12-2002

Forecasting with a Real-Time Data Set for Macroeconomists

Tom Stark

Dean D. Croushore
University of Richmond, dcrousho@richmond.edu

FORECASTING WITH A REAL-TIME DATA SET FOR MACROECONOMISTS

Tom Stark and Dean Croushore
Research Department
Federal Reserve Bank of Philadelphia

July 2001
Revised February 2002

The authors thank Sharon Kozicki, Graham Elliott, Jim Fackler, Ellis Tallman, Tao Zha, and participants at meetings of the 2001 and 2002 American Economic Association, the 2001 Society for Computational Economics, and the 2001 American Statistical Association for their comments. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Philadelphia or of the Federal Reserve System.

FORECASTING WITH A
REAL-TIME DATA SET FOR MACROECONOMISTS

Abstract

This paper discusses how forecasts are affected by the use of real-time data rather than latest-available data. The key issue is this: In the literature on developing forecasting models, new models are put together based on the results they yield using the data set available to the model's developer. But those are not the data that were available to a forecaster in real time.

How much difference does the vintage of the data make for such forecasts? We explore this issue with a variety of exercises designed to answer this question. In particular, we find that the use of real-time data matters for some forecasting issues but not for others. It matters for choosing lag length in a univariate context. Preliminary evidence suggests that the span—or number—of forecast observations used to evaluate models may also be critical: we find that standard measures of forecast accuracy can be vintage-sensitive when constructed on the short spans (five years of quarterly data) of data sometimes used by researchers for forecast evaluation. The differences between using real-time and latest-available data may depend on what is being used as the "actual" or realization, and we explore several alternatives that can be used. Perhaps of most importance, we show that measures of forecast error, such as root-mean-squared error and mean absolute error, can be deceptively lower when using latest-available data rather than real-time data. Thus, for purposes such as modeling expectations or evaluating forecast errors of survey data, the use of latest-available data is questionable; comparisons between the forecasts generated from new models and benchmark forecasts, generated in real time, should be based on real-time data.

FORECASTING WITH A REAL-TIME DATA SET FOR MACROECONOMISTS

I. INTRODUCTION

When economists develop forecasting models, they base their decisions about model structure on the outcomes of testing alternatives using the most recent vintage of historical data available. Within the forecasting literature, it is common to see papers in which researchers generate forecasts using a new model, then comparing the resultant forecast errors to those from alternative models. Occasionally they even compare their new forecasts to those that were made by others in real time or to a consensus forecast such as that from the Philadelphia Fed's Survey of Professional Forecasters. However, in doing so, the model developers have given themselves two advantages: (1) they know what the data look like ex-post, so they have a richer data set for building a model; and (2) the data they are using have been revised over time and differ significantly from the data used by forecasters in real time. There is not much we can do to eliminate the first advantage, since that is inherent to the development of forecasting models and the progress of knowledge. But the second advantage may be looked at more critically, since it means that the new model being developed may have generated much worse forecasts in real time than with latest-available data. To investigate this issue, we have created a real-time data set for macroeconomists, so forecast developers can check their work and be sure that their "new and improved" forecasting models would really have worked with the data that forecasters faced in real time.

Some researchers may wonder: why is it necessary to check a new model's performance against real-time data? After all, aren't the latest-available data the best data to use for testing new models? Our answer is—not necessarily. All models depend on key specification choices that are motivated by economic theory. In the context of VARs, for example, the choice may be as simple as whether to include a particular variable in the system. Additional choices may include whether to model in levels or first-differences, how to impose certain cointegrating relations, and which identifying restrictions are needed. Since these choices are rarely motivated by issues concerning data revisions, we would expect sound model specifications to perform well regardless of the data vintage. This suggests that ex-post forecast simulation is an important part of model-testing. Indeed, the forecasting profession has already adopted this idea, in part, because it is now standard practice to evaluate new models on the basis of their out-of-sample forecast performance. However, that is almost always done on the basis of the latest-available data. We argue that the profession should go one step further, by evaluating models on the basis of their out-of-sample forecast performance *using real-time data*.

Our real-time data set for macroeconomists is basically a snapshot of the macroeconomic data available at any given date in the past. We use the term "vintage" to mean the data available at any particular date in the past. For example, the first vintage in our data set is November 1965, which consists of the quarterly and monthly data that were available to a forecaster on November 15, 1965.

In an earlier paper (Croushore and Stark, 2001) we described the data set in great detail, explored the nature of the revision process of the data, both long term and short term, and provided some basic forecasting results. In another paper (Croushore and Stark, 1999b), we looked at how some key results in the macroeconomic literature are affected by the choice of vintage. In some cases, there is a significant overturning of the results.

In this paper, our focus is on forecasting. We run through a variety of forecasting examples and develop some ways of thinking about how the choice of using real-time data versus latest-available data matters.

II. RELATED RESEARCH

The early research in this field focused on the use of preliminary rather than final data. Cole (1969) provided evidence that certain types of forecasts were strongly affected by data revisions. She found that "the use of preliminary rather than revised data resulted in a *doubling* of the forecast error." More recently, Diebold and Rudebusch (1991) looked at the importance of data revisions for the index of leading indicators. They found that the use of real-time data was crucial, since the variables included in the index of leading indicators were chosen ex-post. In real time, the leading indicators neither indicate, nor lead! Their results were supported by the work of Robertson and Tallman (1998a), who showed that a VAR that uses real-time data from the index of leading indicators produces no better forecasts for industrial production than an AR model using just lagged data on industrial production. However, they also showed that the leading indicators may be useful in forecasting real output (GNP/GDP) in real time.

Numerous researchers have voiced concerns about data revisions. Denton and Kuiper (1965) agreed with Cole's results that the use of preliminary (rather than final) data led to large forecast errors, but Trivellato and Rettore (1986) found effects that were much more modest. Stekler (1967) found that there was some value in early data releases even though they contained errors. Howrey (1978) noted that it was important to adjust for the fact that data within a particular vintage have been revised to differing degrees. Howrey (1996) compared forecasts of the level of GNP to forecasts of growth rates, finding the former more sensitive to data revisions than the latter. Swanson (1996) investigated optimal model selection for a number of variables using real-time data. Robertson and Tallman (1998b) evaluated alternative VAR model specifications for forecasting major macroeconomic variables using a real-time data set. Real-time data have also been used by Koenig and Dolmas (1997) and Koenig, Dolmas, and Piger (1999) in developing a method for forecasting real output growth using monthly data.

In many cases, the results of these studies are limited to specific—and short—sample periods because the authors do not have a complete real-time data set. In this paper, we explore some of the same forecasting issues—but with a collection of vintages that span a much longer sample period, thus allowing for additional generality of our results.

III. THE DATA AND AN ANALYTICAL EXAMPLE

The data set is described in great detail in Croushore and Stark (2001). This section briefly describes the data set and illustrates a few ways in which data revisions may be important.

