WORKING PAPERS
RESEARCH DEPARTMENT

WORKING PAPER NO. 02-3
IS MACROECONOMIC RESEARCH ROBUST TO ALTERNATIVE DATA SETS?

Dean Croushore
Tom Stark
Federal Reserve Bank of Philadelphia

March 2002

FEDERAL RESERVE BANK OF PHILADELPHIA
Ten Independence Mall, Philadelphia, PA 19106-1574• (215) 574-6428• www.phil.frb.org

WORKING PAPER 02-3

IS MACROECONOMIC RESEARCH ROBUST TO ALTERNATIVE DATA SETS?

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

March 2002

This is a substantially revised version of Working Paper No. 99-21. We thank Bill Dalasio, Ryan Hayward, Jim Sherma, and Bill Wong for their fine research assistance. We thank Athanasios Orphanides, Ellis Tallman, Ken West, participants in seminars at the Federal Reserve Bank of Philadelphia, the University of Pennsylvania, the International Finance Division at the Federal Reserve Board, and George Washington University, as well as those at the Midwest Macroeconomics meetings, the Pennsylvania Economics Association, the Federal Reserve System Committee on Macroeconomics, the National Bureau of Economic Research Summer Institute, and the Philadelphia Fed's Conference on Real-Time Data Analysis for their comments.

IS MACROECONOMIC RESEARCH ROBUST TO ALTERNATIVE DATA SETS?

Abstract

This paper uses a real-time data set to analyze data revisions and to test the robustness of published econometric results. The data set consists of vintages, or snapshots, of the major macroeconomic data available at quarterly intervals in real time. The paper illustrates why such data may matter, examines the properties of several of the variables in the data set across vintages, and examines key empirical papers in macroeconomics, investigating their robustness to different vintages.

JEL classification: E00, C82

IS MACROECONOMIC RESEARCH ROBUST TO ALTERNATIVE DATA SETS?

I. INTRODUCTION

Macroeconomists use historical data for a variety of purposes: to test models, to analyze economic events and policy, and to forecast. In many cases, however, the data that should be used in these studies are not the (final, revised) data available from government statistical agencies today, but rather the original, unrevised data available to economic agents who were around at the time. In other cases, the ability to verify published findings and to check the robustness of those findings to different data sets is an important test of the validity of the results.

These reasons motivated us to create a data set that gives snapshots of macroeconomic data available to an academic researcher, policymaker, or forecaster at any given date in the past. Following Swanson (1996), we refer to each data set corresponding to the information set at a particular date as a "vintage" and to the collection of such vintages as a "real-time data set." Further details about the data set and how it was constructed can be found in Croushore and Stark (2001).

This paper focuses on two main aspects of the data set: (1) examining the nature of data revisions; and (2) testing the robustness of some important macroeconomic studies to alternative choices of data vintages.

In thinking about the nature of data revisions, we draw a distinction between two different types of data revisions: (1) Some revisions occur because statistical agencies have additional source information with which to update their initial estimates of aggregates such as real GDP. We call these *information-based* data revisions. (2) Other revisions involve changes in the

structure of the economic data accounting system, such as changes in aggregation methods (for example, fixed weighting or chain weighting), changes in base years (for example, 1982 or 1987) used for calculating real variables, and changes in definitions of the concept being measured (for example, whether to treat government spending on capital goods as an expense or an investment good). We call these *structural* data revisions. These structural revisions (which occur during benchmark revisions) are the main focus of this paper, though it is difficult to separate the two types of revisions in the data. In particular, we are interested in studying what effect such structural revisions have on some important macroeconomic studies.

The first part of this paper looks at the nature of data revisions to see if we can make generalizations about their properties. We do this in two alternative ways: (1) using spectral analysis to analyze the revisions; and (2) testing for "news" and "noise" in the revisions. This paper is the first (that we know of) to use spectral analysis to analyze data revisions, though previous researchers have used other methods to analyze such revisions, most notably Zellner (1958), Mork (1987), Runkle (1998), Swanson, Ghysels, and Callan (1999), and Croushore and Stark (2001). We also expand on the news and noise tests of Mankiw, Runkle, and Shapiro (1984) and Mankiw and Shapiro (1986), looking at different variables over longer time spans.

