# BLS WORKING PAPERS

U.S. Department of Labor
U.S. Bureau of Labor Statistics
Office of Prices and Living Conditions

**REAL-TIME STATE-SPACE METHOD FOR COMPUTING SMOOTHED ESTIMATES OF FUTURE REVISIONS OF U.S. MONTHLY CHAINED CPI**

**Peter A. Zadroznyy**, U.S. Bureau of Labor Statistics

Working Paper 482
April 2016

All views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Bureau of Labor Statistics.

# REAL-TIME STATE-SPACE METHOD FOR COMPUTING SMOOTHED ESTIMATES OF FUTURE REVISIONS OF U.S. MONTHLY CHAINED CPI

**Peter A. Zadrozny****

Bureau of Labor Statistics
2 Massachusetts Ave., NE, Room 3105
Washington, DC 20212
e-mail: zadrozny.peter@bls.gov

April 14, 2016

JEL Classification: C32, C53, E17

Key words: Kalman smoother estimation of delayed and revised data

**ABSTRACT**

Well known CPI of urban consumers is never revised. Recently initiated chained CPI is initially released every month (ICPI), for that month without delay within BLS and for the previous month with one month delay to the public. Final estimates of chained CPI (FCPI) are released every February for January to December of the calendar year two years before. Every month, simultaneously with the release of ICPI, we would like to have a best estimate, given current information, of FCPI for that month, which will not be released until two calendar years later. ICPI and FCPI data may be indexed in historical time by months of occurrence or in current or real time by months of observation or release. The essence of the solution method is to use data indexed in historical time to estimate models and, then, for an estimated model, to use data indexed in real time to estimate FCPI. We illustrate the method with regression and VARMA models. Using a regression model, estimated FCPI is given directly by an estimated regression line; and, using a VARMA model, estimated FCPI is computed using a Kalman smoother.*

---

*The paper has been presented at the following conferences: Computing in Economics and Finance (Montreal, June 2007), Measurement Errors (Birmingham, UK, July 2007), Joint Statistical Meeting (Salt Lake City, August 2007), NBER-NSF Time-Series (Iowa City, Sept. 2007), CIRANO Data Revision in Macroeconomic Forecasting and Policy (Montreal, Oct. 2007), Society for Nonlinear Dynamics and Econometrics (Atlanta, April 2009), UCR Business Cycles (Riverside, April 2009), Econometric Society (Boston, June 2009), and Midwest Macro (Nashville, May 2011). The paper represents the author's views and does not necessarily represent any official positions of the Bureau of Labor Statistics.

**Affiliated as Research Fellow with Center for Financial Studies (CFS), Johann Wolfgang Goethe University, Frankfurt, Germany, and with Center for Economic Studies and Ifo Institute for Economic Research (CESifo), Munich, Germany.

## 1. Introduction.

An initial estimate (ICPI) of chained CPI of urban consumers (CCPIU) is produced and released to the public every month with a month's delay, and final revised estimates of CCPIU (FCPI) are released every February for all months two calendar years earlier. Depending on the month, a final release comes 14-25 months after an initial release. ICPI differs from FCPI on average by about 22% (see the column average of the BENCH model in table 5). The difference occurs because final estimates of expenditure weights are available for a given year only two calendar years later. ICPI is an initial estimate of FCPI that uses two-year-old expenditure data. The present paper describes and illustrates a method for estimating FCPI concurrently, i.e., in the month of an ICPI release, and evaluates the accuracy of the obtained estimates in terms of root mean-squared errors (RMSE). In the method, FCPI estimates are based on estimated regression models and estimated vector autoregressive moving-average (VARMA) models.

In the regression branch of the method, FCPI is given directly by an estimated regression line. In the VARMA branch of the method, FCPI is estimated by applying the Kalman smoother to an estimated VARMA model. Because FCPI is released every February for two calendar years earlier (not in every month with a constant delay) and is observed in relatively few periods, in practice, a regression can include only current and past ICPI as explanatory variables for estimating current FCPI. By contrast, because a VARMA model and the Kalman smoother can easily handle data delays, in particular, annual February releases of FCPI, they can easily use all current and past ICPI and FCPI data when estimating current FCPI, which explains why the VARMA-Kalman-smoothing estimates of FCPI here are more accurate than the regression-line estimates.

The application here uses monthly data from January 1998 to December 2005, where the 1998-1999 data are unofficial and the 2000-2005 data are official. The results (figure 2, tables 5-6) show that the best VARMA model RVAR12 and its associated Kalman smoother estimate of FCPI imply RMSEs of 17% or 23% lower than the RMSE of 22% of ICPI as an estimate of FCPI. However, even VARMA model RVAR0, which can be produced automatically (i.e., without any judgement or intervention) produces a RMSE of 19% or 14% lower than the RMSE of 22% of ICPI. These results suggest that using the Kalman smoother is more important for accurately estimating FCPI than which particular model is used. For example, even though VARMA model RVAR0 and regression model UREG0 are theoretically equivalent, the VARMA-Kalman-smoothing estimates of FCPI yield a RMSE of 19%, whereas the regression-line estimates of FCPI yield a RMSE of 23% (see column averages of UREG0 and RVAR0 models in table 5).

We consider unrestricted and restricted regression models of FCPI regressed on current and lagged ICPI and unrestricted and restricted VARMA models of ICPI and FCPI. In unrestricted models, all coefficients of variables and disturbances lagged 1-12 months are estimated; in restricted models, insignificant coefficients are set to zero in rounds of estimation. To limit the number of estimated models, only pure VAR and pure VMA models are considered.

Regression and VAR models are estimated using linear methods such as ordinary least squares (OLS) and seemingly unrelated regression (SUR). The VMA model is estimated using the nonlinear maximum likelihood estimation (MLE) method. Whereas regression estimates of FCPI are based only on current and past ICPI and a constant regression-line "formula," VARMA-Kalman-smoothing estimates of FCPI are based on current and past ICPI and FCPI and on a monthly-varying estimation "formula," because FCPI is released only in February. In principle, regression estimates could be based on separate regression lines for each month, but there is not enough data to estimate a separate regression for each month. By contrast, a Kalman smoother automatically produces a different estimation "formula" every month, even for a time-invariant model.

Interim estimates of CCPIU (NCPI) are also released every February for the months of the previous calendar year. However, NCPI were not used here, because they are not clearly "interim" between ICPI and FCPI: in even-numbered years ICPI and NCPI are identical to within eight decimal digits and differ only in odd-numbered years. The resulting high degree of multicollinearity between ICPI and NCPI suggests that NCPI would contribute little additional information for estimating FCPI beyond that in ICPI.

The paper continues as follows. Section 2 discusses data in historical versus real-time forms, both of which the application uses. Section 3 discusses transforming the ICPI and FCPI data by logging, differencing, standardizing, and normalizing them, in order to account for trends, reduce seasonality, and simplify and improve model estimation. Section 4 describes the particular regression and VARMA models to be estimated and used for computing estimates of transformed FCPI. Section 5 reviews some related economics and statistics literature. Section 6 discusses the state-space formulation of estimated VARMA models used for computing FCPI estimates using the Kalman smoother. Section 7 reports estimated regression and VARMA models of transformed ICPI and FCPI data and RMSEs of the FCPI estimates based on the estimated models. Section 8 concludes the paper with a summary. Section 9 contains supplemental figure 2 and table 6.