Each vintage of the data set consists of a time-series of the data, with observations from 1947:1 to the quarter before the vintage date, as they existed in the middle of each quarter (on the 15th day of the month). Vintages span November 1965 to the present, though in this paper our empirical work began in August 1999, so that is the last vintage we use.

As one moves from one vintage to the next, two things happen: (1) The later vintage contains additional observations (usually one additional observation for each variable); and (2) The later vintage incorporates any revisions to the observations common to both.

The real-time data set includes the following variables:[1]

Quarterly observations:

- Nominal GNP (vintages before 1992) or GDP (vintages in 1992 and after)
- Real GNP (vintages before 1992) or GDP (vintages in 1992 and after) and components:
  - Consumption and its components:
    - Durables
    - Nondurables
    - Services
  - Components of Investment:
    - Business Fixed Investment
    - Residential Investment
    - Change in Business Inventories (Change in Private Inventories after vintage August 1999)
  - Government Purchases (Government Consumption and Gross Investment, vintages in 1996 and after)
  - Exports
  - Imports
- Chain-Weighted GDP Price Index (vintages in 1996 and after)
- Corporate Profits
- Import Price Index

Monthly observations:
(quarterly averages of these variables are also available in the quarterly data sets)

- Money supply measures:
  - M1
  - M2
- Reserve measures (data from Board of Governors):
  - Total reserves
  - Nonborrowed reserves
  - Nonborrowed reserves plus extended credit
  - Monetary base
- Civilian Unemployment Rate

[1] All the data are available on the web: http://www.phil.frb.org/econ/forecast/reaindex.html. Note that in some vintages of the data, the entire history of the data back to 1947 were not available. In these situations, in this paper we use the last known output growth rates or inflation rates for the periods for which complete data were not known.

Consumer Price Index (seasonally adjusted)
3-month T-bill Rate
10-year T-bond Rate

It is natural to ask: How much does the data vintage matter? In our previous research, we have shown that data revisions may matter significantly for some problems. In Croushore and Stark (2001), we show that data revisions may be significant in both the long run and the short run. For example, suppose you were looking at the growth rate of real consumer spending on durable goods. If you looked at the annual average growth rate over the five years from the end of 1969 to the end of 1974, using the November 1975 data vintage, you would have seen that the annual average growth rate was 1.7 percent. But if you looked at data over the same time span, but using the data vintage from August 1999, you would have seen that the growth rate was much higher—2.8 percent. The size of this revision to a five-year average growth rate might seem extreme, but our research shows that this is fairly typical of the size of revisions to many variables over many periods.

Short-term growth rates can also vary dramatically. For example, if we look at the growth rate of real output for the first quarter of 1977, we see tremendous changes across vintages. The first release of the real output data for that quarter was 5.2 percent (from our May 1977 vintage). It then went on the following journey: 7.5 percent in the vintage of August 1977, 7.3 percent in August 1978, 8.9 percent in August 1979, 9.6 percent in February 1981, 8.9 percent in August 1982, 5.6 percent in February 1986, 6.0 percent in February 1992, 5.3 percent in February 1996, 4.9 percent in May 1997, and 5.0 percent in May 2000, where it stands today (or, at least, at the time this paper was written). The question is: Do such sharp changes in quarterly growth rates have any implications for forecasting?

Although our focus in this paper is empirical, it is of interest to ask whether there are any *a priori* reasons to think that such data revisions should matter. We address this question within the context of a simple AR(2) process, asking under what general conditions data revisions might be important for forecasting. As we'll demonstrate below, many of the insights from this simple model seem to hold in our real-time empirical forecasting exercises.

We suppose that the data generating process for a variable, Y_t, measured without error, is given by the AR(2) specification:

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \quad (1)$$

where μ, ϕ_1, and ϕ_2 are parameters, and ε_t represents a mean zero economic shock with a variance given by σ_ε^2.

Now, consider the forecast made on the basis of a vintage of data that may be subject to revision. Let $Y_{t,v}$ represent the realization for *Y* in period *t* as reported in vintage *v*. Thus, we can represent a revision to a period *t* observation with the expression, $Y_{t,v} - Y_{t,v-1}$. In the national income and product accounts data that we study here, such revisions occur frequently. For example, the first release for a quarter is revised twice over the next two months. Then, each July the observations over the previous three years are revised. Finally, a major revision—called a benchmark revision—occurs every five years. Benchmark revisions affect all observations and may include changes in the definitions of variables and changes in the aggregation method.

In this setup, we will let $Y_{t|t-1,v}$ represent a vintage-specific forecast, that is a forecast for period *t*, made on the basis of information through period *t* – 1 as reported in vintage *v*. In a real-time forecasting situation, *v* would be given by the first vintage to contain the *t* – 1 observation. However, the expression is more general than that because it allows us to represent a forecast made on the basis of any vintage, not just the one that would have been used in real time. The one-step ahead forecasting rule is given by:

$$Y_{t|t-1,v} = \hat{\mu}_v + \hat{\phi}_{1,v} Y_{t-1,v} + \hat{\phi}_{2,v} Y_{t-2,v} . \tag{2}$$

We denote estimated parameters with a circumflex (^), noting that they depend on the data vintage that is used, so they are subscripted with a *v*.

To see how forecasts are affected by data revisions, consider the forecast for period *t* made on the basis of another vintage, *w* > *v*, given by $Y_{t|t-1,w} = \hat{\mu}_w + \hat{\phi}_{1,w} Y_{t-1,w} + \hat{\phi}_{2,w} Y_{t-2,w}$. We can represent the implied forecast revision with the expression:

$$Y_{t|t-1,w} - Y_{t|t-1,v} = (\hat{\mu}_w - \hat{\mu}_v) + (\hat{\phi}_{1,w} Y_{t-1,w} - \hat{\phi}_{1,v} Y_{t-1,v}) + (\hat{\phi}_{2,w} Y_{t-2,w} - \hat{\phi}_{2,v} Y_{t-2,v}) . \tag{3}$$

This expression suggests two influences on forecast revisions: (1) a direct channel because of revisions to initial conditions (by which we mean the data on variables that appear in the forecasting equation, in this case $Y_{t-1,w}, Y_{t-1,v}$, $Y_{t-2,w}$, and $Y_{t-2,v}$); and (2) an indirect channel from changes in estimated coefficients.[2]

To see the direct channel, suppose a revision to the data has no effect on the coefficient estimates, so that $\hat{\mu}_w = \hat{\mu}_v = \hat{\mu}$, $\hat{\phi}_{1,w} = \hat{\phi}_{1,v} = \hat{\phi}_1$, $\hat{\phi}_{2,w} = \hat{\phi}_{2,v} = \hat{\phi}_2$. Then the forecast revision is:

[2] We ignore a third possibility: a change in the specification of the model as we change from vintage *v* to vintage *w*. An example would be a change in the number of lags in the model.

$$Y_{t|t-1,w} - Y_{t|t-1,v} = \hat{\phi}_1 (Y_{t-1,w} - Y_{t-1,v}) + \hat{\phi}_2 (Y_{t-2,w} - Y_{t-2,v}) \,. \tag{4}$$

In this situation, data revisions have a larger effect on forecasts, the larger are the autoregressive coefficients, $\hat{\phi}_1$ and $\hat{\phi}_2$. For example, the forecasts of processes that are nearly white noise ($\phi_1 \approx 0, \phi_2 \approx 0$) are unlikely to be affected very much by data revisions. On the other hand, forecasts of processes that exhibit quite a bit of persistence (as characterized by large and positive autoregressive coefficients) will likely display quite a bit of sensitivity to data revisions.

