International Finance Discussion Papers

Number 318

December 1987

IMPROVING THE FORECAST ACCURACY OF PROVISIONAL DATA: AN APPLICATION OF THE KALMAN FILTER TO RETAIL SALES ESTIMATES

by

B. Dianne Pauls

ABSTRACT

If forecasts of economic activity are to rely on preliminary data, the predictable component of the data revisions should be taken into account. This paper applies the Kalman filter to improve the forecast accuracy of published preliminary estimates of retail sales. Successive estimates of retail sales are modeled jointly as a vector autoregressive process, incorporating panel rotation and calendar effects. Estimates of retail sales based on this model are then combined with the raw Census estimates via the Kalman filter. This technique, which may be applied to other bodies of data, yields a significant improvement in the efficiency of the raw Census data, reducing the mean-squared error by about 1/3.

Improving the Forecast Accuracy of Provisional Data: An Application of the Kalman Filter to Retail Sales Estimates

by

B. Dianne Pauls[1]

1. Introduction

Numerous studies have documented the substantial nature of revisions to macroeconomic data.[2] Because timely forecasts of economic activity often rely on such preliminary data nonetheless, assessing whether the data revisions are forecastable is crucial. A recent study by Mankiw and Shapiro (1986) suggests that, although revisions to GNP are quite large, they can not be forecast based on known information. Thus, the preliminary data represents the best forecast -- in a mean-squared error sense -- of the final estimate given all the available information, and their literal use macroeconomic projections produces efficient forecasts. This result contrasts with earlier work by Howrey (1978) and Conrad and Corrado (1978) which indicates that revisions to disposable income and retail sales are forecastable based on known information. These studies further demonstrate substantial scope for improving

[1] The author is a member of the Division of International Finance, Board of Governors of the Federal Reserve System. This work was begun while the author was in the Division of Research and Statistics. The views expressed herein are the author's own and do not necessarily reflect the views of the Board of Governors or other members of its staff. I have benefited from conversations with Susan Burch, William Cleveland, Neil Ericsson, and several analysts at the Bureau of the Census and from comments by David Pierce on an earlier draft. I am also grateful to Stephen Gordon and William Nelson for research assistance.

[2] See for example, Zellner (1958), Cole (1969), and Mankiw and Shapiro (1986).

forecasting accuracy by using signal extraction methods to exploit this correlation.

This paper follows the work of Conrad and Corrado in using the Kalman filter to derive a better estimate of underlying retail sales activity. Retail sales are a prime candidate for this analysis because the revisions often are sizable relative to the movement in the underlying series. Over the period of this study -- January 1982 to May 1986 -- the mean absolute revision in the month-to-month change in total retail sales (excluding autos, gasoline, and nonconsumption items) is 0.46 percent, while the mean absolute month-to-month change is 0.73 percent. Furthermore, an examination of the Census Bureau's sampling procedures suggests a likely correlation in the data revisions. In particular, monthly estimates of retail sales are based on a rotating panels of stores. Thus, if these panels are distinct, one would expect, at a minimum, a correlation between estimates of sales activity based on a common panel of stores.

The Kalman filter provides an algorithm for signal extraction; the efficiency of the preliminary data is improved by combining it with an alternative information source. The alternative information is the prediction of the final estimates based on a time series model that explicitly incorporates the structure of the data revisions. The resulting estimates are more efficient in the sense of having a smaller mean-squared error than either the preliminary observation or the model-based prediction.

This paper differs from earlier work by Conrad and Corrado primarily in its approach to the time series modeling of retail sales. A more explicit model of the panel rotation bias in retail sales is presented. And, rather than estimate univariate models using seasonally differenced data, successive estimates of retail sales are modeled jointly as a vector autoregressive process. The parameters of this process are estimated simultaneously with calendar effects, such as the composition of the month and the occurrence of Easter. In addition, because the Census Bureau substantially revised its sampling procedures in October 1977, the time series model presented in Conrad and Corrado probably is no longer applicable.

The results yield a significant improvement in the efficiency of the raw Census data, reducing the mean-squared error by about 1/3. Given that retail sales are the primary source data for personal consumption expenditures -- the largest component of GNP -- there are two possibilities for reconciling these findings with those of Mankiw and Shapiro. Either the Bureau of Economic Analysis takes account of the systematic nature of revisions to retail sales data in constructing quarterly personal consumption figures, or revisions to other components of GNP behave in such way as to offset this correlation.

The paper begins with a general discussion of the application of the Kalman filter to preliminary data -- contained in Section 2. Before modeling the data revision process, however, an understanding of the Census Bureau's construction of successive estimates of retail sales is required. Section 3 outlines these procedures. A joint time series model of the final estimates and the data revisions is presented in Section 4. Once the systematic components of the revisions have been incorporated into the model, the Kalman filter is applied in Sections 5 and 6 to reduce the non-systematic components of the revision process. Section 7 contains concluding remarks.

2. The Application of the Kalman filter preliminary data

The Kalman filter provides an algorithm for using the dynamic behavior of a time series to reduce the measurement error in the sample observations of that series. It entails a set of equations that utilizes the information in the latest, possibly error-ridden, observation to update the prediction of the true data based on its own past values and, perhaps, other exogenous variables. The application of the Kalman filter to preliminary data is just a special case of the measurement error problem.

The first step in applying the Kalman filter is to specify the evolution of the final series -- the data without measurement error. This dynamic behavior is described in the so-called state or transition equation. Denoting the final data as x_t, the state equation is written as a first-order Markov process:

$$x_t = Ax_{t-1} + Cz_t + Re_t \tag{1}$$

where x_t is a vector representing the state at time t; and A, C, and R are matrices of known coefficients, z_t is a vector of exogenous variables, and e_t is a vector of random disturbances with mean 0 and covariance matrix Σ. Although the state equation is specified as a first-order Markov process, the model can easily be extended to encompass higher-order autoregressive processes for the data by suitably augmenting the state vector.

