Data Vintages and Measuring Forecast Model Performance

JOHN C. ROBERTSON AND ELLIS W. TALLMAN

Robertson is a visiting scholar and Tallman is a senior economist in the macropolicy section of the Atlanta Fed's research department. They thank Lucy Ackert and Mary Rosenbaum for comments, Robert Parker and Bruce Grimm of the BEA for the NIPA time series data and comments and insights into the bureau's data revision process, Glenn Rudebusch and James Hamilton for the CLI time series data used in their studies, and Amanda Nichols for research assistance.

The data on most economic variables are estimates. These estimates are revised, sometimes frequently, and often they continue to be revised many years after the first estimate appears. For example, on July 31, 1998, the Bureau of Economic Analysis (BEA) of the U.S. Department of Commerce announced that the seasonally adjusted estimate of real gross domestic product (GDP) growth for the second quarter of 1998 was an annualized 1.4 percent. In addition, the July press release contained revised estimates for the real GDP series (and components) from the first quarter of 1995 until the first quarter of 1998. The revision showed an increase in the estimated average year-over-year real GDP growth from 2.9 percent to 3.3 percent for the period from 1995 to 1997.

Chart 1 presents the year-over-year GDP growth estimates for the period from 1995 to 1998 as of June 1998 (referred to as the June 1998 data vintage) together with the corresponding estimates as reported in July 1998 (the July 1998 data vintage).[1] As is apparent from the chart, the growth estimates from the older vintage are systematically lower than those from the more recent vintage. In mid-1996, for example, the year-over-year growth rate was nearly 1 percentage point lower in the June vintage of data than in the revised July vintage.

The timing of revisions to data usually follows a regular schedule, even if the size or the direction of the revisions do not. For example, the BEA usually publishes revisions of the National Income and Product Accounts (NIPA) for the three prior years each July. Moreover, by the time this article appears in print, the 1.4 percent growth estimate for the second quarter of 1998 will have been revised twice—in August and again in September.

Other estimates of economic activity also change over time as new vintages are constructed. For example, the historical data on the seasonally adjusted index of total industrial production (IP) were revised in January 1997 and then again in December 1997. Chart 2 plots the year-over-year percentage change in each of these two data vintages over the period from 1994 until the end of 1996. The difference between the vintages appears sizable; for example, growth during mid-1995 was more than 2 percent lower using the January 1997 version rather than the December 1997 revision.

In a policy context the distinction between vintages of data can be important. For example, Orphanides (1997) shows that a rule-based monetary policy performs dramatically worse when real-time data are used instead of subsequently revised versions of the data. The result—that revised data help make better policy—is an interesting finding; the more relevant issue, though, is that good rule-based policy performance requires the use of data unavailable to the policymaker in real time.

This article finds that the choice of data vintage can be important when comparing the performance of competing forecasting models of real output. Specifically, the research considers a choice between competing forecast models that is based on relative out-of-sample forecast performance. The study requires (1) using data available at the time the forecast would have been made to construct the forecast and (2) using data available not too long after the period being forecast to evaluate the model's performance. For the IP measure of output this approach leads to a quite different conclusion about relative model performance from that derived by using the latest available or most recent vintage of data throughout the analysis.

This result emphasizes the important distinction between an actual real-time forecast analysis and a pseudo real-time forecast analysis. In a pseudo real-time analysis a forecaster uses only the latest available vintage of historical data series in constructing and evaluating the forecasts. In contrast, an actual real-time analysis requires using the vintage of data actually available at the forecast date, together with forecast errors constructed using a vintage of data available soon after the period being forecast. To the extent that future data revisions will be similar to past ones, the results from simulating the past real-time performance of competing models should provide a better guide to a model's subsequent performance than would the results of simulations using only the current vintage of data.

The following section of the article discusses in more detail the distinction between simulating actual real-time forecasts and pseudo real-time forecasts. The position argued is that most results reported in the academic forecasting literature are from pseudo real-time forecast experiments. Of the few studies that have attempted to introduce a real-time aspect into the analysis, most have tended to use the notion either too loosely or too tightly to reflect accurately what a forecaster would have been able to do in real time. The discussion then presents the empirical results of the model comparison exercise using real-time data and contrasts these with the results of using only the most recent data vintage.

Using only the latest vintage of historical data may influence the measured forecast performance in misleading ways.

Real-Time Forecasting

The standard forecast estimation and evaluation strategy is to estimate or fit a model over some period, construct an out-of-sample forecast, and compare this forecast with the actual outcome. Then the forecaster makes a decision, based on the relative size of the resulting forecast errors, about the quality of the model's previous forecast performance. The forecaster hopes that a model that has performed well relative to previous alternatives will continue to do so in the future.

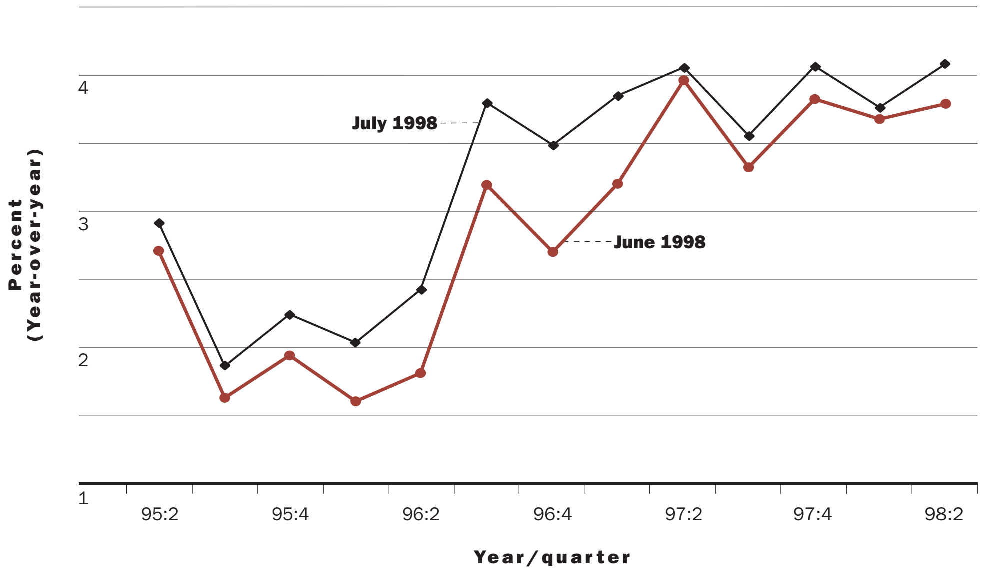
As described in most econometric textbooks as well as in the academic literature, forecast evaluations of a model typically employ the most recent vintage of the relevant time series at each stage of the process. It is possible, however, that using only the latest vintage of historical data may influence the measured forecast performance in misleading ways, and the result may not be a good approximation of forecasting accuracy in real time.

Two potential problems arise when forecast evaluations employ the latest vintage of historical data for both estimating and evaluating. First, in a realistic forecasting situation, one can use only the vintage of historical data available at the time the forecast is made. That even more refined measurements will become available is of little relevance.[2] Thus, forecasts with revised data are not realistic, real-time forecasts.

[1] *A data series vintage or "age" is denoted by the month in which the entire data series existed—when that specific set of numbers was available as data.*

[2] *From today's perspective, it could be argued that the latest available vintage provides the most accurate historical record of series such as gross domestic product or industrial production. But an even more accurate record will likely be available in the future after further revisions have taken place. Consequently, the notion of an "ultimately revised" or "true" history for estimates is somewhat nebulous.*

CHART 1 GDP Growth as of June and July 1998



Source: Data from Bureau of Economic Analysis, Department of Commerce

The second problem that arises from using the latest vintages of data centers on forecast evaluation. A forecaster typically wants to evaluate the model's forecast performance against an outcome that is measured not too long after the month or quarter being forecast. It is unlikely that forecasters or their clients would be prepared to wait for a more revised historical record.