The second part of the paper looks at the impact of data revisions on major macroeconomic studies. This is a unique approach that differs from the focus of the literature, which has mainly dealt with the impact of data revisions on policy analysis and on forecasting. Recently, many economists have engaged in research on the differences in policy rules if based on real-time compared to revised data, such as Orphanides (2001), though the literature dates back to Maravall and Pierce (1986). Other notable contributions include Ghysels, Swanson, and Callan (1999), Bernanke and Boivin (2001), and Rudebusch (2001). Another branch of the

literature looks at the impact of data revisions on forecasts, with contributions by Cole (1969), Howrey (1978), Diebold and Rudebusch (1991), Swanson (1996), Amato and Swanson (2001), and Stark and Croushore (2001). Since the literature on data revisions shows that such revisions can be important, and our work in the first part of this paper suggests that certain types of revisions are hard to characterize, in the second part of the paper we examine the robustness of macroeconomic studies to variations in the data set that a researcher uses.

A priori, the fundamental notion of robustness is that an economic theory that is being tested for its empirical validity should hold up under structural variations in the data set about which most theories have little to say. For example, a theory about how real output behaves over the business cycle should be true whether our measure of output is GNP or GDP, whether data on real GNP is constructed using fixed-weight methods or chain-weight methods, whether the base year for creating real variables changes, or whether GDP is redefined to include new goods or to change the categories for certain components, such as business software or government investment spending. For us to believe that a theory is consistent with the data, empirical research supporting the theory must prove to be robust—that is, it should not be sensitive to variations in the data, about which the theory has little to say.

Once a researcher considers taking his theory to the data, it might seem that the latest version of the data at a researcher's disposal is the best version, and is the only version that should be used. But there are three ways in which that may not be true. First, the most recent version of the data may be misleading because of bad methods used in its construction. For example, whenever the base year was changed for the U.S. national income and product accounts under fixed weighting, growth rates for real output and its components were revised for previous

years, back to 1947. The changing weights caused substitution bias for periods far from the base year and led to poor measures of real output and its components. That problem was finally fixed by the move to chain weighting in 1996. A researcher studying the economic events or policy decisions of the 1970s and 1980s might have been better off using a data set from a vintage in the 1980s rather than one in the first half of the 1990s, which suffered from greater substitution bias.

Second, under chain weighting, the components of real output do not sum up to real output. So some economic concepts, such as the ratio of real consumption to real output, which may be the subject of theory, do not have an obvious empirical counterpart in vintages after 1996. Or, for example, a theory might argue that the levels of real consumption and real output are cointegrated. But that theory can no longer be empirically verified because under chain weighting, the levels of those variables have no intrinsic meaning. This should make macroeconomists rethink some theories—or at least think harder about whether the available data map cleanly to their theoretical counterparts. The tension arises here because the theory was developed for a world with just one good and did not concern itself with aggregation issues and the effects of changes in relative prices. If the outcome of testing an economic hypothesis on data sets of different vintages leads to conflicting results, the researcher may wish to consider which data set is the most appropriate and track down the source of the conflict, though doing so is not easy, as our work in this paper illustrates.

Third, economists such as Dewald, Thursby, and Anderson (1986) have found that many empirical research papers could not be replicated. One reason may be that the empirical results were fragile in the first place, perhaps because they were the result of a specification search or were based on the special properties of some data sample. So, even if the most recent data set

were indeed the best, a researcher might still prosper by asking the question: "What results would I have obtained if I had run this same experiment five years ago?" If the results would have been the same qualitatively and nearly the same quantitatively, the researcher can be assured that the results are not special to one data set.

We proceed as follows. In Section II, we discuss the data set, look at the spectral properties of revisions to selected macroeconomic variables, and examine the degree of news and noise in several variables. In Section III, we look at some key empirical papers in macroeconomics and explore the degree to which the data vintage matters for their results. We draw conclusions from these results in Section IV.