To see the indirect channel, suppose a revision to the data does not affect the variables used in the forecasting equation but affects the coefficient estimates in a significant way, so that $Y_{t-1,w} - Y_{t-1,v} \approx 0$ and $Y_{t-2,w} - Y_{t-2,v} \approx 0$, yet $\hat{\mu}_w \neq \hat{\mu}_v$, $\hat{\phi}_{1,w} \neq \hat{\phi}_{1,v}$, and $\hat{\phi}_{2,w} \neq \hat{\phi}_{2,v}$. This could happen, for example, if data on Y_{t-1} and Y_{t-2} were not revised, but data on Y_{t-3}, Y_{t-4}, and longer lags were revised, which caused the estimated coefficients to change. Then the forecast revision is:

$$Y_{t|t-1,w} - Y_{t|t-1,v} \approx (\hat{\mu}_w - \hat{\mu}_v) + (\hat{\phi}_{1,w} - \hat{\phi}_{1,v}) Y_{t-1,v} + (\hat{\phi}_{2,w} - \hat{\phi}_{2,v}) Y_{t-2,v} \,. \tag{5}$$

In this situation, the impact of the data revision comes from its effect on the coefficient estimates. The size of the autoregressive parameters does not matter—only the magnitude of the change in the parameter estimates.

Once we combine these two effects, it is hard to derive any concrete conclusions on the effect of data revisions on forecasts. Thus, the issue is largely an empirical one. In the following sections, we lay out the case for the importance of data revisions in the context of autoregressive specifications for output and prices.

IV. REPEATED OBSERVATION FORECASTING

To illustrate how the data vintage matters in forecasting, we begin by developing a novel method for showing how the use of different vintages leads to different forecasts for a particular model. We estimate and forecast the quarterly growth rate of real output with an autoregressive model, ARIMA(p,1,0), on the log level of real output using alternative data vintages. From November 1965 to August 1999, we have 136 vintages of data. We use each vintage (to the extent that the vintage contains the initial values) to forecast each observation, from 1965:4 to 1999:3, thus generating "repeated forecasts" for each date. To generate a forecast for some particular date (t), we can use quarterly data from a particular one of the 136 vintages with a sample beginning in 1947:1 and ending at date $t - 1$ (for all vintages dated t and after).[3] The variable being forecast is the quarterly change in the log of real output (which is GNP prior to 1992 and GDP in 1992 and after; we just call this *RGDP* below). Since model selection is important in these types of exercises, we allow each vintage- and time-specific forecast to be based on a different value of p, using the Akaike Information Criterion (AIC) to select that value. We consider for now just one-step-ahead forecasts, though we could easily extend the horizon. To summarize, we construct the double-indexed sequence $\left\{\Delta \ln(RGDP)^{f}_{t+h|(t-1,v)}\right\}_{t=1965:4,1966:1,\ldots,1999:3;v=t-h+1,t-h+2,\ldots,1999:3}$, where the superscript f indicates that these are forecasts, v is the vintage date, and $t+h$ is the date for which the forecast was made. Note that t = 1965:4, 1966:1, . . . , 1999:3, and vintages used for a particular forecast date t are $v = t - h +$

[3] Thus for a forecast for the fourth quarter of 1965, there are 136 vintages of data from which to forecast—each vintage from November 1965 on. But for later forecasts, there are fewer vintages available; for example, for a forecast for the fourth quarter of 1975 there are only 96. Even though vintages dated $t + 1$ and after include actual data on the object being forecast (at date t), we are using only data through $t - 1$. This is a simulation of forecasting exercises in which forecasts for past dates are made using latest-available data, but in which the researcher just uses data through $t - 1$ in order to simulate a real-time forecasting exercise.

1, $t - h$ +2, . . ., 1999:3, where h is the forecast step (for example, $h = 0$ for a current-quarter forecast and $h = 4$ for a four-quarter-ahead forecast) and $t - h$ is the date of the last observation of the sample data being used to generate the forecast.

This method generates a large number of forecasts. To look at one slice of this problem, examine Figure 1, which shows the forecasts for 1970:1 to 1974:4. You will note that there are 20 vertical sets of dots in the figure; each one of those sets of dots corresponds to a different date for which a forecast is made. Look at the first vertical set of dots, which corresponds to forecasts made for 1970:1. The forecasts shown are forecasts of the quarterly growth rate of real GDP (shown on the vertical axis) in 1970:1, made using an ARIMA model with data from the sample period 1947:1 to 1969:4, running the regression equation, then forecasting one step ahead. Each dot corresponds to a forecast made using data for the same sample period but from a different vintage of data. The vintages range from February 1970 to August 1999, so there are actually 119 dots in that first vertical column (but many are identical, so they do not all show up).

Even though many of the forecasts turn out to be the same, there are also quite large differences between the forecasts. For example, the forecasts for 1970:1 range from 1.13 percent to 2.87 percent. Thus, you can see that there is a tremendous disparity in the forecasts for any date. This is particularly surprising when you realize that all of the forecasts for a particular date are generated using the same sample period. Only the vintage of the data changes.

In our earlier study of data revisions (Croushore and Stark, 2001), we noted that the revisions to the national income and product accounts data were particularly large and variable in the second half of the 1970s. This occurred because measures of real output and its components depend on measures of relative prices (such as the relative price of oil), and relative prices changed dramatically in the middle 1970s and early 1980s. Consequently, measures of real output changed dramatically across vintages as revisions were made over time. As you might imagine, the uncertainty in the measurement of real output generates sharply different forecasts when using data from different vintages. As Figure 2 indicates, our simple ARIMA model for real output yields big differences in forecasts for the second half of the 1970s. For example, in the fourth vertical set of dots in the figure, corresponding to forecasts for 1975:4, the forecasts range from 4.89 percent to 10.68 percent for the quarter! For some dates, the range of the forecasts for a given date is much wider than the range of the actuals across vintages for that date.