In an important empirical study Fair and Shiller (1990) describe in detail the necessary conditions that a historical analysis of real-time out-of-sample or ex ante forecasts must satisfy. To be specific, suppose that the goal is to evaluate the accuracy of a particular forecasting model of GDP; the forecast is made using data available in some period, and forecast values are generated for subsequent periods. For these out-of-sample forecasts to be constructed in real time,

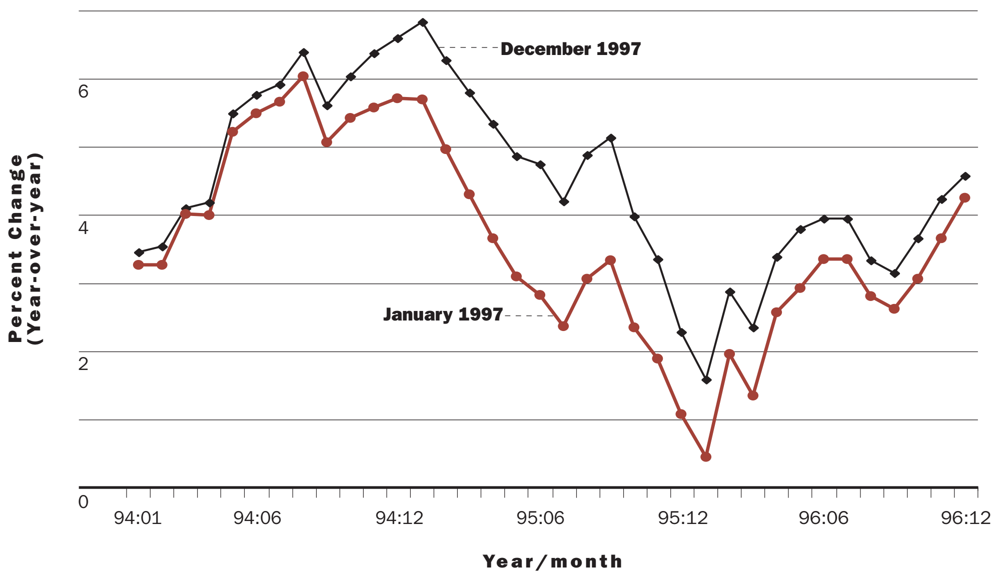
(1) *Future values of variables in the model must be only forecasts.* These forecasts are used in constructing the forecasts of the particular variables of interest. For example, suppose one is interested in forecasting real GDP growth and the federal funds rate (FFR) is believed to affect future real output growth. On any particular date on which a forecast is made, any future values of the FFR used in forming real GDP growth forecasts will themselves be forecasts. To allow actual values of the FFR into the forecasting model is to give the model an unfair advantage. Of course, if the future path for the FFR were known in advance, then it would make sense to use these values when constructing the forecast.

(2) *The coefficients of the model must be estimated only over the sample period up to the time the forecasts are being formed.* For example, suppose there are data from 1959 up until 1998. Estimating the coefficients of the model using all the data through 1998 and then forecasting from 1988 on would be giving the forecast model information from future data observations contained in coefficient estimates that were not actually available in 1988.

(3) *Only data for the period prior to the time the forecast is made can be used in determining the model specification.* Following from the previous example, suppose that the model specification (say, the number of lagged observations to use in the model) is chosen by a criterion that used all the data in the sample through 1998 and the chosen specification is then fitted and forecast from 1988 on. Again, the chosen model's forecasts would have been partially based on information from future data observations. Instead, the model specification should be chosen only on the basis of analysis of observations available through 1988.

(4) *The vintage of the data used to estimate the model and construct the forecasts must be actually available at the time the forecast is made.* This restriction is the focus of this article. Here, the forecasting model is limited to using only the data vintage available at the time the forecast would have been constructed, preventing future information in the form of data revisions from entering into the forecasts. Thus, for example, a forecast formed in

CHART 2 Industrial Production Index as of January and December 1997



Source: Board of Governors of the Federal Reserve System

July 1988 uses only the vintage of data actually available in July 1988.

It is unlikely that any published out-of-sample forecast model evaluation would have failed to satisfy the first two of Fair and Shiller's requirements. Yet it is surprising that a number of studies have ignored the third requirement. In these cases, researchers used the full sample, including the period to be forecast, when determining the model specification (that is, the variables included, the lag length, and so on). Notably, failure to satisfy the fourth requirement is almost universal in the literature. In some studies, like those of Staiger, Stock, and Watson (1997) and Stock and Watson (1998), the researchers are aware that they are only simulating pseudo real-time forecasts when they use the most recent data vintage.[3] In fact, even though Fair and Shiller explicitly address the first three issues in their paper, they admit that they use only the latest revised data in their pseudo real-time forecast evaluation.[4] Presumably the argument is that using real-time vintages of data simply does not matter for the results, but almost no work has been conducted to investigate whether there is evidence to support this proposition.

Moreover, because Fair and Shiller used only the latest vintage of data in their analysis, they did not have to deal with the equally important conceptual issue of which data vintage to evaluate the forecast against. Quantifying forecast accuracy requires a benchmark series against which to compare forecasts. The most recent vintage of data is often suggested as that benchmark because these data give a somewhat cleaner and more accurate measurement. But the frequent redefinition and rebenchmarking of the data series may alter the series properties in ways that a forecaster cannot be expected to predict. Moreover, while using the latest available estimates places the forecasts against measures with the least measurement error, forecasters are most likely to be held accountable for their ability to forecast, say, real GDP growth, using an estimate that is available not too long after the quarter being forecast. This article proposes that a decision about the data vintage that the forecasts are to be evaluated against should be considered prior to beginning a forecasting exercise.

Recent Research on Real-Time Forecast Evaluation

Recent research on the accuracy of forecasting models has moved closer to satisfying the necessary conditions of a real-time exercise as laid out in Fair and Shiller. The key criterion seems simple: if the forecasts cannot be reproduced using the available data

[3] *See Staiger, Stock, and Watson (1997, note 8) and Stock and Watson (1998, note 1).*

[4] *Fair and Shiller, using the Fair model, would have faced a daunting task in compiling a real-time data set of the hundreds of data series involved. Similarly, Stock and Watson (1998) employed more than 200 different time series in their forecasting study.*

BOX 1

The Composite Index of Leading Economic Indicators

The composite leading economic indicator (CLI) series was developed as a tool for business cycle analysis in the late 1960s. Prior to December 1995, the BEA produced the index of leading indicators data series. As of December 1995, the BEA stopped publishing the CLI and the Conference Board took over its production and publication. Detailed information regarding the construction of the CLI is available at the Conference Board Web site (www.tcbindicators.org).

The idea of the CLI is to summarize in one series the data on variables that typically move in a business cycle pattern prior to standard measures of economic output such as GDP. The primary objective is to help detect, ex ante, turning points in the business cycle—that is, whether the economy was likely to enter a recession (or to recover and grow out of a business cycle contraction). The business cycle research of Burns and Mitchell in the 1930s and 1940s helped motivate indicator analysis; however, the predominant researcher associated with the indicators (of which the leading indicators index is only one) was Geoffrey H. Moore (1990). While the CLI is primarily used as a turning point predictor, in recent publications the Conference Board has also suggested that it may be useful for forecasting the growth in economic output over time (Conference Board 1997, 1998).

The CLI is constructed as a weighted average of several publicly available data series. Currently there are ten component series in the index although both the number of series and the specific series used have changed over time. The weights applied to each series in forming the index are occasionally revised, and the index is usually recalculated every year to incorporate historical revisions to the component data.

Changes to the component series and the associated weights are in response to perceived changes in the empirical relationships between the components and the business cycle. The June 1997 issue of *Business Cycle Indicators* discusses in detail how the composition of the CLI has changed over time. The appendix in Beckman (1997) annotates the numerous revisions and improvements to the CLI historical data series.

It appears that changes to the composition of the CLI need not be substantial for them to be important in a real-time sense. Diebold and Rudebusch (1991) show that using revised historical CLI series in tests of the forecast value of the CLI generates spurious results supporting significant forecasting power for the CLI in predicting an index of industrial production. Because the results use revised data, the revised CLI reflects future information in both the choices of the component series as well as in the weights assigned to the component series in the index. The argument is that in the revision process the CLI is designed to maximize its correlation with the business cycle, so it would not be surprising that empirical results using revised data support the forecast power of the CLI more than do real-time vintages of the CLI. In contrast to those results, Hamilton and Perez-Quiros (1996) and the results in this article support a real-time role for the CLI in forecasting growth rates of real GNP and GDP, respectively.

set, then the exercise is not real-time and is unlikely to be a useful evaluation of real-time performance. Several studies that introduce aspects of real-time data construction fail to satisfy this criterion completely. Some other studies use real-time data sets but fail to use the most up-to-date versions available at the time the forecasts were made.