## 2. Data in Historical and Real-Time Forms.

This section discusses data in historical and real-time forms. The idea is that estimating a model is more easily done using the data and model in more compact historical form, but estimating FCPI with partly delayed data is more easily done using the data and model in expanded real-time form, which accounts explicitly for data delays. The historical form is usually used for storing data and estimating models. In the application, we first estimated models in historical form and, then, converted both data and models to expanded real-time form in order to compute VARMA-Kalman-smoothing estimates of FCPI.

**Table 1: Data in Compact Historical Form.**

| Month t of occurrence | $di_t$ | $df_t$ |
|---|---|---|
| 1 | $di_1$ | $df_1$ |
| 2 | $di_2$ | $df_2$ |
| 3 | $di_3$ | $df_3$ |
| ... | ... | ... |
| 10 | $di_{10}$ | $df_{10}$ |
| 11 | $di_{11}$ | $df_{11}$ |
| 12 | $di_{12}$ | $df_{12}$ |
| End of year t = 1,...,12 | | |
| 13 | $di_{13}$ | $df_{13}$ |
| 14 | $di_{14}$ | $df_{14}$ |
| 15 | $di_{15}$ | $df_{15}$ |
| ... | ... | ... |
| 22 | $di_{22}$ | $df_{22}$ |
| 23 | $di_{23}$ | $df_{23}$ |
| 24 | $di_{24}$ | $df_{24}$ |
| End of year t = 13,...,24 | | |

Table 1 depicts $di_t = \ln(ICPI_t)-\ln(ICPI_{t-1})$ and $df_t = \ln(FCPI_t)-\ln(FCPI_{t-1})$ in compact historical form. In table 1, all instances of t in all columns denote the ***historical month of occurrence of a datum***, regardless when it was released. Table 2 depicts $di_t$ and $df_t$ in expanded real-time form. As in table

1, t subscripts of $di_t$ and $df_t$ in table 2 denote the historical month of occurrence of a datum, but, in contrast to table 1, t in column 1 of table 2 denotes the ***real-time month of observation of a datum***.

**Table 2: Data in Expanded Real-Time Form.**

| Month t of observation | $di_t$ | $df_t$ | | |
|---|---|---|---|---|
| 1 | $di_1$ | na | ... | Na |
| 2 | $di_2$ | $df_{-12}$ | ... | $df_{-23}$ |
| 3 | $di_3$ | na | ... | Na |
| ... | ... | na | ... | Na |
| 10 | $di_{10}$ | na | ... | Na |
| 11 | $di_{11}$ | na | ... | Na |
| 12 | $di_{12}$ | na | ... | Na |
| End of year t = 1, ..., 12 | | | | |
| 13 | $di_{13}$ | na | ... | na |
| 14 | $di_{14}$ | $df_0$ | ... | $df_{-11}$ |
| 15 | $di_{15}$ | na | ... | na |
| ... | ... | na | ... | na |
| 22 | $di_{22}$ | na | ... | na |
| 23 | $di_{23}$ | na | ... | na |
| 24 | $di_{24}$ | na | ... | na |
| End of year t = 13, ..., 24 | | | | |

From BLS's viewpoint, $di_t$ is released in the same month in which it occurs, with no monthly delay. Consequently, for $di_t$, there is no distinction between historical and real time, so that the simultaneously historical and real-time values of $di_t$ are in the same rows in tables 1 and 2. By contrast, all 12 monthly values of $df_t$ for a given calendar year are released in February two calendar years later. For example, December values of $df_t$ for a given year are released 14 months later and January values of $df_t$ for the same year are released 25 months later. Because all values of $df_t$ for a given year are released in February two years later and because rows in table 2 are indexed by real-time months of release or observation, we need 12 columns in table 2 in order to depict the 12 monthly values of $df_t$ released each February. Thus, actual values of $df_t$ are placed in months or rows marked 2 and 14, which are the Februaries in table 2. For all other non-February months or rows, values of $df_t$ are denoted "na," which means not available or missing. Consider, for example, the second February or row t = 14 in table 2. In principle, table 2 has 14 columns, so that the dots represent columns 4-13 which are not shown. Row t = 14 in table 2 contains $df_0$, ..., $df_{-11}$, which denote observations on $df_t$ which occured, respectively, in months t = 0, ..., -11 or December to January two calendar years earlier.

**3. Data Transformation, Trend, and Seasonality.**

We now explain how and why we transformed data by logging, differencing, standardizing, and normalizing them in order to account for trends, reduce seasonality, and simplify and improve model estimation.

The $ICPI_t$ and $FCPI_t$ data were obtained from January 1998 to December 2005 (1998:1 - 2005:12) or 96 months. An initial "in-sample estimation" period of 1998:1 - 2002:12 or 60 months was used to estimate models; a subsequent "out-of-sample forecasting" period of 2003:1 – 2005:12 or 36 months was used to estimate $df_t$ and evaluate its accuracy. Before being used for estimating models and $df_t$, the $ICPI_t$ and $FCPI_t$ data were transformed to normalized, standardized, month-to-month first differences of natural logarithms. The transformed data are graphed in figure 1 to get an idea of the nature and extent of autocorrelations and seasonality and, thereby, to get an idea of needed lags in the regression and VARMA models.

The given $ICPI_t$ and $FCPI_t$ data were first transformed to $di_t = \ln(ICPI_t)-\ln(ICPI_{t-1})$ and $df_t = \ln(FCPI_t)-\ln(FCPI_{t-1})$. Before being graphed or used in estimation, the differenced-logged data were standardized by subtracting in-sample means and dividing by in-sample standard deviations and were normalized by replacing outliers more than 3 standard deviations from zero with a missing-data indicator. To make the standardization and normalization realistic, the out-of-sample data were also standardized using the in-sample means and standard deviations, but were not normalized, because, when standing at the end of an in-sample period, ready to estimate $df_t$ out of sample, we should proceed as if we did not know out-of-sample means, standard deviations, and outliers. Accordingly, the realistic practice is to estimate $df_t$ on the assumption that in-sample and out-of-sample data are generated by the same process and to standardize in- and out-of-sample data using the same in-sample means and standard deviations. Estimating models and $df_t$ using standardized data also results in simpler estimation because constant terms are zero.

Removing outliers moves standardized data closer to the standard-normal or N(0,1) distribution. The idea is that, because estimation aims primarily to account for the systematic or nonrandom parts of a data generating process, the estimation improves if outliers are first removed. Leaving outliers in out-of-sample data also makes evaluating estimates of $df_t$ more realistic. For example, figure 2 and table 6 in the appendix report two large unremoved outliers in September and November 2005, due to Hurricane Katrina.