We can provide some additional perspective on the sensitivity of the output-growth forecasts to data revisions by comparing forecast variability to the variability in the realizations. Figure 3 does this for the second half of the 1970s. For each quarter over this period, the figure plots the largest and smallest realizations (upper and lower solid lines) and the corresponding largest and smallest forecasts (upper and lower dashed lines), generated in the same manner described above. If you take the vertical distance between the two dashed lines at each date, you will have one measure of the variability of the forecasts. A similar computation for the solid lines will yield a measure of the variability of the realizations. On average, the forecast ranges are a bit narrower than the ranges of the realizations: over the 20 quarters shown in the figure, the average of the ratio of the forecast range to the realization range is 0.76, indicating that the forecast range is, on average, about three-quarters of the realization range. However, there are several quarters (1975:Q4, 1976:Q1, 1978:Q3, and 1978:Q4) in which the forecast range exceeds the range of realizations. For example, in 1978:Q3, the range of realizations is 1.32 percentage points and the range of forecasts is 3.54 percentage points, indicating that the range of forecasts is almost three times as large as the range of realizations. Given the small size of the coefficients in real output growth autoregressions, such a large forecast range seems implausible. However, relatively large ranges of forecasts can occur whenever there is an unusually large range of revisions to the initial conditions that generate the forecasts.

We have repeated this exercise for all the possible different dates, though the results are too voluminous to present fully here. But a similar pattern shows up, even for forecasts for recent years. For example, Figure 4 shows repeated observation forecasts for the second half of the 1980s, and you can see quite a bit of variation in the forecasts.

We can also repeat these exercises for output price inflation. Using an ARIMA(p,1,0) specification on the log level of the output price deflator (chain-weight price index starting with the February 1996 vintage) and AIC lag selection, we generate one-step-ahead repeated observation forecasts in the same manner as described above for real output. *A priori*, since the rate of inflation tends to exhibit more persistence than output growth, we expect the inflation forecasts to be much more sensitive to data revisions than the output growth forecasts.

In Figure 5, we plot the largest and smallest values of inflation realizations (upper and lower solid lines) and the corresponding largest and smallest one-step-ahead quarter-over-quarter inflation forecasts (upper and lower dashed lines) for each quarter over the period from 1975:Q1 to 1979:Q4. Over the 20 quarters shown in the figure, the ratio of the width of the forecast range to the width of the range of realizations averages 0.89, suggesting that inflation forecast variability is almost nine-tenths the size of the variability in the realizations. Thus, over the second half of the 1970s, inflation forecasts were more sensitive to data revisions than output growth forecasts.

One advantage of our real-time data set is that we can extend our results to much longer time periods, thus eliminating the need to guess about whether our results generalize beyond particular sample periods. In Table 1, we report summary statistics on the forecast-variability to realization-variability ratios described above for the period 1965:Q4 to 1999:Q1. The top panel reports the results for our repeated observation one-step-ahead quarter-over-quarter output growth forecasts, and the bottom panel does the same for the inflation forecasts. In each panel, we report the mean ratio, the median ratio, and the 25th and 75th percentiles (denoted "middle 50%"). We do this for the case in which lag selection is based on the AIC, and again, as a robustness check, for the case in which the Schwartz Information Criterion (SIC) selects the lag length.

We draw two important observations from the table. First, inflation forecasts are more sensitive to data revisions than output growth forecasts, confirming our previous results for the second half of the 1970s. Over the period from 1965:Q4 to 1999:Q1, our measure of the variability of inflation forecasts (relative to the variability of inflation realizations), based on the mean and AIC lag selection, is 0.88, which stands at a much higher level than our measure of 0.62 for output growth forecasts. We draw the same conclusion when we examine the median, the 25th and 75th percentiles, and the case in which lag selection is based on the SIC. This result follows both because inflation is more persistent than output growth, suggesting that revisions to initial conditions have larger effects, as indicated in equation (4), and because, for some forecast dates, there are larger changes in the estimated model coefficients for inflation than for output growth.

In this paper, we do not quantify the separate contributions to forecast variability across vintages of: (1) revisions to initial conditions (data on the variables included in the forecasting equation); (2) coefficient revisions; and (3) model revisions (as represented by differences in the number of lags chosen). However, we have examined the effects of these revisions in several ways. First, our examination of model revisions shows that in the early portion of the expanding-window sample periods, model revisions (as measured by the range of lags chosen by the AIC across vintages for a given sample ending date) are a bit lower for inflation than for output growth. However, beginning in 1975, model revisions for inflation exceed those for output growth. Second, we have also examined variation in each equation's estimated constant term and the sum of the coefficients on the autoregressive lags across vintages for a given sample ending date. In this case, to avoid the effect of model revisions, we examine the coefficients from models in which the number of lags is constant across vintages (at the maximum number of lags that minimize the AIC or SIC for a given sample period) but is allowed to change as the sample is expanded. We find periods in which coefficient variability is larger for inflation and periods in which coefficient variability is larger for output growth. It is therefore likely that all three sources (revisions to initial conditions, coefficient revisions, and model revisions) contribute to the variability of forecasts.

Second, the measures of forecast sensitivity are lower when lag selection is governed by the SIC rather than the AIC. For example, for the output growth forecasts, our measure (based on the mean) falls from 0.62 under AIC lag selection to 0.48 under SIC lag selection. Thus, it appears that the SIC, which penalizes additional lags more heavily than the AIC, affords a degree of insulation to the forecasts from data revisions. In part, this occurs because the use of the AIC leads to more variability in the number of lags across vintages for a given forecast date (that is, larger model revisions), whereas the use of SIC leads to a nearly constant choice of lags across vintages.

Further investigation (holding the number of lags constant across vintages for a given forecast date at the maximum lag across vintages that minimizes the AIC or SIC at that date) shows that most of this effect arises because the longer lags chosen by the AIC appear to magnify the sensitivity of forecasts to data revisions. In particular, when we control for model revisions, we find that the mean of the ratio of forecast variability to realization variability is: for real output growth, 0.54 for AIC and 0.48 for SIC; for inflation, 0.85 for AIC and 0.75 for SIC. Thus, even when we eliminate the influence of model revisions, forecast revisions are larger when the lag length is based on the AIC than when it is based on the SIC.