Research by Makridakis and others (1993) provides a good example of the few studies that undertake a real-time forecast analysis. In their research design, the authors provide real-time data sets to a group of forecasters and ask them to make forecasts several periods into the future. The research then evaluates the accuracy of the forecasts years later, after the actual data for the forecast observations have been released in a relatively final form. This type of research is a valuable contribution to the forecasting literature in that it evaluates forecasts in a true real-time framework.

One drawback to this approach is the long time lag needed to generate forecast accuracy results. The forecasters were given real-time data in 1987, 1988, and 1989, and forecasts were made for up to fifteen months ahead in each case. However, the authors performed the evaluation in 1991, using the vintage of historical data then available. For an academic exercise, the work is useful and informative. From a policy perspective, it is too slow in producing the information necessary to distinguish between good and bad forecasting models.

In related work, McNees (1992, 1995) investigates the GDP (or gross national product [GNP]) forecasting performance of several private- and public-sector macroeconomic forecasters. McNees evaluates the outcomes of the real-time forecasts over several historical periods, comparing the forecasts with revised GDP (or GNP) data. Although McNees discusses the issue of what vintage GDP the forecasts should be compared with, he bases his decision to use the most revised GDP (GNP) series on the idea that this vintage of data has the least measurement error.

McNees's studies are important checks on the forecast accuracy of many macroeconomic forecasters. By maintaining a real-time data set of the actual forecasts, he offers objective evaluation criteria for real-time macroeconomic forecasts. His studies, however, do not focus on constructing or evaluating forecasting models per se but on forecast outcomes. Thus, there is no analysis of the impact of real-time data on models or on model selection.

Recent studies by Swanson (1996) and Swanson and White (1997a, b) also provide a useful benchmark for research on real-time aspects of forecasting. Swanson (1996) collects the initial (or first-reported) estimates for a number of variables. For example, for GDP he creates a vector of all the advance GDP estimates for each quarter. These data clearly could be used in a real-time forecasting environment. However, these data are not what a forecaster would actually use in generating a forecast. For example, rather than using a vector of GDP advance estimates, a forecaster would use a vector defined by the available data vintage, containing the newly released observation (advance, preliminary, or final) along with revised values for the prior observations. Thus, at the end of July 1988, the GDP data through the end of 1982 would be obtained from the BEA's December 1985 benchmark (historical) revision. Those for 1983 are from the July 1986 annual revision, those for 1984 are from the July 1987 annual revision, and those for 1985 through the first quarter of 1988 are from the July 1988 annual revision. In essence, a real-time data set is the time sequence of vintages of data—each vintage is a vector of data values. A newer vintage data vector usually contains more observations than does an older vintage and has also usually been subjected to more revisions.[5]

Swanson (1996) and Swanson and White (1997a, b) have used the data set of initial estimates in a number of empirical analyses. For instance, Swanson (1996) compares a set of statistical tests formed using the initial estimate data with outcomes obtained using the most recent vintage of data. He finds a number of instances in which the test results are substantially different, suggesting that data revision matters. However, Swanson provides a test of a more extreme information restriction than would represent an actual real-time forecasting effort. If Swanson found no difference between using initial estimates and latest available data, then it is doubtful that the difference between real-time data and latest available data would be important for forecast accuracy results.

Recent research on the accuracy of forecasting models has moved closer to satisfying the necessary conditions of a real-time exercise.

Diebold and Rudebusch (1991) and Hamilton and Perez-Quiros (1996) present empirical results of out-of-sample forecast analyses in which one of the two variables in the forecasting models was measured in a real-time context. Diebold and Rudebusch examine whether the composite index of leading economic indicators (CLI) is useful for forecasting real output. Specifically, they investigate how well the CLI can forecast the industrial production index relative to an autoregressive model that uses only current and past values of the IP index. They cite previous research suggesting the CLI series was a strong predictor of IP. However, Diebold and Rudebusch hypothesized that using only the latest revised vintage of historical CLI data might have inflated the significance of the CLI's forecasting potential for output. They argue that the construction of the CLI has been subject to change (see Box 1), and these changes might constitute ex post attempts to better correlate the CLI with output. Using a real-time CLI series, they find that the CLI does not add significant forecasting power to an autoregressive forecasting model for IP.

Notably, Diebold and Rudebusch decided not to use real-time IP data in their empirical analysis. In particular, they used only the latest vintage when estimating the models and evaluating the forecast accuracy. By using the most recent vintage of IP data when forming their forecasts, they were in fact doing something impossible in real time. Diebold and Rudebusch justify this decision by arguing that they are searching for evidence on "the

[5] Over the period covered by this data set, the BEA rebenchmarked the data series several times, making the level of the real GDP series discontinuous. To address this problem, Swanson converts all the initial estimates into a single series based in 1987. Doing so raises the issue of the influence of the benchmark on the behavior of the spliced data series.

ability of the CLI to forecast truth, which is taken to be the final IP value" (1991, 609). They reason that the latest revision is the best estimate of real output. In other words, using real-time IP in fitting the model and constructing the forecasts could mask the fact that the CLI has little "intrinsic" forecasting ability for the "true" IP. Any forecasting ability when using real-time IP would reflect that the CLI was simply compensating for the inadequacies of the real-time IP measure. Of course, this feature is characteristic of any real-time forecasting problem; the best data estimates are not usually available at the time a forecast is made. Moreover, while one can debate which vintage of IP the forecasts should be evaluated against, waiting (up to twenty years in Diebold and Rudebusch's case) for a more refined IP estimate is hardly a realistic strategy for judging a professional forecaster or policy model.

Hamilton and Perez-Quiros (1996) also examine the real-time forecast performance of the CLI, but they focus on forecasting real GNP rather than industrial production. To translate the monthly CLI observations to a quarterly frequency, they use the first revised CLI estimate for the last month in the quarter. This number is usually released late in the second month of the next quarter. For example, they would use the revision of the March CLI that is released late in May, making the March CLI estimate roughly contemporaneous with the preliminary estimate for the first quarter's real GNP. Despite incorporating sufficient detail for making the vintage of CLI approximate real-time data, Hamilton and Perez-Quiros do not take into account the real-time availability of the GNP series in constructing the forecasts. Their justification is that they only "want to evaluate how close the forecast is to the value of GNP as ultimately revised" (1996, 42). However, as argued above, the choice of vintage to evaluate the forecasts against is somewhat different from the problem of constructing forecasts in real time. Hamilton and Perez-Quiros effectively make no distinction between data availability in constructing real-time forecasts and in evaluating the subsequent forecast.[6]

Real-Time Forecasting Experiments—Comparisons with Latest Vintages of Data

A simple experiment helps examine how the data revision process affects the selection and evaluation of economic forecasting models. The objective is to uncover whether the CLI can help forecast economic activity in real time. Separate forecasting models are estimated for two economic output measures—quarterly real GDP growth and the monthly growth rate of IP.

(1) Forecasts are constructed at the end of a month. The approach is to choose the specification of the forecasting model for each output series in real time and then estimate the coefficients of the chosen model using only the data actually available at the time the forecast is made.[7]

(2) Forecasts of GDP growth and IP growth were constructed for each of the next two time periods—quarters for GDP and months for IP.

(3) The forecasts were compared with subsequently announced values of the output series. For the real GDP series, the comparison is with the final estimate of the growth rate reported three months after the advance estimate. For the IP series, it is with the first release of the next month of data, as well as against the next two subsequent revisions of that initial estimate.

(4) Repeating the above steps each period through 1998 generates a sequence of real-time, one- and two-period-ahead forecasts for GDP and IP growth. The last step is to compute summary measures of forecast accuracy from these sets of forecast errors.

To be concrete, suppose the task is to examine a forecast of the growth rate of IP for September 1988 and the forecast is formed at the end of July 1988. This is a two-month-ahead forecast. At the end of July there is available an initial estimate of IP for June, a revised estimate for May, a second revised estimate for April, and a historical series constructed from annual revisions in 1986 and 1987 and the benchmark revision released in December 1985. At the end of July there is also an initial estimate of the CLI for June 1988. This estimate can be combined with a historical CLI series obtained from a major revision in February 1983, a revision to post-1983 data in March 1987, and a revision to the most recent twelve months' data in July 1988. This is the data set used to determine the model specification (lag length), estimate the model coefficients, and make a forecast of monthly IP growth for September 1988. The resulting forecast is compared with the initial estimate of IP for September released in October as well as with the revisions released in November and December.