II. ANALYZING DATA REVISIONS

The Data Set

In a lengthy process over the past nine years, we have developed our real-time data set. It consists of a series of vintages of data, each corresponding to an economist's information set on the date of the vintage. For example, the February 1977 vintage of data contains information on real output and all its components, as well as other macroeconomic variables, just as an economist would have viewed the data on February 15, 1977. There is one of these data sets for each quarter, beginning in November 1965, each containing information that was available on the 15th day of the middle month of the quarter.

Data in each vintage include nominal and real GNP (GDP after 1991); the components of real GNP/GDP, including total personal consumption expenditures, broken down into durables, nondurables, and services; business fixed investment; residential investment; the change in

business inventories; government purchases (government consumption and government investment since 1996); exports and imports; the chain-weighted GDP price index (since 1996); after-tax corporate profits; the M1 and M2 measures of the money supply; total reserves at banks (adjusted for changes in reserve requirements); nonborrowed reserves; nonborrowed reserves plus extended credit; the adjusted monetary base (measures of reserves and the monetary base are from the Federal Reserve Board, not the St. Louis Fed); the civilian unemployment rate; the consumer price index (CPI-U); an index of import prices; the three-month T-bill interest rate; and the 10-year Treasury bond interest rate. The interest rates are included for completeness, even though they are never revised. The vintages are mostly complete; there are some missing data for the money stock, monetary base, and reserves variables. For additional descriptive information about the construction of the real-time data, see Croushore-Stark (2001).[1]

Spectral Analysis of Data Revisions

How big are the revisions to the data? Our previous paper, Croushore-Stark (2001), describes many of the revisions, showing how large those revisions are for both short and long horizons. In this paper, we use a different method, spectral analysis, to investigate the size and importance of data revisions.

First, we pull out from the data set vintages dated November 1975, 1980, 1985, 1991, 1995, and 1998. The first five of these vintages were chosen because they were the last vintages

[1]For complete notes on all the variables and any missing data, see the documentation files on our web page: www.phil.frb.org/econ/forecast/reaindex.html.

prior to a comprehensive revision of the national income and product accounts; the last vintage, November 1998, was the latest available data at the time this article was first written. For ease of exposition, we call these benchmark vintages. Each of the comprehensive revisions that were made after our benchmark vintage dates incorporated major changes to the data, including new source data (information-based revisions) and definitional changes (structural revisions). In addition, the base year was changed for real variables in January 1976 (from 1958 to 1972), in December 1985 (from 1972 to 1982), in late November 1991 (from 1982 to 1987), and in January 1996 (from 1987 to 1992), so some of the differences across the benchmark vintages we look at incorporate base-year changes, which affect real variables. In particular, since the base-year changes in 1976, 1985, and 1991 used the old fixed-weighted index methodology, the change of base year alters the timing of substitution bias; this bias is large for dates further away from the base year. The switch to chain weighting in 1996 means that a change of base year (which is arbitrary under chain weighting) will have no effect on the growth rates of variables, whereas the growth rates changed significantly under the old fixed-weighting method. However, as a consequence, the levels of chain-weighted real variables have no intrinsic meaning—they are simply index numbers.

We use spectral analysis to analyze data revisions.[2] The idea is to make a transformation into the frequency domain, allowing us to look at the spectrum to see where the main action is in the revisions. If the revisions are white noise, the spectrum will be flat. But spectra with peaks at different frequencies show that the revisions are not white noise but follow patterns at the

[2] The present analysis is in the spirit of Sargent (1987, pp. 346-8), who showed that inferences drawn from VAR coefficients can be susceptible to measurement errors in the underlying data.

given frequencies. To estimate the population spectrum, we use nonparametric (kernel) methods described by Hamilton (1994, pp. 165-7).[3] We show figures just for real consumption, though the spectral estimates for other real variables in the national income and product accounts are similar.

We begin by estimating the spectrum of the logarithm of the ratio of real consumption across benchmark revisions (Figure 1), using the following naming conventions. The labels on each plot follow the structure LC#, where L means the logarithm of the variable, C means real consumption, and where # represents the benchmark vintage, with # = 1 for the November 1975 vintage, # = 2 for 1980, # = 3 for 1985, # = 4 for 1991, # = 5 for 1995, and # = 6 for 1998.