Next, the observation or measurement equation relating the preliminary data to the final data is specified. Letting y_t denote the preliminary data, this equation is:

$$y_t = Hx_t + v_t \qquad (2)$$

where y_t is a vector of preliminary observations, of possibly smaller dimension than x_t, H is a matrix of known coefficients, and v_t is a vector of observation equation errors with mean 0 and covariance matrix V.

The one-period prediction problem is to forecast the final data for date t, x_t, conditional on the information available at time t, which includes the preliminary data, y_t, the exogenous variables, z_t, and previous estimates of the state, x_{t-1}. Under the assumption that e_t and v_t are mutually uncorrelated, serially independent multivariate normal processes and the loss function is symmetric, the optimal predictor, $x_{t|t}$ is given as (Harvey (1981a, 1981b)):

$$x_{t|t} = x_{t|t-1} + K_t \; (y_t - Hx_{t|t-1}) \qquad (3)$$

where $x_{t|t-1} = Ax_{t-1|t-1} + Cz_t$ is the prediction of x_t based on the state equation. Denoting the covariance matrix of the state estimate errors, $x_t - x_{t|t-1}$, as $P_{t|t-1}$, the Kalman gain, K_t, can be expressed as:

$$K_t = P_{t|t-1} \; H'(V + HP_{t|t-1}H')^{-1}$$

To complete the filtering equations a method for updating the estimate of the covariance matrix of the state estimate errors as new information becomes available is needed. The one-period ahead projection of the covariance matrix, P, is given by:

$$P_{t|t-1} = AP_{t-1|t-1}A' + R\Sigma R'$$

And the optimal predictor of the covariance matrix using all the information a time t is:

$$P_{t|t} = P_{t|t-1} - P_{t|t-1} H'(V + HP_{t|t-1}H')^{-1} HP_{t|t-1} \tag{4}$$

In applying the above equations, several assumptions underlying the Kalman filter must be met. First, consistent estimates of the coefficient matrices A, C, R, and H, and the covariance matrices Σ and V are needed. These are obtained from prior estimation of the state and observation equations. Second, the model requires that the errors e_t and v_t be serially independent. Given the errors in the state equation are the residuals from a time-series model they should have this property. However, the errors from the observation equation represent data revisions, which for many macroeconomic time series are serially correlated. Therefore, the observation equation may need to be modified to conform with the assumptions underlying the Kalman filter. This problem is addressed in Section 5. First, a description of the data and estimates of time-series models of the final data and the data revisions are presented.

3. Census construction of the retail sales data

Retail sales are receipts plus merchandise sold for credit at establishments that are primarily engaged in retail trade. The latter is defined as the sale of merchandise to consumers for personal and household use. Sales exclude sales and excise taxes paid by customers and finance charges. They do, however, include gasoline, tobacco, liquor and other excise taxes, which are paid by the manufacturer and passed through to the retailer and the customer.

Estimates of retail sales are constructed from a sample that consists of very large businesses, which report every month, plus three panels drawn from the list sample and twelve area panels.[3] Based on the dollar volume of sales, businesses that report on a continuous basis comprise 35 percent of the sample, list panels that report every three months for the preliminary and final estimates account for another 59 percent, and the remaining 6 percent consists of area panels that report every twelve or six months for the preliminary and final estimates.

For each month, there are three successive estimates of retail sales -- referred to as the advance, preliminary, and final estimates. The advance estimate is published roughly ten days after the date to which the data refer and is based on a fixed subsample of the very large businesses reporting on a continuous basis, the list panels, and the area panels. The preliminary and final estimates are issued one and two months, respectively, after the advance estimate appears and are based on the fixed sample of very large businesses together with a series of rotating panels.

Each rotating panel reports two months of data at each enumeration. So, for example, the panel reporting in September gives its sales figures for July and August. In October, another panel records sales for August and September, and so on. Hence, for each month there are two panel estimates of retail sales. The Census Bureau combines each

[3] The list sample is a probability sample selected from the retail employers (SIC categories 52-59) contained in the Census Bureau's Standard Statistical Establishment List, which effectively covers all employers who made social security payments for their employees under FICA.

of these panel estimates with the fixed component of the sample -- those stores reporting monthly -- to obtain what they term the "unbiased" estimates of retail sales. For any given month, t, there are two unbiased estimates, which are denoted as ${}_{t+1}u_t$ and ${}_{t+2}u_t$. The preceeding subscripts denote the time the information becomes available while the succeeding subscripts denote the date to which the data pertain. Thus, ${}_{t+1}u_t$ is based on the panel that reports with a one month lag and the fixed subsample reporting monthly, and ${}_{t+2}u_t$ is based on the panel that reports with a two month lag in addition to the fixed subsample.

The preliminary and final estimates for each three-digit SIC level are constructed from the unbiased estimates according to the following formulas:

$$ {}_{t+1}pre_t = .25\,{}_{t+1}u_t + .75 \; \frac{{}_{t+1}u_t}{{}_{t+1}u_{t-1}} \; {}_{t}pre_{t-1} \qquad (5) $$

$$ {}_{t+2}fin_t = .8 \; {}_{t+1}pre_t + .2 \; {}_{t+2}u_t \qquad (6) $$

where: pre = preliminary estimate

fin = final estimate

These filters are designed to smooth changes in unbiased estimates of retail sales due to variation across panels. Equation (5) indicates that the preliminary estimate is a weighted average of the unbiased estimate reported with a one month lag and an estimate that is obtained by multiplying the previous month's preliminary estimate by the ratio of current-to-previous month's unbiased estimates of sales based on the <u>current</u> panel of stores. This procedure attempts to put the new panel's data on a comparable basis with the previous panel's data. The final estimate is a weighted average of the preliminary estimate for the same month and the new panel's unbiased estimate of sales, ${}_{t+2}u_t$, and thus averages the two panels that report sales in a given month.