**Figure 1: Transformed Data in Historical Form, Autocorrelations, and Spectra.**

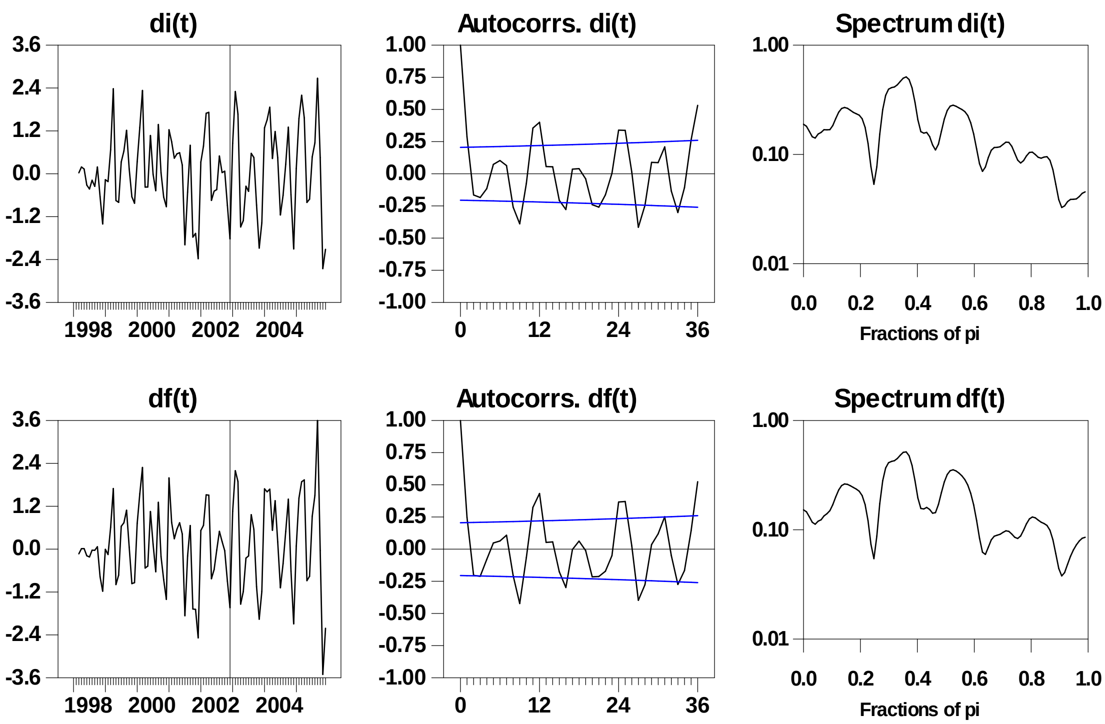



Figure 1 contains 6 graphs: the top 3 graphs depict standardized and normalized $di_t$, its autocorrelations, and its spectrum; the bottom 3 graphs depict standardized and normalized $df_t$, its autocorrelations, and its spectrum. Vertical lines in the leftmost graphs divide the sample into in- and out-of-sample periods. Unshown graphs of original data ($ICPI_t$ and $FCPI_t$) display common upward trends. Leftmost graphs of data show that $di_t$ and $df_t$ are approximately stationary over the sample, with no discernable trends and slightly increasing variance. Middle graphs of autocorrelations of data show significant seasonality, because autocorrelations at seasonal lags of 6, 12, 24, and 36 months are significant (outside of two-standard error confidence bounds about zero). Rightmost graphs of spectra of data, interpreted in terms of table 3, more precisely indicate the nature of the data's seasonality.

Table 3 indicates frequencies and periodicities of harmonic monthly seasonal cycles, with angular frequencies of $\pi/6$ = .5236, ..., $\pi$ = 3.142 radians and periodicities of 12, ..., 2 months. The rightmost graphs in figure 1 show that nearly identical seasonality of $di_t$ and $df_t$ is dominated by seasonal cycles with periodicities of 4, 6, and 12 months.

**Table 3: Frequencies and Periodicities of Harmonic Monthly Seasonal Cycles.**

| Cases | Frequency radians | Frequency $\pi$ radians | Frequency Cycles/mon | Period mon/cycle |
|---|---|---|---|---|
| 1 | .0000 | 0 | .0000 | $\infty$ |
| 2 | .5236 | 1/6 | .0833 | 12 |
| 3 | 1.047 | 1/3 | .1667 | 6 |
| 4 | 1.571 | 1/2 | .2500 | 4 |
| 5 | 2.094 | 2/3 | .3333 | 3 |
| 6 | 2.618 | 5/6 | .4167 | 12/5 |
| 7 | 3.142 | 1 | .5000 | 2 |

In sum, significant seasonality of $di_t$ and $df_t$ in figure 1 suggests that estimated regression and VARMA models should include lagged values up to about 12 months back in order to fit the stationary and seasonal $di_t$ and $df_t$ data adequately.

## 4. Estimated Regression and VARMA Models.

We now define the eight regression and VARMA models. The four regression models are denoted BENCH, UREG0, UREG12, and RREG12 and the four VARMA models are denoted RVAR0, UVAR12, RVAR12, and UVMA12. In section 7, we do not report any restricted-to-zero or estimated parameters of any models, because the parameters have no particular meanings in the context of the application.

Regression model BENCH, whose name means ***benchmark regression*** is defined by

(4.1) $$df_t = di_t + \varepsilon_{ft},$$

with disturbance $\varepsilon_{ft}$ and disturbance variance $\sigma_{ff}$. Although BENCH has no estimated coefficients, we think of it as regression model UREG0 with coefficient $\beta_0$ restricted to one. BENCH has only parameter $\sigma_{ff}$ to be estimated.

All the regression models have the same assumptions on the disturbances, $\varepsilon_{ft}$, that we now make for BENCH: (i) $\varepsilon_{ft}$ is distributed normally, identically, independently, with mean zero, and constant positive variance, $\sigma_{ff} > 0$, or $\varepsilon_{ft} \sim \text{NIID}(0,\sigma_{ff})$; and, (ii) $\varepsilon_{ft}$ is distributed independently of current and lagged regressors, $di_t$, ..., $di_{t-12}$.

Regression model UREG0, whose name means ***unrestricted regression with 0 lags***, is defined by

(4.2) $$df_t = \beta_0 di_t + \varepsilon_{ft},$$

with 2 parameters, $\beta_0$ and $\sigma_{ff}$, estimated using ordinary least squares (OLS).

Regression model UREG12, whose name means ***unrestricted regression with 12 lags***, is defined by

(4.3) $$df_t = \beta_0 di_t + \beta_1 di_{t-1} + \ldots + \beta_{12} di_{t-12} + \varepsilon_{ft},$$

with 14 parameters, $\beta_0$, ..., $\beta_{12}$, and $\sigma_{ff}$, estimated using OLS.

Regression model RREG12, whose name means ***restricted regression with up to 12 lags***, is defined by

(4.4) $$df_t = \beta_0 di_t + \beta_1 di_{t-1} + \beta_4 di_{t-4} + \beta_9 di_{t-9} + \varepsilon_{ft},$$

with 5 parameters, $\beta_0$, $\beta_1$, $\beta_4$, $\beta_9$, and $\sigma_{ff}$, estimated using OLS, as follows. In estimated UREG12, we set to zero the least significant estimated coefficient, with the highest marginal significance level or p value; we reestimated the resulting reduced regression using OLS; we set to zero the least significant resulting estimated coefficient; we continued like this until all estimated coefficients were significant at about the 10% level or $|t_{\hat{\beta}}| \geq 1.645$; and, we estimated $\sigma_{ff}$ residually at the last step.