V. COMPARING FORECASTS WITH REAL-TIME DATA SAMPLES TO LATEST-AVAILABLE DATA SAMPLES

To illustrate more extensively how the data vintage matters in forecasting, we run some simple empirical exercises with many permutations, comparing forecasts based on real-time data to those based on latest-available data (from our August 1999 vintage). We forecast real output growth with an ARIMA model and compare the forecasts generated from models estimated on latest-available data to those generated from models estimated on our real-time data.[4] We

[4] See Croushore and Stark (1999a) for similar results with additional forecasting models, including a univariate Bayesian model, and with the multivariate quarterly Bayesian vector error-correction

proceed in the following manner. First, we estimate a model for real output growth using data from the second quarter of 1948 through the third quarter of 1965 that were known in November 1965. Second, we forecast quarter-over-quarter real output growth for the fourth quarter of 1965 and the following three quarters to the third quarter of 1966. Third, we repeat parts (1) and (2) in a rolling procedure, going forward one quarter each step, adding one more observation to the sample used for estimation. Fourth, we calculate the forecast errors for the forecasts. We carry out this procedure to examine forecasts one quarter ahead, two quarters ahead, and four quarters ahead (we skip three just to conserve space—the results are similar), as well as the average of the four quarterly forecasts.

A key question in forecasting exercises is this: What vintage of the data is to be used to represent "actuals" from which forecast errors are computed? Since the literature on forecasting has no definitive answer to this question, we study it in more detail by using three alternatives: latest-available data (August 1999 vintage), last benchmark data (the last vintage before a benchmark release, which occurs about every five years—November 1975, November 1980, November 1985, November 1991, November 1995, and August 1999), and the vintage of the data four quarters after the fourth-step-ahead forecast associated with each one-to-four-step-ahead forecast string.[5] (The ease with which we can do this is a major virtue of the real-time data set.) We follow this procedure once using as a sample the data from the real-time data set (for

---

(QBVEC) model of Stark (1998). Similar results to those discussed in this paper obtain, though Bayesian methods seem to reduce the impact of data revisions on forecasts.

[5] Note that sometimes in the charts that follow, the full number of observations is not available in each subsample. We lose observations on "actuals" when there are missing data and in the case where we use the "last benchmark" as "actual" when the forecast horizon extends beyond the release date of new benchmark data.

which data revisions are possible as we roll forward each quarter using the next vintage associated with that quarter), and a second time using only the August 1999 vintage.

In setting up this experiment, we thought that the results of this exercise would be obvious. In particular, our prior was that using the latest-available data to forecast and using the latest-available data as "actuals" would lead to lower forecast errors compared with using real-time data to forecast the latest-available actuals. The forecasts based on latest-available data have two advantages in that comparison. First, they contain revisions that reduce the measurement error in the data. Second, the latest-available data being used to forecast contain the definitional changes and benchmark changes that were not available in real time. Given that, our findings, discussed below, are surprising.

Using an ARIMA(4,1,0) model on the log level of real output, we find some interesting comparisons between the forecasts made using real-time data and those using latest-available data (Table 2). In this table, we examine forecast error statistics for the full sample of available data vintages with an expanding window (that is, the sample size increases by one as we move from one vintage to the next). In the table, there are three sets of numbers, corresponding to each alternative choice of "actuals." We compare mean error (ME), mean absolute error (MAE), and root-mean-squared error (RMSE) for both real-time (RT) and latest-available (L) samples, with the number of forecasts shown in the column headed "N." We test the statistical significance of the differences between the RMSEs for real-time and latest-available samples using the Harvey, Leybourne, and Newbold (1997) modification of the test proposed by Diebold and Mariano (1995). (We do not do so for tests of four-quarter averages or when we use the last benchmark as actuals.)

What is perhaps surprising in Table 2 is the lack of difference between forecast error statistics generated using forecasts formed with real-time data compared to those formed with latest-available data. The mean errors are sometimes a bit different, but MAEs and RMSEs are nearly identical (the differences in RMSEs are very small and not statistically significant). That is a surprising result, given our priors. But part of this outcome arises because in subsamples the results can be quite different between real-time and latest-available data, with each doing better in some subsamples. For example, when we look at shorter periods, such as 1970:1 to 1974:4 in Table 3, the differences can be much larger. For example, the difference in RMSEs for one-step-ahead forecasts is fairly large for all three choices of actuals; the difference is statistically significant for the two actuals (latest available and four quarters later) that were tested.

We can justify looking at short subsamples in these tables by thinking about how forecast evaluation is performed in practice. Ideally, of course, we would test forecasting models over long sample periods and judge the models on the basis of their out-of-sample forecast errors. However, in practice, sample periods are short or models have not been used long enough to do anything but make comparisons on short samples. Classic examples in the literature include Litterman (1986), who evaluated Bayesian VARs with five years of forecasts, and McNees (1992), who looked at five-and-one-half years of forecasts. And, of course, a difficulty always faced by those developing forecasting models is the tradeoff between sample size over which to estimate a model and having enough data left to evaluate the model. That tension generally leads model developers to choose a fairly long sample period and a short evaluation period. So models tend to be chosen based on short out-of-sample (or simulated out-of-sample) forecast evaluation periods.

Consequently, Tables 4 and 5 rerun the same exercises for other subsample periods.[6] Notice that there are no consistent results across the sample periods. For example, in Table 4, the results from the second half of the 1970s suggest that the real-time-data-based forecasts (RT) are a bit worse than the latest-available-data-based forecasts (L), whereas in Table 3, the results for the first half of the 1970s showed the opposite result. As you might imagine, the differences between real-time-data-based forecasts and latest-available-data-based forecasts are not very big for recent subsample periods, as in the late 1990s, reported in Table 5, since the data have not been revised much. Of the forecasts that can be tested in Tables 4 and 5, in no cases are the differences between real-time and latest-available forecasts statistically significant.

We have performed the identical experiment for the output price deflator (PGDP). The results over the full sample are shown in Table 6. For prices, the differences between real-time-data-based forecasts and latest-available-data-based forecasts in the full sample are greater than was the case for output, and are significantly different. In some subsamples, the differences between real-time-data-based forecast errors and latest-available-data-based forecast errors are quite large. Perhaps not surprisingly, that is true especially in the second half of the 1970s (Table 7), which also shows many significant differences between real-time and latest-available forecasts.

For the price results, we find confirming evidence for our prior that the root-mean-squared error and mean absolute error would be smaller when using latest-available data compared to using real-time data. That result holds for every subsample. (These results are not reported here, but are available in the Appendix.) We think this result is particularly important

[6] Additional tables, showing the full set of all subsample periods, can be found in an Appendix, available from the authors upon request.

because when researchers develop forecasting models, they usually do their development work on latest-available data. But the resulting forecasts they generate may have significantly lower root-mean-squared errors or mean absolute errors than if the forecasts were generated using real-time data. Thus a claim that a new forecasting model (based on final revised data) is superior to others because it has a lower root-mean-squared error than some forecasts that were generated in real time is simply not tenable.