The goal is to determine whether following the above procedure for replicating real-time forecasts produces results that compel inferences different from those of an analogous simulation using the latest available data vintage throughout. An experiment using the July 1998 vintage of historical time series investigates these differences, following the above steps and acting as if the numbers in this data set were actually available in real time. Thus, in updating the forecasting model, a new observation is added, but the historical observations do not change. Similarly, the accuracy of the resulting forecasting model is always evaluated against the latest, 1998 vintage of historical data.

The specification of the output models that include the CLI uses a bivariate vector autoregression (VAR) of the form

$$\begin{aligned}\Delta y_t &= m_1 + \sum_{i=1}^{p}(a_i \Delta y_{t-i} + b_i \Delta x_{t-i}) + u_t \\ \Delta x_t &= m_2 + \sum_{i=1}^{p}(c_i \Delta y_{t-i} + d_i \Delta x_{t-i}) + v_t\end{aligned} \tag{1}$$

where Δ denotes the first-difference operation. When t represents a quarter, y_t is 400 times the natural logarithm of GDP for quarter t and x_t is 400 times the natural logarithm of the CLI for quarter t. When t represents months, y_t is 1,200 times the natural logarithm of IP for month t and x_t is 1,200 times the natural logarithm of the CLI for month t. The errors u_{t+h} and v_{t+h} for h = 1 and 2 are taken to be unforecastable relative to the current and past of Δy_t and Δx_t and so are set to zero in constructing the forecasts of Δy_{t+1} and Δy_{t+2}. The number of lagged observations to include, p, and the values of the coefficients are all unknowns estimated from the available historical data at the time the forecast is made. The lag length is chosen by selecting the p that minimizes the so-called Akaike information criterion (AIC). The AIC is a statistic that trades off the improved fit of the model to the data, gained by including more lags, with the cost of having to estimate more and more coefficients from a fixed number of observations.

An autoregressive (AR) model for output growth is obtained by imposing the restriction in the first equation of (1) that the $b_i = 0$ for $i = 1, \ldots, p$. The AR model ignores the CLI data completely and relies instead solely on current and past values of the output measure for forecasting future output growth. The AR model thus provides a benchmark against which the VAR's forecasts can be compared, although such a comparison is a rather low hurdle. Notice also that specifying the models in the first differences of the variables precludes the possibility that the levels of the CLI and output might provide additional forecasting ability to the models.[8]

Using the CLI to Forecast Real GDP Growth in Real Time. The experiment examines whether the CLI measured in real time helps forecast real GDP growth over one- and two-quarter forecast horizons. To construct a real-time forecasting test, the study uses only data for both the CLI and real GDP that would have been available at the time the forecasts were made. In essence the experiment is examining the same basic issue that Hamilton and Perez-Quiros (1996) explored, but altering the data set in a number of ways ensures that both the construction and evaluation of the forecasts are closer to real-time exercises. The results from the real-time simulation are then compared with those obtained using the most recent vintage of time series in a pseudo real-time analysis.

Differences in the assessment of forecast performance arise primarily from the choice of series to evaluate against.

As explained in Box 2 on the construction of GDP and Box 1 on the CLI, the historical data on GDP and the CLI can change from month to month. Making the real-time GDP data coincide with the timing of the leading indicator series means considering a number of alternatives. One possibility is to use the second revision of the CLI estimate that corresponds to the last month of the quarter (as opposed to the first revision used in Hamilton and Perez-Quiros 1996). Thus, for example, a second revision of the March CLI is released in late June, and this revision could be aligned with the final estimate of the first-quarter GDP released in mid- to late June.[9]

Deciding to use a real-time data set pins down one aspect of the forecast evaluation exercise, but there are many other choices to be made that may, in principle, qualitatively affect the results. Of course, this difficulty appears precisely because the test is replicating a real-time forecasting problem.

As noted above, the model specification employs growth rates of the CLI and real GDP. The VAR model is first fit to data covering the period from 1959:1 to 1977:1, with a maximum of p = 4 lags of each variable included in the VAR. Respecifying the lag structure and reestimating the model's coefficients for each quarter through 1997:4 provides a framework allowing the most flexibility

[6] *Another curious feature of the Hamilton and Perez-Quiros study is that the authors estimate the chosen model specification up to 1975:3 in order to generate the out-of-sample forecasts from 1975:4 to 1993:2. Thus, they ignore all the interim information in the data that may alter the coefficient estimates. In real time, a forecaster would likely attempt to incorporate more recent information by updating the coefficient estimates periodically.*

[7] *The model specification is described in equation 1.*

[8] *The growth rate forecast results are all qualitatively the same if the VAR and AR models are fitted in levels of GDP, the CLI, and IP. However, the forecast accuracy of each model is always greater when the growth rate model specification is used.*

[9] *By the time these data are collected the current quarter is virtually over, a fait accompli. In essence, then, a one-quarter-ahead forecast amounts to predicting how the BEA will measure that quarter's GDP growth. As an alternative, the experiment also matched, for example, the first revision of March's CLI with the advance GDP estimate for the first quarter, both of which are released late in April. Hence, the forecast could be made only one month into the second quarter. Using this earlier vintage of data did not materially affect the forecast accuracy results described in the next subsection. The issue of matching vintages of the CLI and GDP data does not arise if one uses only the latest available data vintage.*

BOX 2

BEA Revisions of NIPA Economic Data

The Bureau of Economic Analysis (BEA), a division of the U.S. Department of Commerce, puts a substantial amount of its resources into the production of the National Income and Product Accounts (NIPA). Its efforts include both source data gathering—that is, compiling data on the measures that are combined into GDP—and statistical refinement of the data (seasonal adjustment, redefinitions, and rebenchmarking of the price-deflated series). Data production must meet the demands of its users for timely release of data measures but must also take great care to produce accurate data measurement. In such an environment, the BEA tries to satisfy both needs—timeliness and accuracy—by producing three estimates of GDP for the prior quarter.[1]

The first estimate, referred to as the advance GDP estimate, is released toward the end of the month immediately following the quarter to which the data refer. This advance estimate of GDP is based on incomplete source data, but it usually provides a fairly good forecast of the value of future revisions to that quarter's GDP because of how much source data, mainly on consumption, is available. However, because the BEA lacks complete source data for some subcomponents it must make judgmental assumptions about the likely values taken by specific GDP components. This action is simply "forecasting" what the outcome might be.

The so-called preliminary GDP estimate is a revision to the advance estimate that is released toward the end of the second month after the quarter, and a final GDP estimate for the quarter is released toward the end of the third month. The main reason for the revisions of these numbers from the advance to preliminary (and to final) estimates is that the source data for these measures take time to arrive at the BEA to be compiled into the statistics.

The BEA schedules additional revisions that improve the accuracy of the GDP estimate after the release of the final GDP estimate. The revision occurs at this time to improve the estimate of seasonal adjustment of the data. These annual revisions of the data usually occur in July, revising the data in the prior three calendar years as well as the one quarter of data available for the given year.

The BEA then revises the entire history of the quarterly data series (currently back to 1959) approximately every five years in what is known as a benchmark revision. At these benchmark revisions, there are often redefinitions of component data series that revise the entire history of the series. Here, the BEA updates the base year for computing the real measure of GDP and the implicit deflator. Base-year changes often have substantial effects on the overall estimates of economic growth because the initial relative price conditions set in the benchmark year may change dramatically as time passes. In other words, base-year effects alter growth rates in real GDP because the relative prices at which the new real GDP estimate is calculated could differ from those of the previous base year.

The most obvious example of this phenomenon is the 1972 benchmark during the mid-1970s. There were substantial oil price increases in 1973 through 1974 that changed dramatically the relative prices between oil and other goods. Deflating nominal oil prices by using a deflator based on 1972 prices lowered the measured adverse impact of oil imports on real net exports. Looking at GDP measures based on alternative base years provides quite different views of the depth of the economic contraction during the 1974–75 recession. Using a 1977 base year, the relative size of the economic contraction appears larger because oil was a larger share of imports in that year (as a result of both higher prices and the larger quantity of imported oil). Benchmark revision data are generally more accurate because they use more revised source data from which to measure economic activity. For instance, the BEA uses more final information sources because there is more time to check the validity of initial reports.