The estimates in Figure 1 show that the revisions to real consumption exhibit the typical spectral shape of macroeconomic data (Sargent, 1987, pp. 279-83), indicating that most of the power resides at low frequencies.

More interesting are the spectra of the revisions to quarterly growth rates (labeled DLC#), as shown in Figure 2. In some cases there is action at business-cycle frequencies (frequencies between 0.2 and 0.8 correspond to business cycles, with periodicity ranging from roughly eight years for a frequency of 0.2, to two years for a frequency of 0.8), as in the upper right-hand graph (reflecting the revision from benchmark vintage November 1995 to November 1998). In other cases, most of the differences are seasonal, as in the lower left graph, at a frequency of 1.5, which corresponds to a periodicity of four quarters.

[3] In particular, we are using a kernel estimate with a tent-shaped window of width 9.

It is of some interest to examine the relationship of revisions across variables. In the frequency domain, this can be done by examining the squared coherences of the revisions. We show such coherences for real output growth revisions (DLY#) and real consumption growth revisions (DLC#) in Figure 3. In most of the graphs, the coherence is high at business-cycle frequencies, but note that each different set of benchmark vintages seems to have slightly different patterns of coherence, perhaps because of the influence of definitional changes or particular changes in relative prices on the consumption component of output.

All these differences across vintages point to the fact that the benchmark revisions to the data can be characterized neither as white noise nor as a particular ARIMA process, perhaps because benchmark revisions incorporate both information-based revisions and structural revisions. Is it possible to characterize these revisions in terms of their information content? To do so, we examine tests for the news and noise content of revisions.

Tests for News and Noise in Data Revisions

We investigate the information content of data revisions by testing to see if the revisions can be characterized as containing news or reducing noise, as suggested by Mankiw, Runkle, and Shapiro (1984) and Mankiw and Shapiro (1986). The idea is that if the revisions are characterized as containing news, subsequent releases of the data for that date contain new information that was not available in the earlier releases. As a result, the advance release is an efficient estimate of later data. This implies that the revision to the data is correlated with the revised data but not with the earlier data. It also implies that the variance of the data should increase as we look at later and later vintages, since an optimal forecast is smoother than the data.

On the other hand, if data are characterized as reducing noise, subsequent releases of the data just eliminate noise in the earlier release, so the earlier release is the true value plus measurement error that gets reduced over time. In this case, the revision should be uncorrelated with the revised data, but correlated with the advance data. In addition, the variance of the data should decline as it is further revised.

Characterizing the revisions as news or noise may help us model the revision process, which may be useful in estimating economic models. Some researchers, such as Rudebusch (2001) assume that data revisions reduce noise, while others, such as Koenig, Dolmas, and Piger (2001), assume that data revisions contain news.

To formalize this, we use the following notation. Let $X(t, s)$ represent the data for date t as of vintage s. Then a revision of the data from vintage i to vintage j (where $j > i$) is defined as $e(t, i, j) \equiv X(t, j) - X(t, i)$. For example, e(1993Q4, Feb. 1994, Feb. 1995) $\equiv X$(1993Q4, Feb. 1995) $- X$(1993Q4, Feb. 1994). To say that a revision is characterized as containing news means that the revision is uncorrelated (orthogonal) to earlier vintage data, so that $e(t, i, j) \perp X(t, i)$. To say that a revision is characterized as reducing noise means that the revision is uncorrelated with later vintage data, so that $e(t, i, j) \perp X(t, j)$.