VARMA model RVAR0, whose name means ***restricted VAR model with no lags***, is defined by

(4.5) $y_t = \varepsilon_t$,

for bivariate data vector $y_t = (di_t,\ df_t)^T$ and bivariate disturbance vector $\varepsilon_t = (\varepsilon_{it},\ \varepsilon_{ft})^T$, where superscript T denotes vector or matrix transposition.

All the VARMA models have the same assumptions on disturbances that we now make for model RVAR0: (i) disturbance vector $\varepsilon_t$ is distributed normally, identically, independently, with mean zero, and constant, 2×2, symmetric, positive definite, covariance matrix, $\Sigma_\varepsilon = \begin{bmatrix} \sigma_{ii} & \sigma_{if} \\ \sigma_{if} & \sigma_{ff} \end{bmatrix}$, or $\varepsilon_t \sim \mathrm{NIID}(0,\Sigma_\varepsilon)$; and, (ii) $\varepsilon_t$ is distributed independently of any past variables or disturbances, $y_{t-1}$, $\varepsilon_{t-1}$, .... In RVAR0, because standardized $di_t = \varepsilon_{it}$ implies $\sigma_{ii} = 1$, only two parameters, $\sigma_{if}$ and $\sigma_{ff}$, need to be estimated using MLE.

We consider theoretically equivalent UREG0 and RVAR0 as separate models in order to estimate $df_t$ separately using regression lines and Kalman smoothing. We now illustrate this equivalence by transforming UREG0 to RVAR0. Because both $\beta_0$ of UREG0 and $\sigma_{if}$ of RVAR0 account for correlation between $di_t$ and $df_t$, when converting UREG0 to RVAR0, one of these parameters becomes redundant. Thus, we have $\sigma_{ii} = 1$, can set $\sigma_{if} = 0$, and obtain $y_t = \tilde{\varepsilon}_t$, where $\tilde{\varepsilon}_t \sim \mathrm{NIID}(0,\Sigma_{\tilde{\varepsilon}})$ and $\Sigma_{\tilde{\varepsilon}} = \begin{bmatrix} 1 & \beta_0 \\ \beta_0 & \sigma_{ff} \end{bmatrix}$.

VARMA model UVAR12, whose name means ***unrestricted VAR with 12 lags***, is defined by

(4.6) $y_t = \Phi_1 y_{t-1} + \ldots + \Phi_{12} y_{t-12} + \varepsilon_t$,

where the $\Phi_i$ are unrestricted 2×2 VAR coefficient matrices. The 51 parameters of UVAR12, 48 elements of the $\Phi_i$, and 3 nonredundant elements of $\Sigma_\varepsilon$, were estimated by applying OLS to each scalar equation in (4.6).

VARMA model RVAR12, whose name means ***restricted VAR with up to 12 lags***, is defined by

(4.7) $y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \Phi_4 y_{t-4} + \Phi_8 y_{t-8} + \Phi_{11} y_{t-11} + \varepsilon_t$, where the restrictions are limited to setting to zero certain elements of coefficient matrices $\Phi_i$.

The 14 nonzero parameters of RVAR12 were estimated using SUR, analogous to reducing UREG12 to RREG12, as follows. In each equation of estimated UVAR12, we set to zero the least significant estimated coefficient; we estimated the resulting reduced equations using SUR; we dropped from each estimated equation the least significant estimated coefficient; we reestimated the reduced equations using SUR; we continued like this until all estimated coefficients were significant at about the 10% level; and, we estimated $\Sigma_\varepsilon$ residually at the final step.

VARMA model UVMA12, whose name means un***restricted VMA with 12 lags***, is defined by

(4.8) $$y_t = \Theta_1\varepsilon_{t-1} + \ldots + \Theta_{12}\varepsilon_{t-12} + \varepsilon_t,$$

where the $\Theta_i$ are unrestricted 2×2 VMA coefficient matrices. The 51 parameters of UVMA12, 48 elements of the $\Theta_i$ and 3 nonredundant elements of $\Sigma_\varepsilon$, were estimated simultaneously by applying MLE to (4.8). As usual, the estimated UVMA12 model was restricted to be invertible. We tried to estimate a subsequently reduced VMA model with insignificant coefficients set to zero, analogous to reducing UVAR12 to RVAR12, but this failed because the estimation algorithm stepped into parameter regions of inadequately fitting models and failed to converge.

## 5. Structural News-Noise and Unobserved-Components Models.

So far, we have considered nonstructural models, i.e., not motivated by economic or statistical theories, and now consider news-noise and unobserved-components models as examples of economic and statistical structural models. News-noise models have received much attention in economic discussions of data revisions and unobserved-components models are used often in statistics for modelling time series.

First, if

(5.1) $$df_t = di_t + \xi_t,$$

and $di_t$ and $\xi_t$ are generated by separate, orthogonal, scalar, ARMA processes, then, $\xi_t$ is considered the "news" in $di_t$ as an estimate of $df_t$ and $y_t = (di_t, df_t)^T$ is generated by a restricted, structural, bivariate, ARMA process. Alternatively, if $di_t$ and $df_t$ switch roles,

(5.2) $$di_t = df_t + \xi_t$$

replaces (5.1), and $df_t$ and $\xi_t$ are generated by separate, orthogonal, scalar, ARMA processes, then, $\xi_t$ is considered the "noise" in $di_t$ as an estimate of $df_t$ and $y_t$ is generated by a different, restricted, structural, bivariate, ARMA process. Examples of news-noise models are in Mankiw and Shapiro (1986), Sargent (1989), Kishor and Koenig (2009), and Jacobs and van Norden (2011).

Second, if,

(5.3) $$y_t = e \cdot dp_t + \xi_t,$$

where $y_t = (di_t, df_t)^T$, $e = (1, 1)^T$, and $dp_t$ and $\xi_t$ denote unobserved "true" d-form CCPIU and unobserved noise, generated by orthogonal, univariate and bivariate, ARMA processes, then, (5.3) is an unobserved-components model for $y_t$. Examples of unobserved-component models are in Howrey (1978, 1984), Hillmer and Trabelsi (1987), Trabelsi and Hillmer (1989), Shumway and Katzoff (1991), Patterson (1994), and Chen and Zadrozny (2002).