In 1995 the BEA changed the definition of the real GDP measure by moving to a chain-weighted index, which allows for the effects of changes in relative prices and in the composition of output over time (see Landefeld and Parker 1997). This change in the definition of real GDP alters the behavior of the estimated real GDP series relative to the prior, fixed-weight constant-dollar estimates. If a forecaster uses this series as the series against which real-time forecasts are compared, then implicitly the forecaster is attempting to forecast the change in the definition of real GDP. Part of the motivation for this study is to investigate whether the change in definition is a sizable problem for real-time forecasters.

Research by the BEA (such as Young 1993) examines how each subsequent revision in the GDP series has changed the series and how good initial growth rate measures are as estimates of the more recent vintages of the series. Simply stated, the changes from one announcement to the next reflect the tension between the need for data that are both timely and accurate. One expects the later estimate to be more accurate than the advance estimate, but it is not always.

BOX 2 (CONTINUED)

What is the source of the revisions? First, as time passes, the BEA can replace preliminary source data with more revised or comprehensive data. More complete monthly data may be one example of this data revision source. As mentioned above, the advance estimate of real GDP often contains BEA judgmental estimates of measures that the BEA does not possess only one month after the end of the quarter. In the subsequent revisions, these estimates are replaced with source data as they become available. This type of revision occurs with relatively high frequency. One would expect that the BEA would on average be relatively accurate so that the advance and preliminary real GDP data would not possess an obvious bias, for example, an average positive or negative error relative to the final measure for that quarter. Young (1993) offers evidence to support the accuracy of the estimates of real GDP growth rates; in the most recent sample, there appears to be no significant bias in the three announcements or even in the advance estimate relative to a temporally close "latest available" estimate.[2]

As argued by Mariano and Tanizaki (1995), the "true" real GDP measure for any particular quarter is effectively unobservable because the current measure will be subject to future revision. The position held here is that the latest available time series of real GDP is the best historical record currently available. However, no part of this currently available data set would have been available to a researcher in earlier time periods when making forecasts of the then-unknown future values of the series. In the same way, no researcher today has available the data set that will eventually exist when subsequent revisions are made in one or five or ten years.

[1] *The BEA shifted its focus in November 1991 from reporting gross national product (production by U.S. nationals regardless of the location of the factors of production) to GDP (production within the borders of the United States regardless of production factor ownership). Real GNP was long recognized as harder to produce in a timely fashion than real GDP because there are little reliable, timely data on net income from foreign sources. For the United States the numerical difference between the two constructs is relatively small.*

[2] *Importantly, Young (1993) avoids comparing the advance estimate with the most revised, latest available estimate because the latest version is often a substantially revised measure of a possibly even redefined construct. Fleming, Jordan, and Lang (1996) examine accuracy of the measured level of real GDP over the limited sample 1985 to 1991 and find evidence of sizable and systematic measurement bias. Young's results, however, suggest that these findings do not translate into systematic biases for the growth rates.*

for the statistical model to adjust to the new information that arrives with each additional observation. Each time the model was reestimated, new one- and two-quarter-ahead forecasts were generated, yielding a set of eighty-four one-quarter-ahead forecasts and eighty-three two-quarter-ahead forecasts that could be evaluated against final GDP numbers for 1977:2 to 1998:1. Forecasts constructed using real-time data were compared with forecasts from models estimated using the most recent vintage data.

For measuring the accuracy of the forecasts, the decision of which vintage data to compare the forecasts against is an important one. In real time, this choice is clearly important. For example, who knew in 1981 (or even 1991) that the BEA would change to the use of chain-weighted real GDP in 1996? It is perhaps unfair to burden the forecaster in 1981 with the problem of also forecasting definitional changes in the data series. There is a second issue: The most-revised data series takes many years to produce. It is doubtful that a forecaster's accuracy would not be evaluated until years later when the most-revised time series is determined. More likely, the forecasts will be compared with the final GDP growth estimate released approximately three months after the end of the quarter being forecast. Thus, the study compares the actual real-time forecasts to the growth rate implied by the initial final GDP estimate, whereas it compares the pseudo real-time forecasts with the most recent (1998) vintage of data.

Empirical Results for Real GDP. Relative forecast accuracy results are reported based on the root mean squared error (RMSE). This figure is simply the square root of the average of the squared forecast errors, with the square root taken to put the measure back into the units of the variable being forecast (that is, annualized percentage points). The same pattern of results appears using other standard forecast accuracy measures such as the average absolute forecast errors (the average of the sum of forecast errors with sign disregarded). One would expect the variability of the errors to be smaller the more accurate the model is.

Table 1 displays the summary forecast statistics comparing the one- and two-quarter-ahead forecasting performance for the VAR and autoregressive models. For

TABLE 1 RMSE of Pseudo and Real-Time GDP Growth Forecasts

		Forecast Model			
		VAR		AR	
Forecast Type and Timing	Evaluated against	One-Quarter	Two-Quarter	One-Quarter	Two-Quarter
Pseudo (July 1998)	July 1998 vintage	3.02	3.13	3.40	3.53
Real-time (end of quarter)	Initial final estimate	2.56	2.73	2.88	3.00

Source: GDP, Bureau of Economic Analysis; CLI, Conference Board

both forecast horizons and all the data sets, the RMSE of the VAR is considerably less than the RMSE of the AR model. Thus, the model that includes the CLI provides more accurate forecasts than the AR alternative. Consistent with this finding, the real-time forecasts are more highly correlated with the initial final vintage of GDP growth estimates, and the pseudo real-time forecasts are more highly correlated with the latest available vintage of estimates than are either of the corresponding forecasts from the AR model.

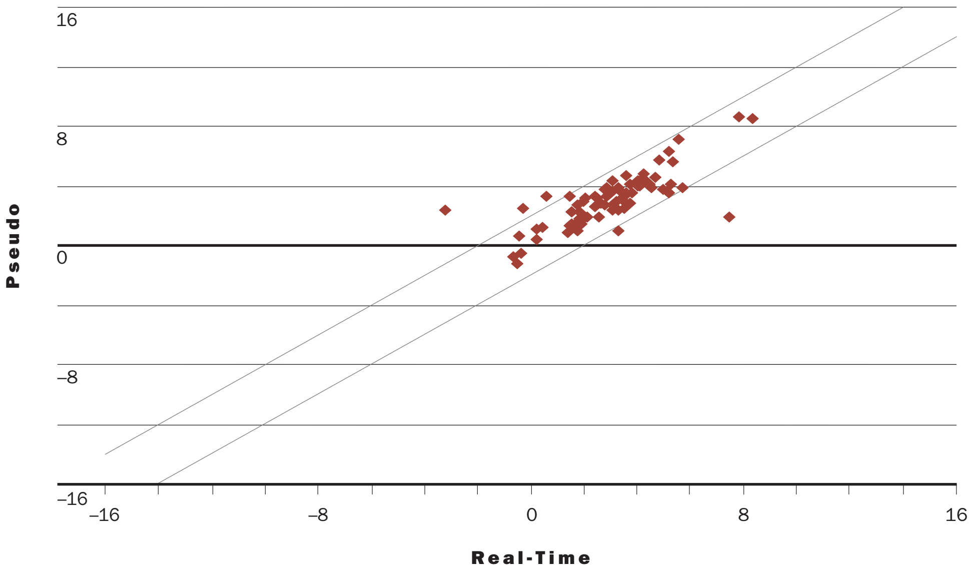
In discussing the results, it is more compact to report the ratio of the RMSE of the VAR model with CLI to the RMSE of the AR model, given that both models employ the same data set restrictions. Using the latest available data vintage, the ratio of RMSE is 0.89 for the one- and two-step horizons. The ratios using real-time data are 0.89 for the one-step horizon and 0.91 for the two-step horizon. The differences between the real-time and latest available data vintages do not seem to change the basic inference that the CLI has some marginal predictive power for GDP.