We can illustrate even more clearly the potential pitfalls of ignoring the real-time data issue by considering the story of Fred Forecaster. Fred, an analyst with a local financial services company, has been given the task of generating forecasts for output and inflation for use in his company's asset allocation decisions. Knowing little about macroeconomic forecasts, Fred does a quick literature review and discovers that simple parsimonious specifications often work best. Fred also discovers a paper reporting benchmark root-mean-square-errors from a model whose specification is not entirely clear. In fact, all that Fred knows is that the author has been tracking model performance for an unusually long time— from 1975:Q1 to 1979:Q4, to be precise—and has reported an RMSE of 4.85 percent for his one-step-ahead output growth forecasts and 1.90 percent for his inflation forecasts. Armed with these benchmark forecast error statistics, Fred collects a time series of latest-available (as of August 1999) observations on real output growth and inflation, and uses an expanding window of observations to estimate an AR(4) model for each. His one-step-ahead out-of-sample RMSEs, computed over the same sample period as that reported for the benchmark forecasts, are 4.36 percent for output growth and 1.53 percent for inflation, quite a bit lower than those reported in the literature. Thus, Fred reckons that his specifications are superior to those used in the literature. However, imagine Fred's embarrassment when he later discovered that the benchmark forecasts were generated from the same specification—an AR(4)—that he used on his latest-available data!

According to our research results, Fred's experience is not uncommon. In Table 8, we extend the experiment outlined above to include several additional sample periods, and we also tabulate results for the full period. Using our real-time data set, we study two types of forecast evaluation exercises for real output growth and inflation. First, we use the latest- available vintage (August 1999) to estimate an AR(4) model, forecast one step ahead, and evaluate forecast errors. We denote this case "Latest $\rightarrow$ Latest," to indicate that latest-available data are used to forecast latest-available observations. Second, we repeat the exercise using real-time data to estimate and forecast. However, because each real-time vintage does not contain the realization required to compute the out-of-sample forecast error, we use an appealing real-time alternative—our "last benchmark" realization, as described above. We denote this case "Real Time $\rightarrow$ Last Bench," to indicate that real-time data are used to forecast the observation contained in the last vintage prior to a benchmark revision. (Note that this experiment differs slightly from the one we used at the beginning of this section. In this experiment, we use latest-available and real-time data to generate two sets of forecasts. However, here, we evaluate each using a different realization, latest-available and last-benchmark, respectively.)

Looking at the results in Table 8 for the full sample (1965:Q4 to 1999:Q2, excluding the observations for 1975:Q4, 1980:Q4, 1985:Q4, 1991:Q4, and 1995:Q4, for which there are no realizations under the last-benchmark measure), we see little difference in the RMSE for the output growth forecasts (3.63 percent for Latest $\rightarrow$ Latest vs. 3.61 percent for Real Time $\rightarrow$ Last Bench), but somewhat larger differences for the inflation forecasts. Notably, the inflation forecasts are deceptively lower when the latest-available data are used (1.35 percent) than when the real-time data are used (1.73 percent). For the sub-periods, the results for the output forecasts are hard to characterize: over some periods, real-time data yield higher RMSEs than the latest-available data; in other periods, the reverse holds. In contrast, the sub-period results for the inflation forecasts are easy to characterize: in all periods, latest-available data yield lower RMSEs than the real-time data.

We believe the results of this section represent a shot across the bow of the ship of conventional model selection and the methodology of forecast evaluation. Using our real-time data set, we find that conventional forecast-error statistics can be sensitive to the choice between latest-available and real-time data. The choice of data vintage may also affect the model-selection decision. Importantly, the results are sensitive to the forecast evaluation period, the span of observations in the period, and the variable being modeled. Clearly, additional research on our real-time data set is required to pin down more specific guidelines on when real-time data issues are most important. In our view, though, there is enough evidence to date to suggest that the only safe way to evaluate forecasting models is with real-time data.

VI. INFORMATION CRITERIA AND FORECASTS

When economists create forecasting models, they often allow criteria like the Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC) to guide some aspects of model specification. For example, Swanson and White (1997) use information criteria to select the number of variables and the number of lags on those variables for use in designing good forecasting models. But, in recent years, the use of lag selection criteria has expanded beyond that in forecasting models. Granger, King, and White (1995) identify several shortcomings in classical hypothesis testing methodology, arguing that structural hypothesis testing could benefit from model selection procedures, such as the AIC and SIC. Notably, we find that the choices obtained from those criteria can be sensitive to the choice of data vintage, complicating the implementation of the Granger-King-White methodology.

Consider the following experiment. Suppose we run our repeated-observation-forecasting experiment on an ARIMA(p,1,0) model for log of real output as we described above. For each sample ending date and for each vintage of data, we use the AIC or SIC to choose the lag length, p, for the model, allowing p to vary from 1 to 8. As we change the sample ending date, of course, the choice of p is likely to change. But the choice also depends on which vintage of data is used. The results of this experiment are shown in Figure 6. Panels in the left column use the AIC; those in the right column use the SIC. Panel 1.A. plots the maximum and minimum lag lengths across vintages that minimize the AIC at each sample ending date from 1965:3 to 1999:2; Panel 2.A. does the same for the SIC. The range between the maximum and minimum number of lags chosen for each sample ending date is shown in Panel 1.B.—it is just the maximum number of lags at each date minus the minimum number of lags shown in Panel 1.A. Similarly, Panel 2.B. does the same for the SIC. Finally, the bottom panels show the number of times that there is a change across vintages in the lag length that minimizes the AIC (Panel 1.C.) or SIC (Panel 2.C.) for each sample ending date. This gives an idea of how sensitive the lag length is to changes in vintage. So, the choice of lags to use in a simple univariate model may be quite different depending on the vintage of data used for the AIC. However, we also note that, at least in this experiment, the SIC does not face this problem, except for one sample ending date.

Thus, it appears that the SIC is less sensitive to data revisions than the AIC. We reach a similar conclusion when we run the same experiment for inflation (Figure 7). In this case, the number of lags shows some variability even when using the SIC, but using the AIC shows much greater variability.

VII. CONCLUSIONS

This paper describes some experiments in forecasting, comparing the use of real-time data to latest-available data. Our strongest and clearest results are that the choice between using real-time data or latest-available data matters in important ways. Forecasts made for a particular date can be quite different, depending on the vintage of data used, as we show in our repeated observation forecasts. The root-mean-squared errors and mean absolute errors in forecasts can differ between using real-time data and latest-available data when only short spans of observations are used. We also find that inflation forecasts seem more sensitive to the choice between real-time and latest-available data than forecasts for real output.

These results caution researchers against developing forecasting models (for some purposes) with latest-available data and only a small number of out-of-sample observations for forecast comparisons, whose MAEs and RMSEs may be misleadingly low. We also find that the choice of lag length can be affected by the use of latest-available compared to real-time data.

We hope these exercises will convince those developing forecasts to focus their efforts on real-time data rather than latest-available data. Clearly, additional research is required to pin down the conditions under which real-time data issues are most important. It is our intention to keep adding additional variables to the data set over time to allow a wider range of real-time forecasting exercises to be performed.