The advance estimate is also constructed by applying a filter to the reported data. Recall that the advance estimate is based on a fixed subsample of the universe represented in the preliminary and final estimates. In an attempt to put the advance estimate on a comparable basis with the more comprehensive sample the Census Bureau computes it as:

$${}_t adv_t = \frac{{}_t ua_t}{{}_{t-1} ua_{t-1}} \quad {}_t pre_{t-1} , \tag{7}$$

where ${}_t ua_t$ denotes the "unbiased" estimate of sales in month t based on the (fixed) advance subsample.

The data used in this study begin with the revision in the Census Bureau's comprehensive sample in January 1982 and extend through mid-1986. Although this period contains several revisions in the advance sample, which is redrawn every two years, these changes are not thought to significantly hamper modeling the systematic component of the data revisions.

4. Forecasting and observation models for retail sales

In this application, the state equation describes the behavior of the preliminary and final estimates for the retail control grouping of stores, defined as total retail sales excluding automotive sales, gasoline sales, and non-consumption items. Rather than model the preliminary and final series directly, the filters used by the Census

Bureau were inverted to recover the raw or unbiased estimates of retail sales because the systematic correlations induced by the sampling procedures were thought to be more evident in the raw data.[4] That is, equations (5) and (6) were solved to obtain:

$$ {}_{t+1}u_t = \frac{{}_{t+1}pre_t}{.25 + .75 \ \dfrac{{}_{t}pre_{t-1}}{{}_{t+1}u_{t-1}}} \qquad (5') $$

$$ {}_{t+2}u_t = 5 \ {}_{t+2}fin_t - 4 \ {}_{t+1}pre_t \qquad (6') $$

Hereafter, the two estimates of retail sales for a given month will be referred to as u1 and u2, where u1 = ${}_{t+1}u_t$ and u2 = ${}_{t+2}u_t$.

In modeling the time-series behavior of retail sales, nonseasonally adjusted data were used and seasonal patterns were estimated as part of the model. In addition to the usual calendar effects associated with monthly data, the panel rotation procedures described previously introduce a correlation at three month intervals -- corresponding to the period between successive reports by a given panel. Furthermore, because the panels reporting every three months comprise nearly 60 percent of the sample, a correlation at the panel rotation frequency should be noticeable in the data. A second influence on

[4] Recall the Census filters are applied at a disaggregated level and then preliminary and final estimates are aggregated across categories representing major kinds of business to obtain sales for the retail control grouping of stores. However, in the procedure above, the unbiased estimates were reconstructed by applying equations (5') and (6') to the final and preliminary estimates for <u>total</u> retail control. This methodology will not reproduce exactly the original unbiased series because the filter given in equations (5) and (6) is nonlinear in the unbiased estimates.

seasonal patterns in retail sales is the calendar variation in the occurrence of major holidays, such as Easter. To assess these possible influences on retail sales, a basic model of the seasonality was estimated and the residuals were examined for the presence of a panel rotation and an Easter effect. The model included a constant, a time trend, a series of variables representing the composition of the month, and monthly seasonal dummies. In estimating this model the time series u1 and u2 were stacked, imposing the restriction that the coefficients be the same in the two equations. However, the mean and time trend were allowed to vary across the two series.

Plots of the spectra of the residuals from this regression are shown in Figures 1 and 2. The markings on the horizontal axis indicate the seasonal frequencies for monthly data. As the monthly dummies should remove most of the power at these frequencies, the spectrum generally takes on small values at exact multiples of $\pi/6$. The one exception occurs at $2\pi/3$, which is the frequency corresponding to a three-month cycle in the data. The remaining power at this frequency represents the effect of the panel rotation.

Another prominent feature of the spectral plots is that, although the power is removed at the seasonal frequencies, relatively large peaks in the spectra remain near the seasonal frequencies. This pattern suggests a seasonal influence that does not quite occur at the same interval each year. In the case of retail sales, the variable timing of Easter is probably important in accounting for the sharp peaks adjacent to the seasonal frequencies.

To examine the Easter effect, we compared the behavior of April versus March retail sales -- after removing a deterministic trend and

Figure 1

Spectrum of U1 Series of Retail Sales: Constant, Trend, Trading Day and Monthly Effect Removed