The next step examines the differences between the measured accuracy of the actual and pseudo real-time VAR forecasts in a little more detail. Chart 3 is a scatter diagram of the one-quarter-ahead forecasts from the VAR for the period from 1977:2 to 1998:1. In the chart, each point reflects the actual real-time forecast on the x-axis and the corresponding pseudo real-time forecast on the y-axis. Points along the 45-degree line indicate that the two forecasts are the same. Only six of the eighty-four forecasts (7 percent) differ by more than 2 percentage points, emphasizing that the forecasts are quite similar despite being based on different vintages of historical data.[10] Chart 4 is a scatter diagram of quarter-on-quarter GDP growth rate estimates for 1977:2 to 1998:1. The July 1998 vintage of GDP growth estimates is on the y-axis, and the initial final vintage is on the x-axis. Examining this chart reveals that thirteen observations out of the eighty-four, slightly more than 15 percent of the revisions, are more than 2 percentage points apart. Comparing Chart 3 with Chart 4 makes it clear that the difference in the measured forecast accuracy arises primarily from the variation between the versions of data being forecast rather than the forecasts themselves.[11] The mean and standard deviation of the most recent vintage are 2.8 and 3.5, whereas the same statistics for the initial final estimates are 2.6 and 3.0, respectively. It is notable that the later vintage of real GDP growth is also more variable.

This empirical evidence indicates that the CLI helps predict real GDP growth, but the forecast accuracy statistics themselves are rather unimpressive. To illustrate this fact consider what happens if one uses the advance estimate of real GDP growth as the forecast for that quarter. The advance estimate is available approximately three weeks after the real-time VAR/AR forecasts are formed at the end of the quarter. Thus, for example, instead of a forecast of second-quarter GDP growth formed at the end of June, the advance estimate can be thought of as a second-quarter forecast formed late in July. In the first case, the advance forecast generates a humbling 0.75 percent RMSE when evaluated against the resulting final estimate of real GDP. This percentage is substantially lower than the RMSE of 2.56 obtained from the real-time VAR model, as reported in Table 1. Moreover, the correlation of the advance estimate with the final estimate is approximately 0.97, as compared with only 0.54 for the real-time VAR model forecasts. The strong results are understandable since the advance GDP estimate uses the same BEA data measurement design and much of the same source data are used to generate the final estimate released only two months later.

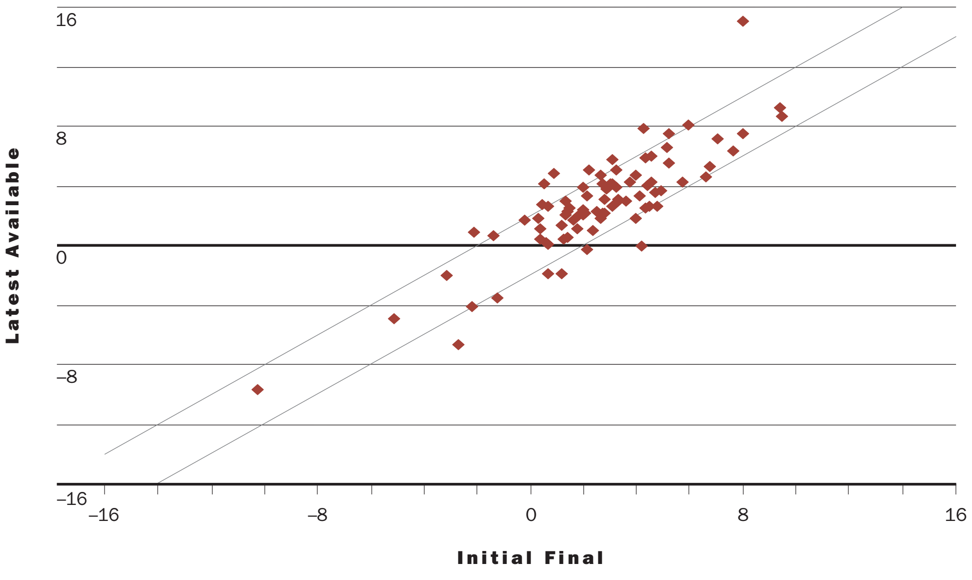
When the advance estimate is compared with the latest available vintage of GDP growth for the current quarter, the forecast error increases substantially—the RMSE is 1.9 percent—and the correlation with the latest vintage of estimated GDP growth drops to 0.84. The increase in the forecast error and the decline in the correlation emphasize the importance in the choice of the estimate against which forecasts will be evaluated. Revision and redefinition of the GDP series over time almost guarantees that real-time forecasts will worsen when forecast accuracy statistics are taken relative to the most recent vintage of data. Nonetheless, the advance

CHART 3 One-Quarter-Ahead VAR GDP Growth Forecasts, 1977:2–1998:1



Source: GDP, Bureau of Economic Analysis; CLI, Conference Board

CHART 4 GDP Growth Estimates, 1977:2–1998:1



Source: Bureau of Economic Analysis

[10] *The fact that the forecasts simply are not very different is emphasized in observing that comparing the real-time forecasts with the most recent vintage of data yields RMSE values that are virtually the same as those from the pseudo out-of-sample forecasts.*

[11] *The finding that real-time GDP forecasts are more accurate for the penultimate data than the pseudo real-time forecasts are for the most recent data vintage is consistent with McNees (1988), who notes that forecast errors are generally smaller when real-time forecasts are compared with less-revised vintages of data.*

BOX 3

The Index of Industrial Production

The Board of Governors (BOG) of the Federal Reserve System produces the index of industrial production (IP) on a monthly basis; it is one of the few measures of output produced monthly.[1] The index estimates output in a number of industrial sectors that, combined, currently account for approximately 25 percent of total output. That portion of total output (as measured by real GDP) has diminished over time but remains an oft-cited economic statistic. It measures the change in output in the following industrial sectors: manufacturing, mining, and electric and gas utilities. Output is measured in physical units, rather than by price and quantity. The index excludes output in other activities, such as agriculture and services. Thus, the index numbers that the BOG releases midmonth are estimates of the monthly level of total output of the nation's factories, mines, and gas and electric utilities.

The construction of the IP index involves a substantial degree of estimation. Typically, less reliable source data are available on a more timely basis than are the more accurate measures used to revise estimates of IP. But even in the composition of IP, there are three different types of input series for estimating the index: physical product, production-worker hours, and electrical power use by industry. For some industries the monthly estimates of production are based on measures of physical output, and that is the most desirable measure of output. Physical product counts the physical output in quantity. These data are not often available in a timely manner. For industries in which direct measurement of physical product is not readily available, the BOG estimates (infers) a measure of output from production-worker hours per industry from numbers produced by the Bureau of the Census or from electrical power use. These data are used in lieu of the physical product data that are not yet available.

The IP figures are available in the middle of the month following the month they measure. The BOG issues preliminary data for the preceding month, and these data are subsequently revised in the next three months. Annual revisions are made in the fall. The BOG revises the series to a greater degree on a periodic basis, linking or benchmarking the individual industrial production series to more comprehensive data sources. One of the major sources for benchmark revisions is the Census of Manufactures, which is released every five years. The IP index was built, for the most part, in five-year segments, each with value-added weights taken from the census year. Now, like real GDP, the IP index is a chain-weighted index.

The major revisions are in the IP series (1971, 1976, 1985, 1990, and in 1997). The BOG completed a revision of its measures of industrial production in January 1997. The primary feature of that particular revision was a new formulation for aggregating the index using weights that are updated annually instead of every five years. The revisions of the data series went back as far as 1977, but some additional changes were made to data from 1976 back to 1967 to improve their consistency with the new data formulation. In addition, the revision also involved the rebasing of the total IP series back to the initial observation (1919); the data are now expressed as percentages of output in 1992.

The IP index, despite covering only about 20 percent of total U.S. output, measures industries that may account for a large proportion of output volatility during a business cycle. Typically, the IP index rises more during economic expansions, and contracts more during economic downturns, than the aggregate real GDP series. Also, it is released more frequently than other output measures (like real GDP). However, the timeliness of the series must also be compared with its measurement error: relative to real GDP, IP estimates appear to have more substantial measurement error.

[1] *Frumkin (1994) and Rogers (1998) provide detailed information on the construction, release timing, and revision schedule of various economic indicators including the IP index, as well as their standard uses and interpretations.*

TABLE 2 RMSE of Pseudo and Real-Time IP Growth Forecasts

		Forecast Model			
		VAR		AR	
Forecast Type and Timing	Evaluated against	One-Month	Two-Month	One-Month	Two-Month
Pseudo (July 1998)	July 1998 vintage	5.50	5.39	6.10	6.00
Real-time (end of month)	First revised estimate	5.41	5.40	5.33	5.49

Source: IP, Board of Governors of the Federal Reserve System; CLI, Conference Board

estimate performs considerably better than either of the one-quarter-ahead real-time forecasting models.