Table 1. Statistics on the Sensitivity of ARMA Forecasts to Data Revisions
(forecast range ÷ realization range)

Sample: 1965:Q4 to 1999:Q1

1.A. One-Step-Ahead Q/Q Output Growth

	Lag Selection Method	
	AIC	SIC
Mean	0.62	0.48
Median	0.50	0.40
Middle 50%	0.33 – 0.81	0.29 – 0.60

1.B. One-Step-Ahead Q/Q Output-Price Inflation

	Lag Selection Method	
	AIC	SIC
Mean	0.88	0.75
Median	0.67	0.60
Middle 50%	0.42 – 1.10	0.41 – 0.87

Table 2. Forecast Error Statistics, Full Sample
Model: ΔLog(RGDP) ~ ARIMA(4,0,0)

2.A. Actual Value: Latest Available

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	135	-0.23	-0.47	2.60	2.63	3.59	3.62	0
2	134	-0.48	-0.65	2.60	2.67	3.72	3.73	0
4	132	-0.80	-0.83	2.70	2.68	3.76	3.74	0
4Q Avg	132	-0.57	-0.69	1.77	1.76	2.38	2.40	X

2.B. Actual Value: Last Benchmark

Forecast Step	N	ME		MAE		RMSE	
		RT	L	RT	L	RT	L
1	130	-0.32	-0.58	2.51	2.51	3.61	3.58
2	124	-0.70	-0.90	2.63	2.66	3.93	3.83
4	112	-0.94	-0.99	2.78	2.78	4.12	4.06
4Q Avg	112	-0.70	-0.82	2.00	2.03	2.80	2.85

2.C. Actual Value: Four Quarters Later

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	129	-0.47	-0.72	2.54	2.49	3.64	3.52	0
2	129	-0.75	-0.93	2.61	2.62	3.84	3.73	0
4	129	-1.07	-1.09	2.63	2.66	3.85	3.81	0
4Q Avg	129	-0.82	-0.95	1.95	1.97	2.67	2.70	X

Note: Significance (Sig.) in the far right column shows if there is a statistically significant difference between the real-time and latest-available RMSEs. An asterisk (*) shows significance at the 10% level, a 0 shows no significant difference, and an X shows that the test was not performed.

Table 3. Forecast Error Statistics, 1970Q1 – 1974Q4
Model: ΔLog(RGDP) ~ ARIMA(4,0,0)

3.A. Actual Value: Latest Available

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	20	-0.79	-0.95	3.83	4.17	4.39	4.86	*
2	20	-1.11	-1.26	3.90	4.04	4.58	4.82	0
4	20	-1.43	-1.47	4.27	4.17	4.89	4.94	0
4Q Avg	20	-1.10	-1.17	2.57	2.64	2.91	2.97	X

3.B. Actual Value: Last Benchmark

Forecast Step	N	ME		MAE		RMSE	
		RT	L	RT	L	RT	L
1	20	-1.27	-1.43	3.65	4.19	4.78	5.11
2	20	-1.59	-1.75	3.93	4.13	5.12	5.28
4	20	-1.92	-1.95	4.31	4.34	5.51	5.51
4Q Avg	20	-1.37	-1.43	3.16	3.23	3.62	3.71

3.C. Actual Value: Four Quarters Later

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	20	-1.53	-1.69	3.63	4.21	4.60	4.95	*
2	20	-1.76	-1.92	3.83	4.05	4.94	5.11	0
4	20	-2.10	-2.13	4.30	4.31	5.51	5.49	0
4Q Avg	20	-1.53	-1.57	3.22	3.29	3.68	3.76	X

Note: For details on significance test (Sig.) see Table 2.

Table 4. Forecast Error Statistics, 1975Q1 – 1979Q4
Model: ΔLog(RGDP) ~ ARIMA(4,0,0)

4.A. Actual Value: Latest Available

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	20	0.51	0.09	3.43	3.10	4.70	4.28	0
2	20	0.18	0.10	3.26	3.15	4.64	4.39	0
4	20	-0.44	-0.07	3.34	3.01	4.56	4.31	0
4Q Avg	20	-0.29	-0.32	2.50	2.25	3.03	2.84	X

4.B. Actual Value: Last Benchmark

Forecast Step	N	ME		MAE		RMSE	
		RT	L	RT	L	RT	L
1	19	0.49	-0.16	3.44	3.13	4.85	4.57
2	18	-0.05	-0.40	3.54	3.37	5.57	4.96
4	16	-0.69	-0.59	3.69	3.48	5.63	5.36
4Q Avg	16	-1.13	-1.26	2.97	2.87	4.16	4.15

4.C. Actual Value: Four Quarters Later

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	20	0.30	-0.12	3.65	3.02	4.98	4.23	0
2	20	-0.05	-0.14	3.48	3.13	5.07	4.41	0
4	20	-0.79	-0.41	3.26	3.00	4.68	4.41	0
4Q Avg	20	-0.75	-0.78	2.66	2.41	3.59	3.45	X

Note: For details on significance test (Sig.) see Table 2.

Table 5. Forecast Error Statistics, 1995Q1 – 1999Q2
Model: ΔLog(RGDP) ~ ARIMA(4,0,0)

5.A. Actual Value: Latest Available

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	18	0.14	0.10	1.27	1.25	1.57	1.50	0
2	18	0.22	0.18	1.17	1.13	1.50	1.44	0
4	18	0.30	0.21	1.18	1.17	1.45	1.43	0
4Q Avg	18	0.24	0.17	0.67	0.66	0.76	0.77	X

5.B. Actual Value: Last Benchmark

Forecast Step	N	ME		MAE		RMSE	
		RT	L	RT	L	RT	L
1	17	0.33	0.28	1.27	1.25	1.54	1.49
2	16	0.40	0.38	1.22	1.20	1.49	1.44
4	14	0.46	0.43	1.13	1.13	1.33	1.31
4Q Avg	14	0.57	0.51	0.57	0.54	0.64	0.63

5.C. Actual Value: Four Quarters Later

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	12	-0.31	-0.34	1.46	1.43	1.77	1.67	0
2	13	-0.22	-0.27	1.46	1.51	1.71	1.72	0
4	15	-0.26	-0.38	1.39	1.42	1.62	1.68	0
4Q Avg	15	-0.18	-0.27	0.65	0.73	0.83	0.89	X

Note: For details on significance test (Sig.) see Table 2.