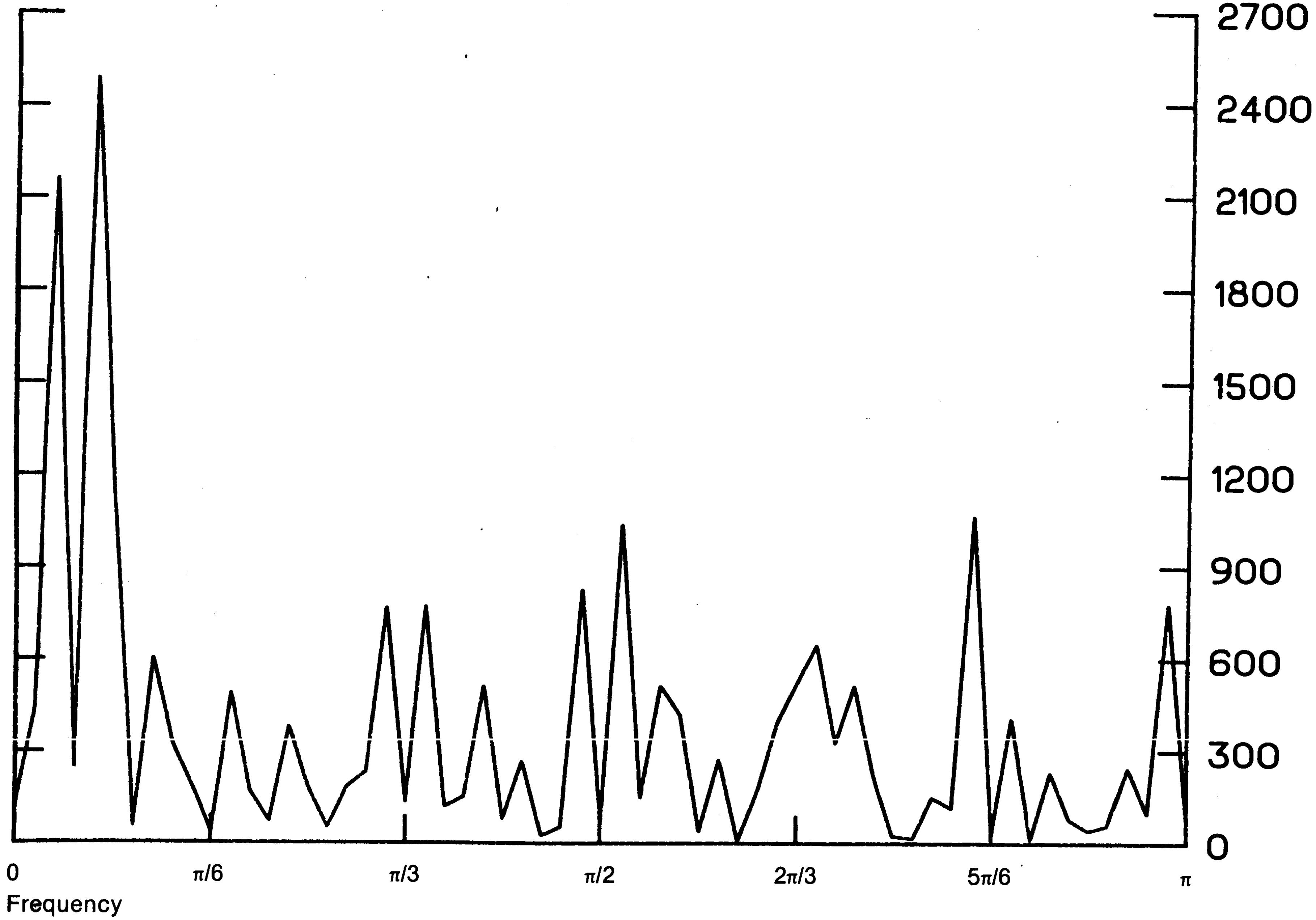
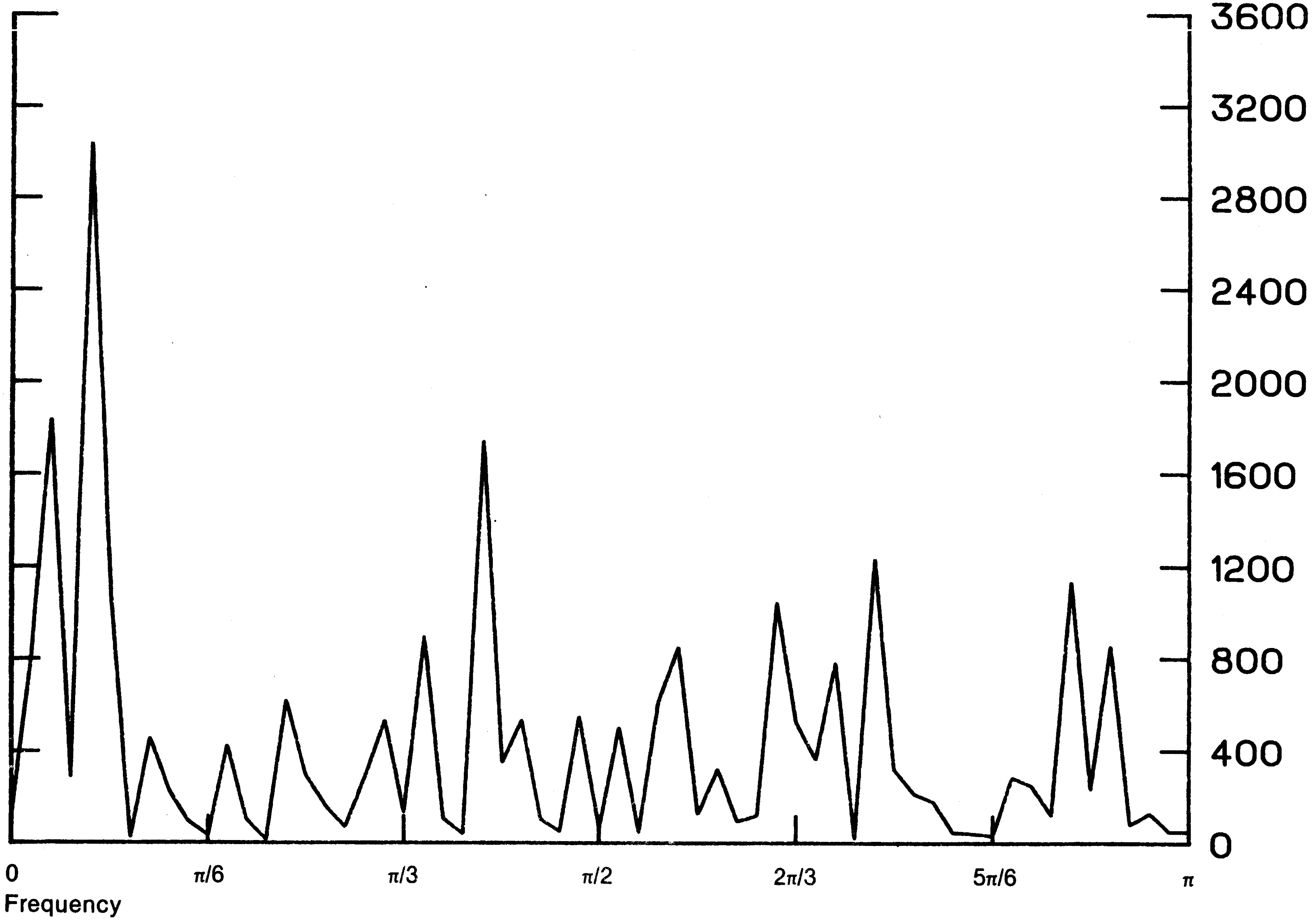


Figure 2

Spectrum of U2 Series of Retail Sales: Constant, Trend, Trading Day and Monthly Effect Removed



trading day effects -- with the date of Easter. A plot of the number of days between April 1 and Easter versus the difference between calendar-adjusted April and March sales is given in Figure 3. Although we have relatively few observations on the timing of Easter, the relative magnitude of April versus March sales appears to be a linearly increasing function of the distance of Easter from April 1 over the interval April 1 to April 14. Furthermore, Cleveland and Grupe (1983), in a study of this same series though with more observations, also observed a linear relationship between the date of Easter and the relative magnitude of April versus March sales.