After assuming that right-side terms in equations (5.1)-(5.3) are generated by VARMA processes, we can restate the equations as VARMA models of $y_t$, subject to nonlinear structural restrictions on their parameters in terms of the parameters of the VARMA processes of the right-side terms. For example, if $di_t$ and $\xi_t$ in equation (5.1) are generated by the orthogonal, univariate, ARMA processes $\alpha(L)di_t = \beta(L)\varepsilon_{1t}$ and $\delta(L)\xi_t = \gamma(L)\varepsilon_{2t}$, where L is the lag operator, the $\varepsilon_t$'s are orthogonal white noises, and $\alpha(L)$, $\beta(L)$, $\delta(L)$, and $\gamma(L)$ are polynomials of finite degree in nonnegative powers of L, then, $y_t$ is generated by $A(L)y_t = B(L)\varepsilon_t$, where $A(L) = \begin{bmatrix} \alpha(L)\delta(L) & 0 \\ 0 & \alpha(L) \end{bmatrix}$, $B(L) = \begin{bmatrix} \delta(L)\beta(L) & \alpha(L)\gamma(L) \\ \beta(L) & 0 \end{bmatrix}$, $\varepsilon_t = (\varepsilon_{1t}^T, \varepsilon_{2t}^T)^T$, and similarly for (5.2) and (5.3).

Although, in principle, a structural model could produce more accurate estimates of $df_t$, the narrow range of $RMSE_f$ in table 5 suggests that this is unlikely here. The estimates of $df_t$ based on estimated VARMA models and Kalman smoothing are, among the references cited here, most closely related to the smoothed estimates in Shumway and Katzoff (1991).

## 6. Real-Time Expanded State-Space Formulation for Kalman Smoothing.

We estimated models using the data and models in compact historical forms (table 1; equations (4.1)-(4.8)) and estimated $df_t$ using the data and models in expanded real-time forms. For the second step, we needed the unrestricted VARMA(26,12) model

$$y_t = \Phi_1 y_{t-1} + \ldots + \Phi_{26} y_{t-26} + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \ldots + \Theta_{12} \varepsilon_{t-12}, \tag{6.1}$$

with 26 AR lags and 12 MA lags, in order to express VARMA models (4.5)-(4.8) as special cases of (6.1) and to account for up to 25-month delays in observing $df_t$.

To compute Kalman-smoothed estimates of $df_t$ in terms of (6.1), we needed to express it in state-space form, which comprises a state equation, which specifies the dynamics of state variables, and an observation equation, which specifies how state variables are observed. Accordingly, we wrote (6.1) as state equation

$$x_t = F x_{t-1} + G\varepsilon_t, \tag{6.2}$$

$$F = \begin{bmatrix} \Phi_1 & \cdots & \cdots & \cdots & \cdots & \Phi_{26} & \Theta_1 & \cdots & \cdots & \cdots & \cdots & \Theta_{12} \\ I & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots & \vdots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & & & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 & \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & 0 & I & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & & \vdots & I & \ddots & & & & \vdots \\ \vdots & & & & & \vdots & 0 & \ddots & \ddots & & & \vdots \\ \vdots & & & & & \vdots & \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & & & \vdots & \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 & I & 0 \end{bmatrix}, \quad G = \begin{bmatrix} I \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

where $x_t = (y_t^T, \ldots, y_{t-25}^T, \varepsilon_t^T, \ldots, \varepsilon_{t-11}^T)^T$ is a 76×1 state vector, F is a 76×76 transition matrix, G is a 76×2 input matrix, and I and 0 are 2×2 identity and zero matrices.

The data vector is $\overline{y}_t = (y_t^T, \ldots, y_{t-25}^T)^T$ and has observation equation

(6.3) $\overline{y}_t = Hx_t$,

where $H = [I_{52x52}, 0_{52x24}]$ is the 52×76 observation matrix and $I_{52x52}$ and $0_{52x24}$ denote 52×52 identity and 52×24 zero matrices. An observation equation can also have an additive observation error, although this isn't needed here. Zero restrictions on $\Phi_i$ and $\Theta_i$ reduce (6.2)-(6.3) to an expanded real-time state-space representation of one of models (4.5)-(4.8) and account for up to 25-month delays in observing $df_t$.

Because models (4.5)-(4.8) have at most 12 AR lags, in the application we set $\Phi_{13}, \ldots, \Phi_{26}$ to zero in the state equation, but keep lags 13-26 of $y_t$ in the state vector and the state equation in order to account for 14-25 month delays in observing $df_t$. Complete state equation (6.2) needs to be used only for Kalman smoothing with VMA model UVMA12. Otherwise, for pure VAR models RVAR0, RVAR12, and UVAR12, we restricted $x_t$ to $\overline{y}_t$ and used only the upper-left 52×52 quadrant of F and the 52×2 upper part of G.

Let $df_{t|t} = E[df_t|I_t]$ denote the Kalman-smoothed estimate of $df_t$, where $E[df_t|I_t]$ denotes the expectation of $df_t$ conditional on current information $I_t$, which comprises current and past observations on $di_t$ and $df_t$ and the estimated model. For any k = 1, ..., 25, $x_{2k+2,t}$ is element 2k+2 of $x_t$ and equals $df_{t-k}$, so that

(6.4) $df_{t|t} = E[x_{2k+2,t+k}|I_t]$.

Thus, smoothed estimates $df_{t|t}$ can be computed, for any k = 1, ..., 25, as k-month-ahead forecasts of element 2k+2 of $x_t$. Genuine forecasts, $df_{s|t}$, for $s > t$, and smoothed estimates, $df_{s|t}$, for $t - 25 \le s < t$, can also be computed as forecasts of elements of $x_t$. $RMSE_f$ in table 5 reflect Kalman-smoothed estimates of $df_t$ made in out-of-sample months, but could alternatively reflect Kalman smoothed estimates of $df_t$ made in the last out-of-sample month.

Anderson and Moore (1979) thoroughly discuss Kalman smoothing and smoothing in cases of no missing data. In the present real-time analysis, $df_t$

is missing in all non-February months, which requires using a missing-data extension of the Kalman smoother, which was first described for scalar time series by Jones (1980), was extended to vector time series by Ansley and Kohn (1983), and was described for and applied to mixed-frequency data by Zadrozny (1988, 1990).

In the application, we computed $df_{t|t} = E[x_{4,t+1}|I_t]$, using $k = 1$, and evaluated its accuracy in terms of standardized root mean-squared errors,

(6.5) $$RMSE_f = \sigma_{f,\eta} / \sigma_{f,f},$$

where $\sigma_{f,f}$ and $\sigma_{f,\eta}$, respectively, denote out-of-sample standard deviations of $df_t$ and of its estimation error, $\eta_t = df_t - df_{t|t}$.

We define $R^2_{e,f}$ and $R^2_{f,f}$ of $df_t$ to reflect the percentage of $df_t$ accounted for by current and lagged $di_t$ and lagged $df_t$ and relate them to $RMSE_f$ in order to establish bounds for evaluating $RMSE_f$ in table 5. Thus, we define in-sample-estimation-period $R^2_{e,f} = 1 - \sigma^2_{e,\varepsilon_f} / \sigma^2_{e,f}$, where first subscript "e" refers to "estimation," $\sigma^2_{e,\varepsilon_f}$ denotes in-sample variance of residuals of $df_t$ and $\sigma^2_{e,f}$ denotes in-sample variance of $df_t$; and, we define out-of-sample-forecasting-period $R^2_{f,f} = 1 - RMSE^2_f$, where first subscript "f" refers to "forecasting."