Table 6. Forecast Error Statistics, Full Sample
Model: ΔLog(PGDP) ~ ARIMA(4,0,0)

6.A. Actual Value: Latest Available

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	135	0.65	0.43	1.15	1.02	1.54	1.35	*
2	134	0.89	0.67	1.49	1.33	1.98	1.83	*
4	132	1.30	1.13	2.05	1.91	2.73	2.55	*
4Q Avg	132	0.99	0.79	1.51	1.34	1.98	1.79	X

6.B. Actual Value: Last Benchmark

Forecast Step	N	ME		MAE		RMSE	
		RT	L	RT	L	RT	L
1	130	0.50	0.30	1.32	1.23	1.73	1.61
2	124	0.77	0.56	1.66	1.56	2.17	2.06
4	112	1.27	1.14	2.16	2.07	2.92	2.80
4Q Avg	112	0.97	0.79	1.58	1.42	2.12	1.96

6.C. Actual Value: Four Quarters Later

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	129	0.59	0.36	1.27	1.14	1.68	1.54	*
2	129	0.84	0.61	1.60	1.46	2.14	2.00	*
4	129	1.19	1.02	2.07	1.97	2.80	2.66	*
4Q Avg	129	0.91	0.69	1.46	1.31	2.02	1.84	X

Note: For details on significance test (Sig.) see Table 2.

Table 7. Forecast Error Statistics, 1975Q1–1979Q4
Model: ΔLog(PGDP) ~ ARIMA(4,0,0)

7.A. Actual Value: Latest Available

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	20	1.35	0.73	1.61	1.33	2.06	1.57	*
2	20	1.97	1.30	2.26	1.75	2.55	2.09	*
4	20	3.41	2.76	3.41	2.78	3.55	3.01	*
4Q Avg	20	2.63	2.02	2.64	2.13	2.90	2.52	X

7.B. Actual Value: Last Benchmark

Forecast Step	N	ME		MAE		RMSE	
		RT	L	RT	L	RT	L
1	19	0.84	0.33	1.48	1.34	1.90	1.76
2	18	1.48	0.92	2.06	1.85	2.46	2.24
4	16	3.38	2.84	3.38	2.90	3.63	3.25
4Q Avg	16	2.67	2.21	2.67	2.25	2.94	2.67

7.C. Actual Value: Four Quarters Later

Forecast Step	N	ME		MAE		RMSE		
		RT	L	RT	L	RT	L	Sig.
1	20	1.03	0.42	1.55	1.25	2.00	1.68	0
2	20	1.71	1.03	2.12	1.69	2.62	2.23	*
4	20	3.04	2.39	3.09	2.52	3.38	2.92	*
4Q Avg	20	2.34	1.74	2.40	2.02	2.72	2.40	X

Note: For details on significance test (Sig.) see Table 2.

Table 8.
ΔLog(Real Output)~ARIMA(4,0,0): 1-Step-Ahead RMSEs

	70-74	75-79	80-84	85-89	90-94	Full
Latest → Latest	4.86	4.36	5.18	1.61	2.41	3.63
Real Time → Last Bench	4.78	4.85	4.99	1.94	1.99	3.61

ΔLog(Output Price)~ARIMA(4,0,0): 1-Step-Ahead RMSEs

	70-74	75-79	80-84	85-89	90-94	Full
Latest → Latest	2.30	1.53	0.91	0.70	0.85	1.35
Real Time → Last Bench	2.88	1.90	1.55	1.30	1.22	1.73

REFERENCES

Cole, Rosanne, "Data Errors and Forecasting Accuracy," in Jacob Mincer, ed., *Economic Forecasts and Expectations: Analyses of Forecasting Behavior and Performance*. New York: National Bureau of Economic Research, 1969, pp. 47-82.

Croushore, Dean, and Tom Stark. "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics* 105 (2001), pp. 111–30.

Croushore, Dean, and Tom Stark. "Does Data Vintage Matter for Forecasting?" Federal Reserve Bank of Philadelphia Working Paper 99-15, October 1999a.

Croushore, Dean, and Tom Stark. "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?" Federal Reserve Bank of Philadelphia Working Paper 99-21, December 1999b.

Denton, Frank T., and John Kuiper. "The Effect of Measurement Errors on Parameter Estimates and Forecasts: A Case Study Based on the Canadian Preliminary National Accounts," *Review of Economics and Statistics* 47 (May 1965), pp. 198-206.

Diebold, Francis X., and Roberto S. Mariano. "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13 (July 1995), pp. 253-63.

Diebold, Francis X., and Glenn D. Rudebusch. "Forecasting Output With the Composite Leading Index: A Real-Time Analysis," *Journal of the American Statistical Association* 86 (September 1991), pp. 603-10.

Granger, Clive W.J., Maxwell L. King, and Halbert White. "Comments On Testing Economic Theories And the Use Of Model Selection Criteria," *Journal of Econometrics* 67 (May 1995), pp. 173-187.

Harvey, David, Stephen Leybourne, and Paul Newbold. "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting* 13 (June 1997), pp. 281–91.

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

Howrey, E. Philip. "Forecasting GNP With Noisy Data: A Case Study," *Journal of Economic and Social Measurement* 22 (1996), pp. 181-200.

Koenig, Evan F., and Sheila Dolmas. "Real-Time GDP Growth Forecasts," Federal Reserve Bank of Dallas Working Paper 97-10, December 1997.

Koenig, Evan F., Sheila Dolmas, and Jeremy Piger. "The Use and Abuse of 'Real-Time' Data in Economic Forecasting," Federal Reserve Bank of Kansas City, manuscript, September 1999.

Litterman, Robert B. "Forecasting with Bayesian Vector Autoregressions—Five Years of Experience," *Journal of Business and Economic Statistics* 4 (January 1986), pp. 25-38.

McNees, Stephen K. "How Large Are Economic Forecast Errors?" *New England Economic Review* (July/August 1992), pp. 25-33.

Robertson, John C., and Ellis W. Tallman. "Data Vintages and Measuring Forecast Model Performance," Federal Reserve Bank of Atlanta *Economic Review* (Fourth Quarter 1998a), pp. 4-20.

Robertson, John C., and Ellis W. Tallman. "Real-Time Forecasting with a VAR Model," Manuscript, Federal Reserve Bank of Atlanta, November 1998b.

Stark, Tom. "A Bayesian Vector Error Corrections Model of the U.S. Economy," Federal Reserve Bank of Philadelphia Working Paper 98-12, June 1998.

Stekler, H.O. "Data Revisions and Economic Forecasting," *Journal of the American Statistical Association* 62 (June 1967), pp. 470-83.

Swanson, Norman. "Forecasting Using First-Available Versus Fully Revised Economic Time-Series Data," *Studies in Nonlinear Dynamics and Econometrics* 1 (April 1996), pp. 47-64.

Swanson, Norman R. and Halbert White. "A Model Selection Approach To Real-Time Macroeconomic Forecasting Using Linear Models And Artificial Neural Networks," *Review of Economics and Statistics* (November 1997), pp. 540-550.

Trivellato, Ugo, and Enrice Rettore. "Preliminary Data Errors and Their Impact on the Forecast Error of Simultaneous-Equations Models," *Journal of Business and Economic Statistics* 4 (October 1986), pp. 445-53.