Based on these features of the data, the following fixed effects model was specified for the two series of unbiased estimates of retail sales:

$$\log(u1) = \alpha_1 \text{ TIME} + \sum_{i=1}^{7} \beta_i \text{ TD}_i + \gamma \text{ EASTER} \tag{8}$$

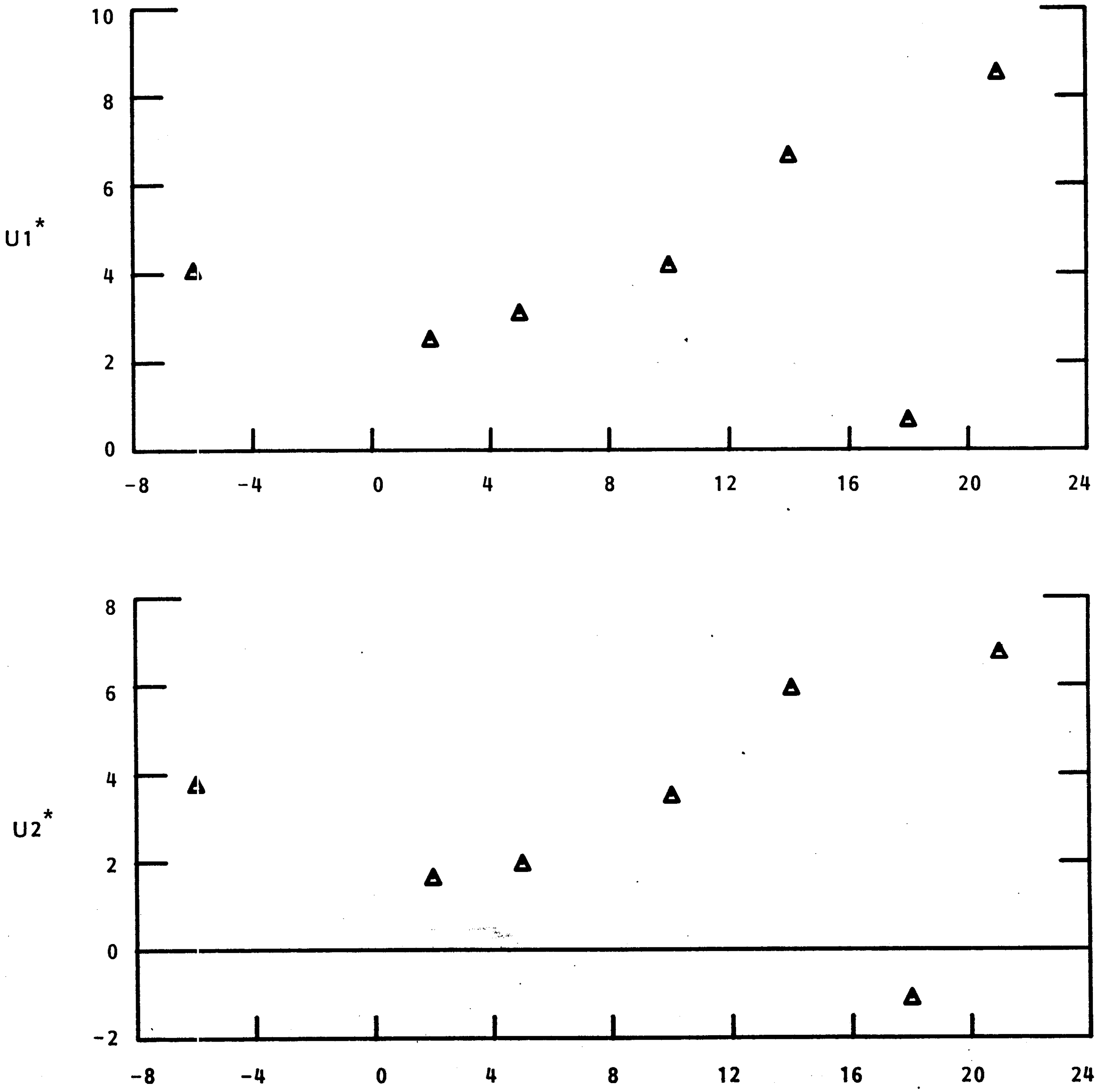
$$+ \sum_{i=1}^{3} \sum_{j=0}^{3} \delta_i \, \phi_{i+3j} \text{ MONTH}_{i+3j}$$

$$\log(u2) = \alpha_2 \text{TIME} + \sum_{i=1}^{7} \beta_i \text{ TD}_i + \gamma \text{ EASTER} \tag{9}$$

$$+ \sum_{i=1}^{3} \sum_{j=0}^{3} \delta_i \, \phi_{i+3j-1} \text{ MONTH}_{i+3j-1}$$

Figure 3

EASTER EFFECT IN RETAIL CONTROL



DAY OF EASTER MINUS APRIL 1

*APRIL MINUS MARCH RESIDUAL SALES (BILLIONS OF DOLLARS) where: TD_i = trading day factors.

EASTER = a dummy variable for April that equals the number of days in a 2-week interval prior to Easter falling in March.

$MONTH_{i+3j}$ = monthly dummies.

