RTS, with and without certainty cases included. While we might expect theoretically some reduction in signal extraction variance from removing certainty cases, this may not be borne out in practice. Removal of certainty cases from a series may produce a series which is much more erratic in behavior and, thus, more difficult to model. In addition, the lowering of variances is contingent upon whether or not the certainty cases are positively correlated or uncorrelated with the noncertainty cases, loosely speaking. This is an issue that will have to be looked at more closely in the future. With regard to the series used for this paper, we found an improvement in signal extraction results for one series with certainty cases removed, and apparently worse results for the other two series. In addition, though we are not reporting the results here, we looked at four additional series and found similar mixed results.

It does appear, however, that signal extraction in and of itself is beneficial. Table 1 presents a list of coefficients of variation (CV) for HT, composite, and signal extraction estimators (minimum and maximum over time for the latter). The signal extraction CVs (with and without certainty cases) are lower than those for HT and composite estimates. Thus, while it appears signal extraction techniques can improve estimates in repeated surveys, further work is needed to better assess the effect of removing certainty cases on time series signal extraction estimates.

8. DISCLAIMER

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

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

Bell, W. R. and Hillmer, S. C. (1989), "Modeling Time Series Subject to Sampling Error," Research Report 89/01, Statistical Research Division, Bureau of the Census.

_____(1990a), "The Time Series Approach to Estimation for Periodic Surveys," Survey Methodology, to appear.

_____(1990b), "Initializing the Kalman Filter for Nonstationary Time Series Models," Journal of Time Series Analysis, to appear.

Bell, W. R. and Wilcox, D. W. (1990) "The Effect of Sampling Error on the Time Series Behavior of Consumption Data," paper presented at the NBER Conference on Seasonality in Econometric Models, University of Montreal.

Isaki, C. T., Wolter, K. M., Sturdevant, T. R., Monsour, N. J., and Trager, M. L. (1976), "Sample Redesign of the Census Bureau's Monthly Business Surveys," Proceedings of the American Statistical Association, Business and Economic Statistics Section, 90-98.

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 Observarions," Technometrics, 22, 389-395.

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

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.

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

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

Wolter, K. M., Isaki, C. T., Sturdevant, T. R., Monsour, N. J., and Mayes, F. M. (1976), "Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys," American Statistical Association, Proceedings of the Business and Economic Statistics Section, 99-109.

Figure 1a.
Composite Estimates
for Household Appliances

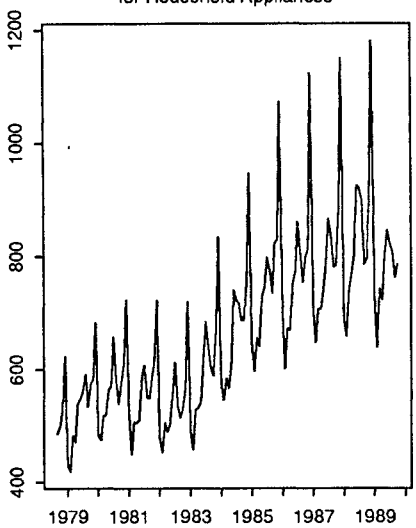

Figure 1b.
Composite Estimates
for Men's and Boy's Clothing

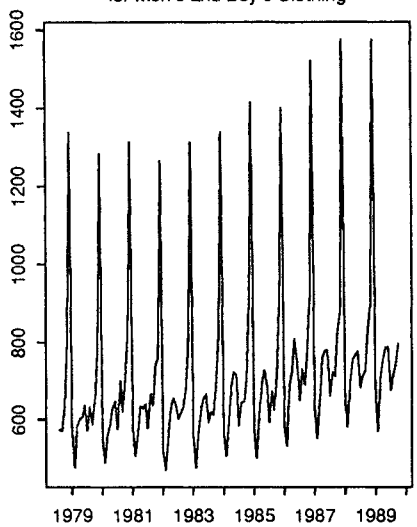

Figure 1c.
Composite Estimates
for Radios and TVs

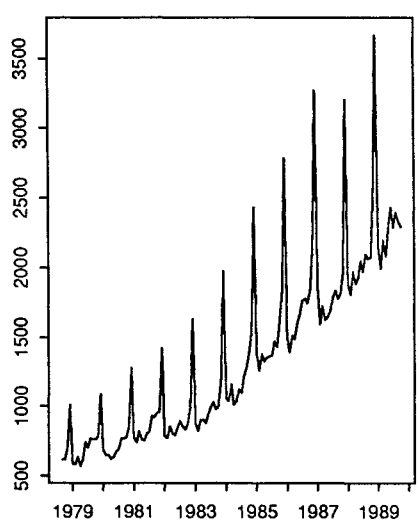

Figure 2a.
Composite Estimates
for Household Appliances
Excluding Certainty Cases