We begin by looking at four different data sets, each consisting of quarterly growth rates of a variable. One data set (labeled *initial*) consists of the growth rate for each date from the first vintage in our data set in which the variable appears for that date. For all variables that we analyze here, this occurs in the quarter after the observation period, that is, data for the fourth

quarter of 1968 are first reported in our data vintage from the first quarter of 1969.[4] So, these observations are described in our notation as $X(t, t+1)$, where $t+1$ refers to the vintage 1 quarter after date t. The second data set we use, labeled the *1-year-later* estimate, consists of the growth rate for a quarter based on a data set with a vintage one year after the initial vintage, or five quarters after date t, $X(t, t+5)$; the third (*3-years-later* estimate) is based on a vintage three years after the initial vintage or 13 quarters after date t, $X(t, t+13)$. The fourth data set (*latest*) consists of the November 1998 vintage of data, $X(t,$ Nov. 1998). From these data sets, we construct the corresponding *revisions* to the data from the initial release to 1 year later, $e(t, t+1, t+5)$; from 1 year to 3 years later, $e(t, t+5, t+13)$; and from 3 years later to the latest data, $e(t, t+13,$ Nov. 1998).

Are the revisions to the data best characterized as containing news or reducing noise? To find out, we run tests like those of Mankiw and Shapiro. First, we examine the standard deviation of quarter-to-quarter real growth rates from the four different data sets in Table 1. If the revisions contain news, the standard deviation should increase from initial, to 1-year, to 3-year, to latest data sets; if the revisions reduce noise, the standard deviation should decline as we move down the rows from initial to latest.

As Table 1 shows, for real consumption, the standard deviation rises from initial to 1 year, then falls in each successive revision. So, the initial to 1-year revision contains news, while the 1-year to 3-year and 3-year to latest revisions reduce noise. The same pattern is true for real

[4] For variables other than real and nominal output, for some observations over the period 1965Q3 to 1969Q3, our initial observations are taken from the preliminary report of the national income and product accounts, not the advance report, because of reporting delays. These observations do not appear in our real-time data set but are available on request.

business fixed investment and real residential fixed investment. For real output, consumption of durables, and consumption of services, the standard deviation rises from initial to 1-year later, falls from 1-year later to 3-years later, then rises from 3-years later to latest. Again, some revisions contain news, others reduce noise. Only nominal output shows a consistent pattern, with the standard deviation rising for each subsequent data set, which suggests that the revisions to nominal output contain news and do not reduce noise. Also, note that the standard deviation rises between initial release and 1-year later, for all the variables, which suggests that the first year's worth of revisions contains news. Later revisions appear to be a mix of adding news and reducing noise, for variables other than nominal output.

Next, we examine the correlation between the revisions and the quarter-to-quarter growth rates in Table 2.[5] Consistent with the standard deviations for real consumption, only the initial to 1-year revision can be characterized as containing news because it is correlated with later data and uncorrelated with earlier data. The other five revisions can be characterized as reducing noise because they are correlated with some earlier data and uncorrelated with later data. Overall, one could argue that revisions in the first year to the initial consumption data contain news and that subsequent revisions reduce noise. The same is true for all the other variables examined here, except for nominal output. In each case, the initial to 1-year later revision contains news, but many of the other revisions reduce noise, or, in same cases, may both reduce noise and contain news. The one exception is nominal output, which shows no signs of reducing

[5]The t-statistics reported in parentheses in Table 2 are based on the method of Newey and West (1987) to correct for heteroskedasticity and serial correlation, using a truncation lag equal to the longest lag for which there is a significant correlation (which ranges from 0 to 8 across variables).

noise—the revisions appear only to contain news, consistent with the pattern in Table 1. Thus, it appears that the process of gathering nominal data is one that adds news.

The results of the news-noise tests are that: (1) revisions between the initial release and one-year later can be characterized as containing news, a result that is consistent with the notion that these are information-based revisions; and (2) revisions after 1 year, including revisions to create final data, can not be easily characterized.

Both the spectral analysis and the news-noise tests suggest that revisions to the data can be hard to characterize, particularly because of structural revisions that occur across benchmark vintages. In many cases, this should be expected because different types of structural changes are made in each benchmark revision. This, of course, makes it extremely hard to predict how a benchmark revision will affect published results. We have learned a bit about the process governing the revisions, but not enough to make generalizations about them in terms of an ARIMA model that could be used as a data-generating process, at least for real variables after the first year of revisions. With that in mind, we now turn to examining the extent to which such structural revisions matter for macroeconomic studies.