The δ_i's represent the panel effects, and the ϕ_i's capture the usual monthly effects. Notice that these two coefficients enter multiplicatively, so that a given month's estimate can be viewed as the interaction of a panel rotation and a monthly effect. Exploiting the fact that a given panel reports the current month's sales in the u1 series and the previous month's sales in the u2 series, the δ_i's are identified by imposing cross-equation restrictions. Because this procedure only allows us to identify relative factors for one panel versus another rather than three separate panel effects, a normalization was chosen in which one of the δ_i's was set equal to one.[5]

The forecasting model for retail sales consists of the fixed effects model given by equations (8) and (9) together with a time-series

[5] To see the difficulty in identifying three separate panel effects, consider unconstrained estimation of each equation. In that case, 12 monthly factors are estimated for each series. If each of these monthly factors is really the product of a panel effect and a monthly effect, then, under the hypothesis that the underlying monthly effects are the same for the two series, the panel effects can be recovered indirectly by forming ratios of the estimated monthly factors in the two equations. Denoting the estimated monthly factors for the two series as $a_i(1)$ and $a_i(2)$,

$$\frac{a_i(1)}{a_i(2)} = \frac{\delta_{j+1}}{\delta_j}$$

Notice that this procedure only provides estimates of the ratio of successive panel effects.

model of the residuals. Because u1 and u2 both contain estimates of retail sales for a given month, a vector autoregressive process, which includes residuals from both series was specified for the residuals. The seasonal parameters and the coefficients of the time series model were estimated simultaneously. To determine the appropriate lag length, a general model that includes lag parameters at the low frequencies as well as at the panel rotation and seasonal frequencies was tested against more restricted alternatives. Denoting the residuals from the fixed effects model, which are in logs, as $\tilde{\tilde{u}}1$ and $\tilde{\tilde{u}}2$, the model selected was:

$$\tilde{\tilde{u}}1 - \tilde{\tilde{u}}2_{-1} = \rho_{11}(\tilde{\tilde{u}}1_{-1} - \tilde{\tilde{u}}2_{-2}) + \rho_{12}\,(\tilde{\tilde{u}}1_{-2} - \tilde{\tilde{u}}2_{-3}) + e1 \qquad (10)$$

$$\tilde{\tilde{u}}2 = \rho_{21}\tilde{\tilde{u}}2_{-1} + \rho_{22}\tilde{\tilde{u}}1_{-2} + \rho_{23}\tilde{\tilde{u}}2_{-3} + e2 \qquad (11)$$

Estimates for this model are displayed in Table 1. The seasonal and trading day effects all appear statistically significant and of reasonable magnitudes. December implicitly has the largest seasonal factor and the trading day coefficients generally increase approaching the end of the week with Friday and Saturday having the largest weights. The Easter effect is significant and negative as an early Easter depresses April retail sales. Recalling that the factor for panel one was normalized to 1, the relative factor for the second panel is significantly different from the factors for panels one and three at the 95 percent confidence level, while panels one and three are insignificantly different.

Table 1: Time Series Model of Unbiased Estimates of Retail Sales

Variable	Coefficient	Standard Error
constant	9.698	.225
time	.0045	.0003
td1 (Sun.)	.0316	.0080
td2 (Mon.)	.0422	.0072
td3 (Tues.)	.0494	.0085
td4 (Wed.)	.0365	.0075
td5 (Thurs.)	.0487	.0079
td6 (Fri.)	.0532	.0074
td7 (Sat.)	.0531	.0081
Easter	-.0019	.0008
panel 2	.949	.0154
panel 3	1.013	.0170
January	-.407	.0055
February	-.321	.021
March	-.331	.0042
April	-.280	.011
May	-.297	.0057
June	-.274	.0091
July	-.330	.0055
August	-.299	.0061
September	-.293	.0090
October	-.301	.0056
November	-.201	.010
$\tilde{\tilde{u}}1_{-1} - \tilde{\tilde{u}}2_{-2}$	-.925	.127
$\tilde{\tilde{u}}1_{-2} - \tilde{\tilde{u}}2_{-3}$	-.672	.132
$\tilde{\tilde{u}}2_{-1}$	.366	.138
$\tilde{\tilde{u}}1_{-2}$	-.452	.343
$\tilde{\tilde{u}}2_{-3}$	.659	.327

$$(\tilde{\tilde{u}}1-\tilde{\tilde{u}}2_{-1}) = \rho_{11}(\tilde{\tilde{u}}1_{-1}-\tilde{\tilde{u}}2_{-2}) + \rho_{12}(\tilde{\tilde{u}}1_{-2}-\tilde{\tilde{u}}2_{-3}) + e1$$

$$\tilde{\tilde{u}}2 = \rho_{21}\tilde{\tilde{u}}2_{-1} + \rho_{22}\tilde{\tilde{u}}1_{-2} + \rho_{23}\tilde{\tilde{u}}2_{-3} + e2$$

Table 1 (cont.)

where:

$$\tilde{\tilde{u}}1 = \log\ (u1) - \Big[c + \alpha\ \text{time} + \sum_{i=1}^{7} \beta_i \text{td}_i + \gamma\ \text{Easter} + \sum_{j=0}^{3} (\phi_{1+3j}\text{MONTH}_{1+3j} + \delta_2\phi_{2+3j}\text{MONTH}_{2+3j}) + \sum_{j=0}^{2} \delta_3\phi_{3+3j}\text{MONTH}_{3+3j})\Big]$$

$$\tilde{\tilde{u}}2 = \log(u2) - \Big[c + \alpha\ \text{time} + \sum_{i=1}^{7} \beta_i \text{td}_i + \gamma\ \text{Easter} + \sum_{i=2}^{3} \sum_{j=0}^{3} \delta_i\phi_{i+3j-1}\text{MONTH}_{i+3j-1} + \sum_{j=0}^{2} \phi_{3+3j}\text{MONTH}_{3+3j}\Big]$$

Turning to the time series properties of the data, given that u1 and $u2_{-1}$ represent the same panel of stores, equation (10) describes the behavior of the month-to-month change in the unbiased estimate of sales based on a consistent panel. According to the parameter estimates shown in Table 1, the month-to-month change displays significant negative second-order autocorrelation. Equation (11) indicates that the second estimate of sales in a given month follows a third-order autoregressive process. The inclusion of $\overset{\approx}{u}1$ at lag 2 rather than the own series residuals as well as $\overset{\approx}{u}2$ at lag 3 captures any remaining panel effects as $\overset{\approx}{u}2$ and $\overset{\approx}{u}1_{-2}$ and $\overset{\approx}{u}2_{-3}$ are all derived from the same panel of stores.