What values of $RMSE_f$ and $R^2_{f,f}$ are good or bad? We answer this question by stating formal (based on mathematics) and expected (based on heuristics) bounds on $RMSE_f$ and $R^2_{f,f}$. First, by construction, $RMSE_f \geq 0$ and $R^2_{f,f} \leq 1$, which is neither good nor bad. We would like $RMSE_f \cong 0$ and $R^2_{f,f} \cong 1$, which is very good, but, frequently, $RMSE_f > 1$ and $R^2_{f,f} < 0$, which is bad. Second, $df_t$ is estimated efficiently, which is good, only if all conditioning information is utilized fully. If $df_t$ is estimated efficiently, then, $df_{t|t}$ and $\eta_t = df_t - df_{t|t}$ are uncorrelated, $RMSE_f < 1$, and $R^2_{f,f} > 0$. Finally, if the data-generating process is constant over the whole sample period, then, we expect (heuristically) that $RMSE_f \cong \sqrt{1 - R^2_{e,f}}$ and that $R^2_{f,f} \cong R^2_{e,f}$; otherwise, if the data-generating process changes from the in-sample to the out-of-sample periods, then, we expect that $RMSE_f > \sqrt{1 - R^2_{e,f}}$ and $R^2_{f,f} < R^2_{e,f}$. Thus, we expect (heuristically) that

$$(6.6) \qquad RMSE_f \geq \sqrt{1 - R^2_{e,f}}$$

or, equivalently, that $R^2_{f,f} \leq R^2_{e,f}$. If $RMSE_f \cong$ or $< \sqrt{1 - R^2_{e,f}}$, then, $df_t$ is estimated (in the out-of-sample period, using an in-sample estimated model), respectively, about as well as or better than we can expect.

**7. Computing and Evaluating Real-Time Smoothed Estimates of $df_t$.**

Table 4 reports statistics from estimating the models using 60 months of data for 1998:1 - 2002:12: in-sample or estimation-period $R^2$ of $df_t$ ($R^2_{e,f}$), Akaike's (1973) information criterion (AIC), Schwarz's (1978) Bayesian information criterion (BIC), Ljung-Box Q statistics for testing serial correlations of residuals of $di_t$ and $df_t$ at lags 1-36, and their marginal significance levels or p values underneath. To be compatible with the $R^2_{e,f}$ of the regression models, the $R^2_{e,f}$ of a VARMA model was increased to include the explanatory effect of $di_t$ on $df_t$, already in the $R^2_{e,f}$ of the regression models. This was done by taking the basic VARMA $R^2_{e,f}$, which accounts for variations in current $df_t$ in terms of variations in lagged $di_t$ and $df_t$ and adding the effects of variations in current $di_t$, according to the estimated correlation between the $di_t$ and $df_t$ residuals and the transformation between the UREG0 and RVAR0 models discussed in section 4. Because of their equivalence, the UREG0 and RVAR0 models have identical IC and Q statistics.

The estimated model that minimizes AIC and BIC is considered the best one. Table 4 reports that RVAR12 has the lowest AIC and that RVAR0 has the lowest BIC. BIC has the more stringent penalty function and usually prefers the model with fewer parameters. Except for $df_t$ in UREG12, $di_t$ in UVAR12, and $df_t$ in UVMA12, all residuals in table 4 have highly significant Q statistics. However, because the significant residual serial correlations do not match those of VARMA models, adding lags of variables or disturbances seems unlikely to produce lower Q statistics. Thus, we considered no other estimated models and picked RVAR0 and RVAR12 as the best models. However, any choice of best models based on in-sample AIC or BIC is tentative, because the final objective is to minimize out-of-sample $RMSE_f$.

**Table 4: Regression and VARMA model summary estimation statistics.**

| Model Number | Model Name | Estm Mthd | $R^2_{e,f}$ | | AIC | BIC | #Est Pars | $Q_i$ | $Q_f$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | BENCH | --- | .9472 | | --- | --- | 0 | 178.3<br>.0000 | 138.9<br>.0000 |
| 2 | UREG0 | OLS | .9479 | | -143.4 | -139.2 | 2 | 178.3<br>.0000 | 138.9<br>.0000 |
| 3 | RREG12 | OLS | .9693 | | -160.5 | -151.0 | 5 | 178.3<br>.0000 | 62.83<br>.0002 |
| 4 | UREG12 | OLS | .9742 | | -151.2 | -121.8 | 14 | 178.3<br>.0000 | 77.54<br>.8027 |
| | | | | | | | | | |
| 5 | RVAR0 | MLE | .9576 | | -143.4 | -139.2 | 2 | 178.3<br>.0000 | 138.9<br>.0000 |
| 6 | RVAR12 | SUR | .9795 | | -164.1 | -134.8 | 14 | 65.79<br>.0001 | 72.80<br>.0000 |
| 7 | UVAR12 | OLS | .9845 | | -150.6 | -43.84 | 51 | 36.05<br>.1414 | 64.10<br>.0001 |
| 8 | UVMA12 | MLE | .9327 | | -99.62 | 7.192 | 51 | 61.63<br>.0003 | 20.15<br>.8591 |

The $RMSE_f$ results in table 5 lead to the following seven conclusions.

1. BENCH is the best regression model, with the lowest $RMSE_f$ average and spread (over the twelve months) of .2167 and .3791; and, RVAR12 is the best VARMA model and best overall model, with the lowest $RMSE_f$ average and spread of .1746 and .3674. RVAR12 has a 19.4% lower $RMSE_f$ average and a 3.1% lower $RMSE_f$ spread than BENCH.

2. The close $RMSE_f$ of RVAR0 and RVAR12 suggest that accuracy of $df_{t|t}$ depends mostly on correlations between contemporaneous $di_t$ and $df_t$ and on delays in observations and less on correlations across current and lagged $di_t$ and $df_t$ as accounted for by RVAR12 but not by RVAR0.

3. Except for UVAR12, the VARMA models produce lower $RMSE_f$ average and spread than the regression models, either because the VARMA models are better models or because the Kalman smoother is applied to them. For the same model, the Kalman smoother should estimate $df_t$ more accurately than a regression line, because it uses all current and past initial and final data, whereas a regression line uses only current and past initial data. Comparing $RMSE_f$ of

equivalent UREG0 and RVAR0 models illustrates this point: it seems to matter less which particular model is used than that the Kalman smoother is used.