Using the CLI to Forecast Growth of Industrial Production in Real Time. Another examination of the forecasting properties of the CLI looks at its marginal forecasting contribution for IP, a more frequently released output measure. As described in Box 3, IP is a less general output measure than real GDP—it measures only output in mining, manufacturing, and electric and gas utilities—comprising somewhere around 25 percent of total U.S. output. Still, IP is an oft-cited economic measure and was central to Diebold and Rudebusch's (1991) research on the predictive power of the CLI.

The real GDP forecasting example combines a more frequent activity indicator (CLI) with the quarterly real activity measure. In that application, it was necessary to choose the CLI measure for the month in the quarter that coincided with the relevant release of quarterly real GDP growth data. For the IP forecasting application, the same statistical framework is used, but the model is fitted to the CLI and IP on a monthly frequency. A monthly IP measure for a given month is released in the middle of the subsequent month. The corresponding CLI number is released at the end of the month following the month to which it refers. Because of this slight staggering in the releases within the month, it is assumed that the IP forecasts are formed at the end of the month so that both figures are available for the prior month. This decision rule, then, means that at the end of July 1988, say, there are measures of the CLI and IP up to June 1988, and IP growth is forecast for July and August. The Board of Governors of the Federal Reserve System releases an initial IP estimate for July in mid-August, or about fifteen days after the forecast is made. A first revision of this estimate is released one month later in mid-September, and a second revision is reported after another month has elapsed. As noted above, the real-time forecasts are compared with the initial estimate, the first revision, and the second revision. However, these different vintages have little impact on the results for the real-time forecasting models, so results report only what is based on comparison with the first revised data. In contrast, the pseudo real-time forecasts are always compared with the vintage of historical data available in July 1988.

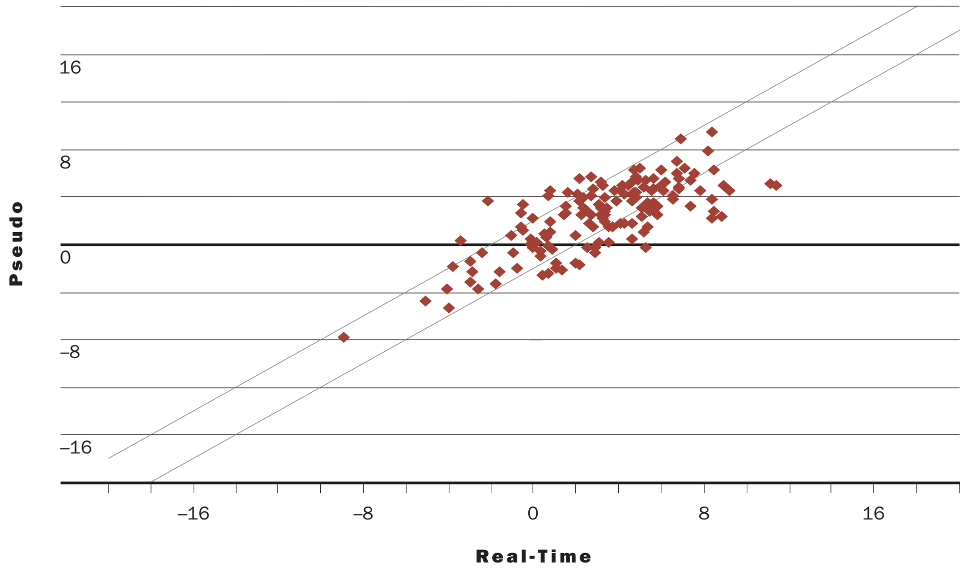
The statistical model shown in equation 1 is employed to relate the monthly growth rates of the IP and CLI series. The models are first fit to data covering the period from 1959:01 to 1985:12, with a maximum of $p = 12$ lags of each variable included in the VAR. The actual lag length is chosen via the AIC. The lag structure is then respecified and the model's coefficients are reestimated for each month through 1998:02. Each time the model was reestimated new one- and two-month-ahead forecasts were generated, yielding a set of 147 one-month-ahead forecasts and 146 two-month-ahead forecasts that could be evaluated against IP growth rates for 1986:01 to 1998:03.[12]

To analyze the contribution of the CLI to forecasts of IP, the forecast accuracy of models that use only past observations of IP (AR models) is compared with that of models that also exploit the CLI data (the VAR model). This simple criterion provides the same low hurdle for the CLI data that were examined with the real GDP data: if the CLI data contribute to the forecast accuracy of IP, then the VAR model that includes the CLI will generate forecasts with lower RMSE than those of the AR models.

Table 2 presents the summary forecast statistics comparing the one- and two-month-ahead forecasting performance for the VAR and autoregressive models. For the pseudo real-time forecasts the RMSE of the VAR is considerably less than that of the AR model at both forecast horizons. The RMSE ratio is 0.90, suggesting that the use of the CLI helps reduce forecast error by about 10 percent. For the two-step horizon, the ratio is 0.89. In a separate experiment combining real-time CLI

[12] *The study notes that Diebold and Rudebusch (1991) examine whether the level of the CLI improves the one-step-ahead forecasts for the level of IP. The model examines the ability of growth rates (percentage changes) in the CLI to help forecast the growth in IP so that the two sets of results are not directly comparable. The results with IP, however, produce inferences that are comparable to those made in their research.*

CHART 5 One-Month-Ahead VAR IP Growth Forecasts, 1986:01–1998:03



Source: IP, Board of Governors of the Federal Reserve System; CLI, Conference Board

data along with latest available IP data (similar to the work of Diebold and Rudebusch), the RMSE ratio moves to 0.95 and 0.93. Thus, using a hybrid data set that mixes data vintages reduces the measured forecast improvement, but it still suggests that the CLI provides a very marginal improvement in forecasting accuracy over the forecasting model that simply uses lags of IP.

The one-month-ahead forecasts from the real-time VAR and AR models evaluated against the first revised IP estimates produced a RMSE ratio of 1.03, suggesting that real-time CLI actually worsens the VAR model's forecast accuracy relative to a simple AR model. The ratio for the two-step-forecasting horizon is 0.98.[13] Thus, it seems that the CLI does not help forecast IP growth in a real-time setting. This result is noteworthy because the statistical results using the most recently revised series, and even those that combine revised and real-time data, favor including the CLI in a forecasting model of IP. Hence, forecast evaluation tests using the most recently revised data series for the CLI and IP will generate inferences that suggest a positive contribution of the CLI to forecasts of IP, and these inferences will not hold up in real-time applications.

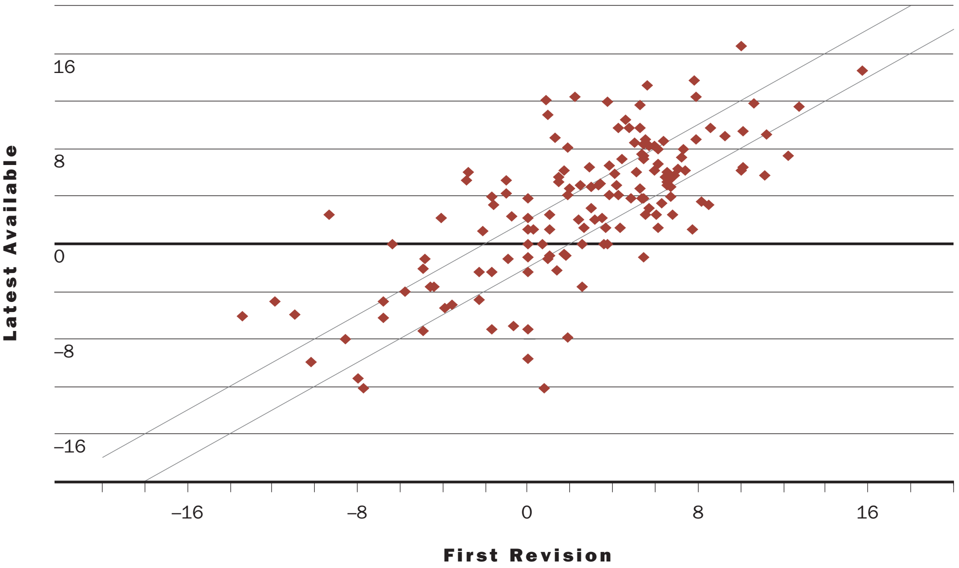
The differences between the real-time and pseudo real-time VAR forecasts are illustrated in Chart 5. This scatter diagram presents the one-month-ahead forecasts for the period 1986:01 to 1998:03, generated from the real-time VAR model and the VAR constructed using the most recent data vintage. There is considerably more variation between the forecasts due to data vintage than between the corresponding GDP forecasts; 46 out of 147 forecasts (31 percent) are different by more than 2 percentage points. The real-time one-month-ahead forecast has a correlation of 0.35 with the next month's revised estimate while the corresponding pseudo real-time forecasts have a correlation of 0.42 with the July 1998 vintage of estimates. Both are considerably lower than the corresponding correlation for the GDP forecasts.