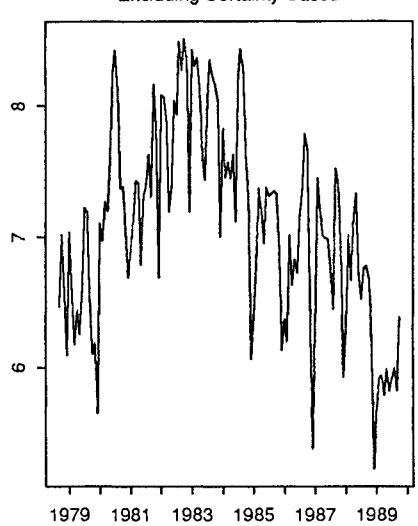

Figure 2b.
Composite Estimates
for Men's and Boy's Clothing
Excluding Certainty Cases

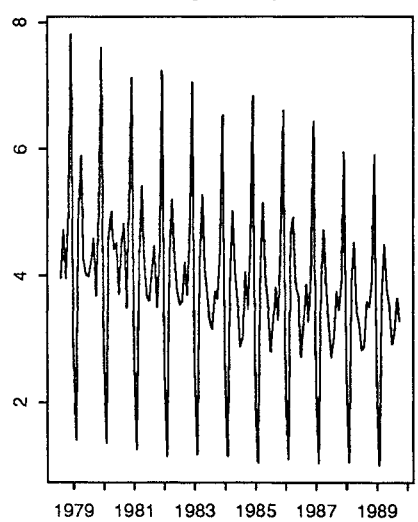

Figure 2c.
Composite Estimates
for Radios and TVs
Excluding Certainty Cases

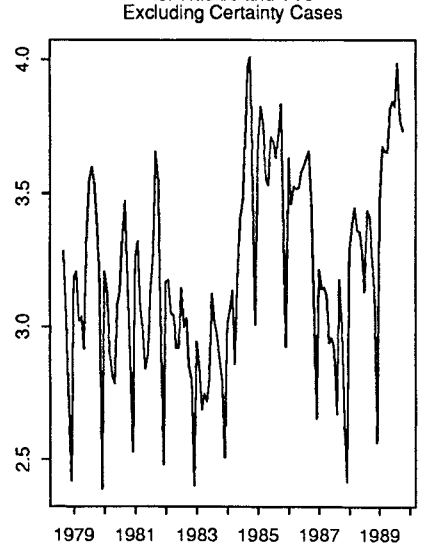

Figure 3a.
Variance Estimates for
Signal Extraction Results
for Household Appliances
With Certainty Cases - Solid
Without Certainty Cases - Dotted

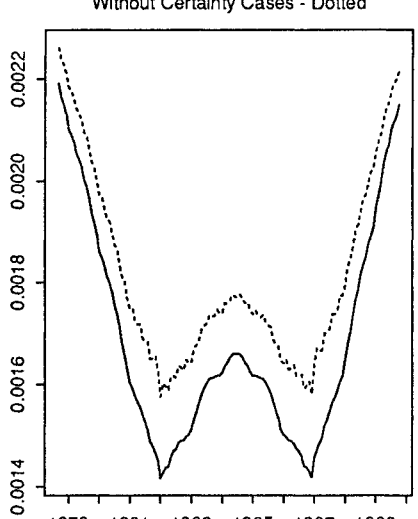

Figure 3b.
Variance Estimates for
Signal Extraction Results
for Men's and Boy's Clothing
With Certainty Cases - Solid
Without Certainty Cases - Dotted

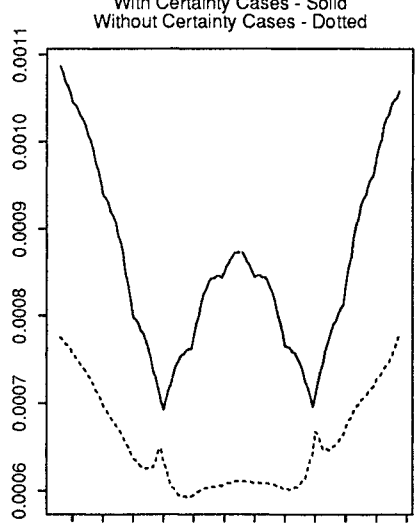

Figure 3c.
Variance Estimates for
Signal Extraction Results
for Radios and TVs
With Certainty Cases - Solid
Without Certainty Cases - Dotted

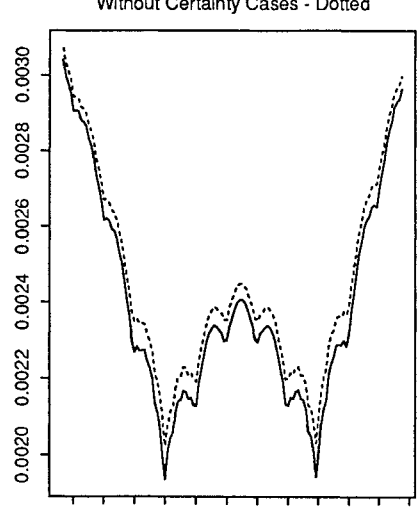

Table 1.

Coeffiecients of Variation

	HT	Composite	Signal Extraction with certainty cases		Signal Extraction without certainty cases	
			min	max	min	max
Household Appliances	7.8	5.1	3.7	4.7	4.0	4.8
Men's and Boy's Clothing	5.1	3.6	2.6	3.3	2.4	2.8
Radios and TVs	15.7	6.1	4.4	5.5	4.5	5.5