**Table 5: $RMSE_f$ of smoothed estimates of $df_t$ occuring in 2003:1-2005:12.**

| Month | BENCH | UREG0 | RREG12 | UREG12 | | RVAR0 | RVAR12 | UVAR12 | UVMA12 | Row Avrg |
|---|---|---|---|---|---|---|---|---|---|---|
| Jan | .1918 | .2076 | .1489 | .1423 | | .1856 | .0965 | .1958 | .1689 | .1672 |
| Feb | .0832 | .0745 | .1363 | .1797 | | .0609 | .0887 | .2215 | .0800 | .1156 |
| Mar | .1916 | .1801 | .2223 | .2515 | | .1478 | .1677 | .2457 | .1567 | .1954 |
| Apr | .1660 | .1892 | .3356 | .2835 | | .1637 | .2471 | .3409 | .1872 | .2392 |
| May | .0984 | .1083 | .0553 | .0559 | | .0914 | .1203 | .1735 | .0976 | .1001 |
| Jun | .1210 | .1157 | .1914 | .1557 | | .0984 | .0692 | .1362 | .1096 | .1247 |
| Jul | .2158 | .2186 | .1338 | .1681 | | .2039 | .1513 | .1824 | .1761 | .1813 |
| Aug | .3027 | .3130 | .2766 | .2355 | | .2994 | .2382 | .2777 | .2901 | .2792 |
| Sep | .4623 | .4921 | .4848 | .4962 | | .4430 | .4561 | .5204 | .4697 | .4781 |
| Oct | .2276 | .2311 | .1125 | .1258 | | .2008 | .1086 | .1683 | .1815 | .1695 |
| Nov | .3788 | .4030 | .4046 | .3995 | | .3479 | .2562 | .3133 | .3336 | .3546 |
| Dec | .0975 | .1023 | .1479 | .1334 | | .0837 | .0952 | .0952 | .0969 | .1065 |
| Col Avrg | .2167 | .2268 | .2259 | .2238 | | .1939 | .1746 | .2392 | .1957 | .2121 |
| Sprd | .3791 | .4176 | .4295 | .4403 | | .3821 | .3674 | .4252 | .3897 | |

Spread = maximum $RMSE_f$ - minimum $RMSE_f$.

4. Average $R^2_{e,f}$ in table 4 of the best $df_t$-estimating VARMA models, RVAR0, RVAR12, and UVMA12, is .9739, which implies $\sqrt{1 - R^2_{e,f}}$ = .1616. RVAR12, RVAR0, and UVMA12 imply, respectively, column-average $RMSE_f$ of .1746, .1939, and .1957, which are not much above .1616 and, hence, suggest that $di_t$ and $df_t$ are generated by a relatively constant process over the whole sample period and that $df_{t|t}$ based on these models, especially on RVAR12, is approximately efficient.

5. The monthly pattern of $RMSE_f$ in table 5 could be seasonal, because the data are seasonal and releases of $df_t$ are seasonal (only in February), or,

it could be random, because the short smoothing period covers only three years. $RMSE_f$ is lowest in May and highest in September and November, the latter because Hurricane Katrina struck in August 2005.

6. The narrow range of $RMSE_f$ in table 5 suggests that other models, such as structural models, are unlikely to produce lower $RMSE_f$ for this data.

7. The $df_{t|t}$ could be computed by Kalman smoothing as in Shumway and Katzoff (1991). Kalman smoothers are extensions of Kalman smoothers with numerous implementations (Anderson and Moore, 1979). Smoothed estimates can be computed using the more compact historical forms of the data and models.

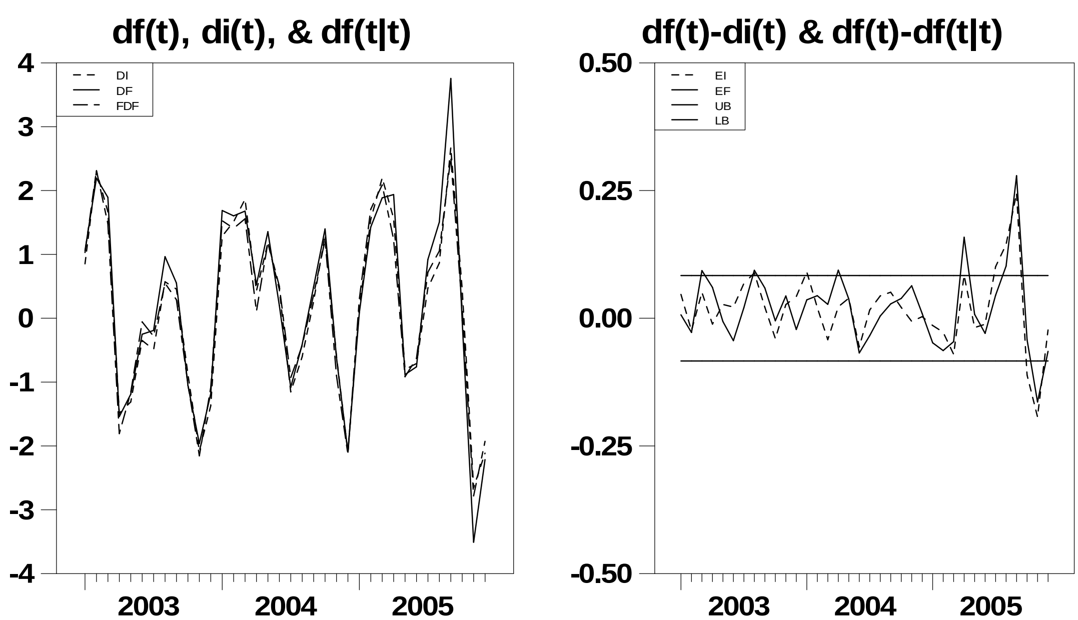
## 8. Conclusion.

The paper has described and applied regression and VARMA modeling methods for estimating the current value of a variable which is observed intermittently with delay and is contemporaneously and serially correlated with another variable which is observed every period without delay. In the regression branch of the method, the delayed variable is regressed on current and lagged values of the undelayed variable and the estimated regression line estimates the current value of the delayed variable. In the VARMA modeling branch of the method, a bivariate VARMA model is estimated for the variables, the Kalman smoother is applied to the estimated model, and an element of the one-step-ahead forecast of the state vector estimates the current value of the delayed variable. The methods were applied to U.S. monthly chained CPI of urban consumers (CCPIU) from January 1998 to December 2009, with initial and final CCPIU, respectively, being the undelayed and delayed variables. The results in table 5 show that estimated VARMA models and Kalman smoothing produced lower average root-mean-squared errors of estimates of final CCPIU than estimated regression models, namely, .2009 compared with .2233 or about 10% lower.

## 9. Supplemental Figure 2 and Table 6.

Figure 2 and table 6 supplement table 5 by graphing and tabulating $df_t$, $di_t$, $df_{t|t}$ based on RVAR12, errors $e_{i,t} = df_t - di_t$ and $e_{f,t} = df_t - df_{t|t}$, and 2-standard-error confidence bounds of $e_{f,t}$. To be consistent with table 5, before being graphed in figure 2, the errors in table 6 and standard-error bounds produced by the Kalman smoother were standardized by division by .2259.

**Figure 2: $df_t$, $di_t$, $df_{t|t}$, $e_{i,t}$, $e_{f,t}$, and 2-standard-error confidence bounds.**




In the left graph, $df_t$, $di_t$, $df_{t|t}$ are depicted, respectively, by solid, evenly-dashed, and unevenly-dashed lines; in the right graph, $df_t$-estimation errors, $e_{i,t} = df_t - di_t$ and $e_{f,t} = df_t - df_{t|t}$, and 2-standard-error confidence bounds of $e_{f,t}$ are, respectively, depicted by evenly-dashed and solid lines.