In modeling the early observations or the advance estimates of retail sales, the Census filter given by equation (7) was inverted to yield:

$$\frac{ua_t}{ua_{t-1}} = {}_t adv_t \ - {}_t pre_{t-1} \qquad (7')$$

Notice only the ratio of the raw estimates for successive months can be recovered from the published advance estimate.

Taking logs, the relationship between the monthly change in sales based on the advance subsample and the monthly change based on the same rotating panel of stores was examined. Defining DUA_t as the revision in the estimate of the monthly percentage change in retail sales,

$$DUA_t = \Delta\log(ua_t) - [\log(u1_t) - \log(u2_{t-1})]$$

a third-order autoregressive model, shown below, was found to adequately describe the data revisions.

$$DUA_t = -.267\ DUA_{t-1} + .027\ DUA_{t-2} - .142\ DUA_{t-3} + e3 \qquad (12)$$
$$\qquad\quad (.144) \qquad\qquad (.149) \qquad\qquad (.144)$$

$$R^2 = .108$$

As a final diagnostic, the autocorrelation function of the residuals from the forecasting and observation models, equations (10) - (12), were examined. For all three equations the Box-Pierce Q(12) and Q(24) statistics failed to reject the hypothesis that the residuals are serially independent at the 5 percent critical level.

5. Problem respecification

Given the presence of systematic components in the data revisions, some modification in the formulation of the problem is required. Essentially, the state vector is expanded to incorporate the model of the data revisions, purging the observation equation errors of any systematic components in the data revision process. The state vector is expressed as:

$$x = \begin{bmatrix} \log u1_t \\ \log u2_t \\ \log u1_{t-1} \\ \log u2_{t-1} \\ \log u2_{t-2} \\ DUA_t \\ DUA_{t-1} \\ DUA_{t-2} \end{bmatrix}$$

According to the description of the retail sales data in Section 3 the observation at time t, y_t, consists of:

$$y_t = \begin{bmatrix} \Delta \log ua_t \\ \log u1_{t-1} \\ \log u2_{t-2} \end{bmatrix}$$

Treating the actual data as the sum of the advance estimate and the data revision, the matrix H in the observation equation is defined as:

$$H = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Thus, the observation equation consists of a set of identities, implying that the observation equation errors are zero, apart from rounding. That is,

$$y_t = Hx_t$$

The state equation remains:

$$x_t = Ax_{t-1} + Cz_t + Re_t$$

where Cz_t represent the calendar and panel rotation effects. According to the forecasting and data revision models estimated in the preceeding section the matrix A is specified as:

$$A = \begin{bmatrix} \rho_{11} & 1 & \rho_{12} & -\rho_{11} & -\rho_{12} & 0 & 0 & 0 \\ 0 & \rho_{21} & \rho_{22} & 0 & \rho_{23} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \eta_1 & \eta_2 & \eta_3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

where η_1, η_2, η_3 represent the coefficients from the data revision process -- equation (12).

The matrix R and the residual vector e are defined as:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad e_t = \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \end{bmatrix}$$

where e_{1t}, e_{2t}, e_{3t} are the residuals from equations (10), (11), and (12), respectively.

6. Filtered estimates of retail sales

Using the method outlined above, filtered estimates of the unbiased figures for retail sales -- u1 and u2 -- were calculated over the period May 1982 to May 1986. Recalling that u1 and u2 are available with one and two period lags respectively, $u1_t$, $u2_t$, and $u2_{t-1}$ were estimated based on the information available at time t. Estimates of the preliminary and final series were then reconstructed from the filtered data using equations (5) and (6).

An analogous procedure is used to obtain estimates of preliminary and final retail sales figures based solely on the time series model described in equations (10) and (11). First, model forecasts of the unbiased estimates were calculated using the same information as before. Thus, projected values of $u2_{t-1}$ were used in forming estimates of ul_t and $u2_t$.[6] Next, these estimates were recombined according to equations (5) and (6) to obtain model-based projections of preliminary and final retail sales figures.

The performance of the filtered estimates is compared with that of the model forecast and the Census Bureau's early published data -- labelled the raw observation -- in Table 2. The basis for comparison is the mean-squared error. The results show that the filtered estimates yield about a 30 to 35 percent reduction in the mean-squared error of the observation, depending on which revision is considered. In projecting economic activity, one is most concerned with an estimate of the month-to-month change in retail sales -- column 4 of the table.

The size of the gains achieved by the filter depends on the quality of the forecasts from the alternative information source -- the time series model of retail sales. If these projections are very poor relative to the quality of the raw observation, then little improvement is likely to be achieved through the use of the filter. The forecast errors of the model are of roughly the same order of magnitude as the raw observation errors, as indicated by a comparison of the mean-squared errors in Table 2. Thus, the filter combines two independent forecasts of roughly comparable quality and thereby yields considerable improvement in the quality of the estimates of retail sales activity.

[6] The estimates of ul_t and $u2_t$ formed in this way differ from the one-period ahead model-based projection used in the filter -- $x_{t|t-1}$. The latter uses last period's filtered estimates of $u2_{t-1}$ rather than the model forecast in forming estimates of ul_t and $u2_t$.

Table 2: Results of Kalman filtering of retail sales estimates

	adv to	adv to	pre to	adv-pre$_{-1}$ to
MSE x 10^5	final	pre	final	pre-fin$_{-1}$
1. Raw observation	3.76	2.52	.76	1.64
2. Model projection	3.18	2.93	.63	2.07
3. State estimate	2.48	1.75	.49	1.07
4. Percent reduction from raw observation to state estimate	34	31	36	35

7. Conclusion

If forecasts of economic activity are to rely on preliminary data, the predictable component of the data revisions should be taken into account. For some economic time series, the data revisions are not forecastable based either on their own past or on other information. In these instances, the raw preliminary data provides an efficient forecast of that variable.