Chart 6 is a scatter diagram of the latest available and the first-released monthly IP growth rates for the period from 1986:01 to 1998:03. Observe that 83 of the 147 observations (56 percent) differ by more than 2 percentage points, emphasizing that the latest available vintage of historical data over the forecast horizon differs substantially from the corresponding initial estimates. Comparing Chart 5 with Chart 6 makes it clear that the difference in the measured forecast accuracy arises primarily from the variation between the vintages of data being forecast rather than the forecasts themselves. The mean and standard deviation of the growth rates computed using the most recent data vintage are 2.95 and 6.04, whereas the same statistics for growth rates computed using the first revised estimates (and the second revision of the estimate for the preceding month) are 2.40 and 5.58, respectively. This pattern is the same one found for GDP data: the latest vintage of data is more variable and has a higher average than the less-revised estimates.

Conclusion

This article describes what historical real-time forecast evaluation should look like and how it is conceptually different from what is referred to here as a pseudo real-time forecast evaluation. The results suggest

CHART 6 IP Growth Estimates, 1986:01–1998:03



Source: Board of Governors of the Federal Reserve System

that using real-time vintages of data is a basic ingredient for generating valid out-of-sample forecast evaluations.

The practical question is whether a failure to use real-time data sets leads to inferences different from those made using only the latest available vintage of data. In principle, the specification and estimation of the forecasting model may differ due to the choice of the vintage of the data set, as may the evaluation of the model's forecasts. To shed some light on the practical importance of the issue, the article examines the ex ante forecast performance of two separate vector autoregressive (VAR) models. The discussion of both examples examines whether the CLI helps forecast measures of economic activity—real GDP growth and IP growth, respectively.

For real GDP, the results indicate (1) that the use of the latest vintage of historical time series on the CLI and real GDP does not cause the fitted VAR's forecasts to be much different from those of a VAR fitted using the actual historical data available at the time the forecasts were made and (2) that the choice of vintage of the real GDP data does alter the measured forecast accuracy of the VAR model but does not change the model's ranking. Relative to an autoregressive model for real GDP, knowing the value of the CLI within the current quarter leads to more accurate forecasts of GDP growth over each of the next two quarters.

For IP, the findings show (1) that using the latest vintage of the CLI and IP data does not cause the fitted VAR's forecasts to be much different from those from a VAR fitted using the actual historical data available at the time the forecasts were made and (2) that the pseudo real-time forecast results, when evaluated against the latest available data, suggest that the CLI can help predict the growth in IP. Using real-time data on the CLI, combined with latest available data vintage for IP (comparable to Diebold and Rudebusch 1991), generates weaker but still supportive results of the predictive power of the CLI for IP when compared with the latest available vintage of data. When the real-time forecasts are evaluated against the next available and nearby IP estimates, the results suggest that a VAR actually produces less accurate forecasts than does a simple AR model of IP. For the models considered here, failure to use real-time data in constructing and evaluating the forecasts was not too serious a problem for real GDP, but it produced an apparently misleading inference for the IP model.

Differences in the assessment of forecast performance arise primarily from the choice of series to evaluate against. The revisions to IP vintages of historical data are of a larger magnitude and are more extensive than those made to real GDP data. However, in both cases the differences among the data revisions are much larger than the differences among the forecasts. This insight reflects the fact that the models do not generate forecasts that vary greatly across vintages of historical data.

13. The ratio of the VAR to AR RMSE is always slightly greater than 1 when the real-time IP forecasts are compared with the initial IP growth estimate.

In the present case this finding affects the magnitude of the measures of accuracy for a given model as well as across models. Of course, the models used here involve only two series. It remains to be seen whether these empirical results generalize to more realistic forecasting models that typically involve a larger number of variables. Still, the article highlights the finding that the accuracy of results clearly depends upon the target series chosen as a forecast accuracy criterion.

REFERENCES

Beckman, Barry A. 1997. "Reflections on BEA's Experience with Leading Economic Indicators." Bureau of Economic Analysis, unpublished manuscript.

The Conference Board, Inc. 1997. "Using the Individual Leading Indicators to Predict Growth." *Business Cycle Indicators* 2 (April): 3–4.

———. 1998. "Leading Indicators and the Prospects for Growth in 1998." *Business Cycle Indicators* 3 (May): 3–4.

Diebold, Francis X., and Glenn Rudebusch. 1991. "Forecasting Output with the Composite Leasing Index: A Real-Time Analysis." *Journal of the American Statistical Association* 86:603–10.

Fair, Ray C., and Robert J. Shiller. 1990. "Comparing Information in Forecasts from Econometric Models." *American Economic Review* 80, no. 3:375–89.

Fleming, Martin, John S. Jordan, and Kathleen M. Lang. 1996. "The Impact of Measurement Error in the U.S. National Income and Product Accounts on Forecasts of GNP and Its Components." *Journal of Economic and Social Measurement* 22:89–102.

Frumkin, Norman. 1994. *Guide to Economic Indicators.* Armonk, N.Y.: M.E. Sharpe.

Hamilton, James D., and Gabriel Perez-Quiros. 1996. "What Do the Leading Indicators Lead?" *Journal of Business* 69:27–49.

Landefeld, J. Steven, and Robert P. Parker. 1997. "BEA's Chain Indexes, Time Series, and Measures of Long-Term Economic Growth." *Survey of Current Business* 77 (May): 58–68.

Makridakis, Spyros, Chris Chatfield, Michele Hibon, Michael Lawrence, Terence Mills, Keith Ord, and LeRoy F. Simmons. 1993. "The M2-Competition: A Real-Time Judgmentally Based Forecasting Study." *International Journal of Forecasting* 9:5–22.

Mariano, Roberto S., and Hisashi Tanizaki. 1995. "Prediction of Final Data with Use of Preliminary and/or Revised Data." *Journal of Forecasting* 14:351–80.

McNees, Stephen K. 1988. "How Accurate Are Macroeconomic Forecasts?" Federal Reserve Bank of Boston *New England Economic Review* (July/August): 15–36.

———. 1992. "How Large Are Economic Forecast Errors?" Federal Reserve Bank of Boston *New England Economic Review* (July/August): 25–42.

———. 1995. "An Assessment of the 'Official' Economic Forecasts." Federal Reserve Bank of Boston *New England Economic Review* (July/August): 13–23.

Moore, Geoffrey H. 1990. *Leading Indicators for the 1990s.* Homewood, Ill.: Dow Jones-Irwin.

Orphanides, Athanasios. 1997. "Monetary Policy Rules Based on Real-Time Data." Board of Governors of the Federal Reserve System, unpublished manuscript.

Rogers, Ralph Mark. 1998. *Handbook of Key Economic Indicators.* New York: McGraw-Hill.

Staiger, Douglas, James H. Stock, and Mark W. Watson. 1997. "The NAIRU, Unemployment, and Monetary Policy." *Journal of Economic Perspectives* 11 (Winter): 33–50.

Stock, James H., and Mark W. Watson. 1998. "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." National Bureau of Economic Research Working Paper No. 6607, June.

Swanson, Norman. 1996. "Forecasting Using First-Available versus Fully Revised Economic Time-Series Data." *Studies in Nonlinear Dynamics and Economics* 1, no. 1:47–64.

Swanson, Norman, and Halbert White. 1997a. "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks." *Review of Economics and Statistics* 79:540–50.

———. 1997b. "Forecasting Economic Time-Series Using Flexible versus Fixed Specification and Linear versus Nonlinear Econometric Models." *International Journal of Forecasting* 13:439–61.

Young, Allan H. 1993. "Reliability and Accuracy of the Quarterly Estimates of GDP." *Survey of Current Business* 73 (October): 29–43.