**Table 6: Numbers Underlying Table 5 and Figure 2.**

| Month | $df_t$ | $di_t$ | $df_{t\|t}$ | $df_t - di_t$ | $df_t - df_{t\|t}$ |
|---|---|---|---|---|---|
| 2003:1 | 1.059426 | 0.852800 | 1.029448 | 0.206626 | 0.029978 |
| 2003:2 | 2.195151 | 2.302536 | 2.317998 | -0.107385 | -0.122847 |
| 2003:3 | 1.893475 | 1.668810 | 1.480396 | 0.224665 | 0.413079 |
| 2003:4 | -1.539022 | -1.488333 | -1.806657 | -0.050689 | 0.267635 |
| 2003:5 | -1.190376 | -1.310190 | -1.162496 | 0.119814 | -0.027880 |
| 2003:6 | -0.250663 | -0.347280 | -0.056592 | 0.096617 | -0.194071 |
| 2003:7 | -0.190663 | -0.491805 | -0.281821 | 0.301142 | 0.091158 |
| 2003:8 | 0.964597 | 0.571549 | 0.548911 | 0.393048 | 0.415686 |
| 2003:9 | 0.551247 | 0.456338 | 0.289162 | 0.094909 | 0.262085 |
| 2003:10 | -1.052098 | -0.876101 | -1.030397 | -0.175997 | -0.021701 |
| 2003:11 | -1.961222 | -2.079998 | -2.154754 | 0.118776 | 0.193532 |
| 2003:12 | -1.182770 | -1.371769 | -1.085642 | 0.188999 | -0.097128 |
| 2004:1 | 1.687414 | 1.286495 | 1.527973 | 0.400919 | 0.159441 |
| 2004:2 | 1.604391 | 1.510979 | 1.408610 | 0.093412 | 0.195781 |
| 2004:3 | 1.678370 | 1.864005 | 1.558038 | -0.185635 | 0.120332 |
| 2004:4 | 0.527502 | 0.427789 | 0.110470 | 0.099713 | 0.417032 |
| 2004:5 | 1.355959 | 1.182507 | 1.188051 | 0.173452 | 0.167908 |
| 2004:6 | 0.201992 | 0.451038 | 0.502721 | -0.249046 | -0.300729 |
| 2004:7 | -1.081764 | -1.153462 | -0.931304 | 0.071698 | -0.150460 |
| 2004:8 | -0.408177 | -0.601900 | -0.427486 | 0.193723 | 0.019309 |
| 2004:9 | 0.490708 | 0.264952 | 0.366244 | 0.225756 | 0.124464 |
| 2004:10 | 1.399156 | 1.304695 | 1.227766 | 0.094461 | 0.171390 |
| 2004:11 | -0.613947 | -0.586540 | -0.895612 | -0.027407 | 0.281665 |
| 2004:12 | -2.090220 | -2.104176 | -2.130870 | 0.013956 | 0.040650 |
| 2005:1 | 0.085752 | 0.148536 | 0.298636 | -0.062784 | -0.212884 |
| 2005:2 | 1.427234 | 1.547942 | 1.707116 | -0.120708 | -0.279882 |
| 2005:3 | 1.885741 | 2.197780 | 2.088557 | -0.312049 | -0.202816 |
| 2005:4 | 1.937359 | 1.566135 | 1.235226 | 0.371224 | 0.702133 |
| 2005:5 | -0.882269 | -0.801716 | -0.916265 | -0.080553 | 0.033996 |
| 2005:6 | -0.760514 | -0.708369 | -0.629521 | -0.052145 | -0.130993 |
| 2005:7 | 0.920044 | 0.473169 | 0.721203 | 0.446875 | 0.198841 |
| 2005:8 | 1.506380 | 0.866402 | 1.053808 | 0.639977 | 0.452571 |
| 2005:9 | 3.755989 | 2.673409 | 2.520455 | 1.082580 | 1.235534 |
| 2005:10 | -0.037918 | 0.454298 | 0.148269 | -0.492216 | -0.186187 |
| 2005:11 | -3.507909 | -2.653479 | -2.782071 | -0.854430 | -0.725838 |
| 2005:12 | -2.213681 | -2.112154 | -1.927204 | -0.101527 | -0.286477 |

**REFERENCES**

Akaike, H. (1973), "Information Theory and Extension of the Maximum Likelihood Principle," pp. 267-281 in Second International Symposium on Information Theory, B.N. Petrov and F. Csaki (eds.), Budapest, Hungary: Akademia Kiado.

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

Ansley, C.F. and R. Kohn (1983), "Exact Likelihood of Vector Autoregressive Moving-Average Process with Missing or Aggregated Data," Biometrika 70: 275-278.

Chen, B. and P.A. Zadrozny (2002), "Real-Time Quarterly Signal-Plus-Noise Model for Estimating 'True' GDP," American Statistical Association 2001 Proceedings of the Business and Economic Statistics Section: 173-178.

Hillmer, S.C. and Trabelsi, A. (1987), "Benchmarking of Economic Time Series," Journal of the American Statistical Association 82: 1064-1071.

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

Howrey, E.P. (1984), "Data Revision, Reconstruction, and Prediction: An Application to Inventory Investment," Review of Economics and Statistics 66: 386-393.

Jacobs, J.P.A.M. and S. van Norden (2011), "Modelling Data Revisions: Measurement Error and Dynamics of 'True' Values," Journal of Econometrics 161: 101-109.

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

Kishor, N.K. and E.F. Koenig (2009), "VAR Estimation and Forecasting When Data are Subject to Revision," Working Paper, Research Department, Federal Reserve Bank of Dallas, Dallas, TX.

Mankiw, N.G. and M.D. Shapiro (1986), "News or Noise: An Analysis of GNP Revisions," Survey Current Business 66: 20-25.

Patterson, K.D. (1994), "A State Space Model for Reducing the Uncertainty Associated with Preliminary Vintages of Data with an Application to Aggregate Consumption," Economics Letters 46: 215-222.

Sargent, T.J. (1989), "Two Models of Measurement and the Investment Accelerator," Journal of Political Economy 97: 251-287.

Schwarz, G. (1978), "Estimating the Dimension of a Model," Annals of Statistics 8: 147-164.

Shumway, R.H. and M.J. Katzoff (1991), "Adjustment of Provisional Mortality Series: The Dynamic Linear Model with Structured Measurement Errors," JOurnal of the American Statistical Association 86: 611-617.

Trabelsi, A. and Hillmer, S.C. (1989), "A Benchmarking Approach to Forecasting Combination," Journal of Business and Economic Statistics 7: 353-362.

Zadrozny, P.A. (1988), "Gaussian Likelihood of Continuous-Time ARMAX Models when Data are Stocks and Flows at Different Frequencies," Econometric Theory 4: 109-124.

Zadrozny, P.A. (1990), "Estimating a Multivariate ARMA Model with Mixed-Frequency Data: An Application to Forecasting U.S. GNP at Monthly Intervals," Working Paper 90-6, Research Department, Federal Reserve Bank of Atlanta, Atlanta, GA.