This study, however, suggests that revisions to retail sales are more appropriately characterized as classical errors in variables. The results indicate that substantial improvement in the information content of the early data can be obtained by combining it with a time series model of underlying retail sales activity via the Kalman filter. For the monthly change in retail sales -- the most relevant concept for forecasting economic activity -- the sum of squared errors in the early data was reduced by over 30 percent.

These results contrast with work by Mankiw and Shapiro examining the properties of revisions to GNP data. As personal consumption expenditures for goods constitute roughly 1/3 of GNP, either the systematic nature of revisions to retail sales is taken into account in constructing quarterly data on personal consumption expenditures or revisions to other components of GNP behave in such a way as to offset this correlation.

Moreover, the Kalman filter likely can be used to improve the information content of other bodies of provisional data. In assessing its performance, however, one caveat is in order. Although the improvement in forecast accuracy presented here is considerable, it is potentially overstated because the data that is being forecast was used to fit the parameters of the model -- that is, the forecast comparisons are made in-sample. A better gauge of the forecasting ability of the filter would be an examination of its out-of-sample performance, but this was not feasible because the amount of data available on a consistent basis is quite small.

References

Cleveland, W. P., and M. R. Grupe, "Modeling Time Series When Calendar Effects are Present," Applied Time Series Analysis of Economic Data, ed. A. Zellner, 1983, pp. 57-67.

Cole, R., Errors in Provisional Estimates of Gross National Product, New York: National Bureau of Economic Research, 1969.

Conrad, W. and C. Corrado, "The Application of the Kalman Filter to Revisions in Monthly Retail Sales Estimates," Journal of Economic Dynamics and Control, vol. 1, pp. 177-198.

Harvey, A. C., Time Series Models, New York: Wiley, 1981.

, "The Kalman Filter and Its Applications in Econometrics and Time Series Analysis," Mimeo, London School of Economics, 1981.

Howrey, E. P., "The Use of Preliminary Data in Econometric Forecasting," Review of Economics and Statistics, vol. 60 (May 1978), pp. 193-200.

Mankiw, N. G., and M. D. Shapiro, "News or Noise? Revisions of the Preliminary Gross National Product Data," Harvard University Working Paper #1228, April 1986.

U.S. Bureau of the Census, Advance Monthly Retail Sales, various releases, (Washington, D. C.)

, 1982, Revised Monthly Retail Sales and Inventories: January 1972-December 1981 (Washington, D. C.)

Zellner, A., "A Statistical Analysis of Provisional Estimates of Gross National Product and Its Components, of Selected National Income Components, and of Personal Saving, Journal of the American Statistical Association, vol. 53 (1958), pp. 54-65.

International Finance Discussion Papers

IFDP NUMBER	TITLES 1987	AUTHOR(s)
318	Improving the Forecast Accuracy of Provisional Data: An Application of the Kalman Filter to Retail Sales Estimates	B. Dianne Pauls
317	Monte Carlo Methodology and the Finite Sample Properties of Statistics for Testing Nested and Non-Nested Hypotheses	Neil R. Ericsson
316	The U.S. External Deficit: Its Causes and Persistence	Peter Hooper Catherine L. Mann
315	Debt Conversions: Economic Issues for Heavily Indebted Developing Countries	Lewis S. Alexander
314	Exchange Rate Regimes and Macroeconomic Stabilization in a Developing Country	David H. Howard
313	Monetary Policy in Taiwan, China	Robert F. Emery
312	The Pricing of Forward Exchange Rates	Ross Levine
311	Realignment of the Yen-Dollar Exchange Rate: Aspects of the Adjustment Process in Japan	Bonnie E. Loopesko Robert E. Johnson
310	The Effect of Multilateral Trade Clearinghouses on the Demand for International Reserves	Ellen E. Meade
309	Protection and Retaliation: Changing the Rules of the Game	Catherine L. Mann
308	International Duopoly with Tariffs	Eric O'N. Fisher Charles A. Wilson
307	A Simple Simulation Model of International Bank Lending	Henry S. Terrell Robert S. Dohner
306	A Reassessment of Measures of the Dollar's Effective Exchange Value	B. Dianne Pauls William L. Helkie
305	Macroeconomic Instability of the Less Developed Country Economy when Bank Credit is Rationed	David F. Spigelman

Please address requests for copies to International Finance Discussion Papers, Division of International Finance, Stop 24, Board of Governors of the Federal Reserve System, Washington, D.C. 20551.

International Finance Discussion Papers

IFDP NUMBER	TITLES	AUTHOR(s)
304	The U.S. External Deficit in the 1980s: An Empirical Analysis	William L. Helkie Peter Hooper
303	An Analogue Model of Phase-Averaging Procedures	Julia Campos Neil R. Ericsson David F. Hendry
302	A Model of Exchange Rate Pass-Through	Eric O'N. Fisher
301	The Out-of-Sample Forecasting Performance of Exchange Rate Models When Coefficients are Allowed to Change	Garry J. Schinasi P.A.V.B. Swamy
300	Financial Concentration and Development: An Empirical Analysis of the Venezuelan Case	Jaime Marquez Janice Shack-Marquez
299	Deposit Insurance Assessments on Deposits at Foreign Branches of U.S. Banks	Jeffrey C. Marquardt
	1986	
298	The International Debt Situation	Edwin M. Truman
297	The Cost Competitiveness of the Europaper Market	Rodney H. Mills
296	Germany and the European Disease	John Davis Patrick Minford
295	The United States International Asset and Liability Position: A Comparison of Flow of Funds and Commerce Department	Guido E. van der Ven John E. Wilson
294	An International Arbitrage Pricing Model with PPP Deviations	Ross Levine
293	The Structure and Properties of the FRB Multicountry Model	Hali J. Edison Jaime R. Marquez Ralph W. Tryon
292	Short-term and Long-term Expectations of the Yen/Dollar Exchange Rate: Evidence from Survey Data	Jeffrey A. Frankel Kenneth A. Froot
291	Anticipated Fiscal Contraction: The Economic Consequences of the Announcement of Gramm-Rudman-Hollings	Robert A. Johnson