III. DOES DATA VINTAGE MATTER FOR KEY MACROECONOMIC RESULTS?

It is clear that the vintage of the data makes a difference for growth rates in different periods, but does it matter for empirical work? We now take several empirical exercises from the economics literature, rerun them with differing vintages of data, and see how much the vintage matters. We examine empirical work by Kydland and Prescott (1990), Hall (1978), and Blanchard and Quah (1989).

Kydland and Prescott (1990)

Kydland and Prescott examine the correlation of real GNP with lags and leads of itself and other variables. They filter the data with an HP filter, then calculate the cross correlations. They use data from a 1990 vintage; we compare our results for data vintages from February 1990, February 1994, and February 1998 to their results (Table 3) for output autocorrelations and cross-correlations between real GNP and the price deflator, real consumption, and M2. As the table shows, although there are some quantitative differences, the qualitative pattern is quite similar across all the vintages. There are a few exceptions—for example, the contemporaneous correlation between real output and the deflator varies across vintages from –0.48 to –0.66, while the contemporaneous correlation between real output and real consumption varies from 0.82 to 0.88. Other examples of notable differences include the correlation between real output and the fifth lag of the deflator, which varies across vintages from –0.35 to –0.51, and the correlation between real output and the fifth lag of real consumption, which varies from 0.12 to 0.24.

The HP-filtered cyclical data on real output from the three vintages do not change much across vintages; the biggest differences across vintages are on the order of one percentage point and occur only in the 1950s. Trend real output growth also behaves similarly across vintages, though the four-quarter average of trend output growth can differ as much as 0.5 percentage point at times, as shown in Figure 4. Part of the differences across vintages for real output could be attributable to the switch between GNP and GDP that occurred between the 1990 and 1994 vintages. So it is useful to also examine other variables, for which the revision pattern may be different. For example, real consumption (not shown) was not revised as much as real output between the 1990 and 1994 vintages.

Altogether, however, since the purpose of Kydland and Prescott's research was to establish general business-cycle facts, it is hard to conclude that the data vintage matters. The general business-cycle facts are not very sensitive to data vintage. This result may not be surprising since our spectral analysis in part II suggested that most revisions to the log levels of macroeconomic variables occur at low frequencies, which are filtered out by the HP filter used in Kydland and Prescott's analysis.

Hall (1978)

Hall found evidence supporting the life-cycle/permanent-income hypothesis using data on U.S. consumption spending. Although Hall's results have been challenged and modified in a variety of ways, in such papers as those by Flavin (1981) and Deaton (1987), an even more fundamental question is: are Hall's empirical results robust to different data sets? That is, would we get significantly different outcomes depending on what vintage of data we used?

Hall's original data set included observations on consumption from 1948Q1 to 1977Q1, so we assume that he had data of vintage May 1977. Hall begins by testing to see if consumption can be predicted from its own past values. Under the pure life-cycle/permanent-income hypothesis, only the first lagged value of consumption should help predict current consumption. Hall regresses consumption on four lags of consumption, testing to see if the last three lags are jointly zero.[6] His original result is shown in the first line of Table 4. In the table, the coefficient estimates are given, with standard errors in parentheses. The column labeled s shows the standard error of estimate; DW is the Durbin-Watson statistic; and F is the value of the F-statistic testing the hypothesis that the coefficients on the second, third, and fourth lags of consumption

[6] The variable used is real consumption of nondurables and services divided by the population.

are jointly zero, with the p-value for the test shown in parentheses. The F-test shows that you cannot reject the hypothesis at the 5 percent level.

Using our real-time data set with consumption data from the May 1977 vintage, we are able to replicate Hall's results fairly closely, as the second line of the table shows. Our replication confirms Hall's finding that the coefficients on the second, third, and fourth lagged terms are jointly zero.

However, when we rerun the test on the same sample period (1948Q1 to 1977Q1) using vintage data from February 1998, the coefficients change dramatically, and the F-test now rejects the hypothesis that the second-through-fourth lagged consumption terms are jointly zero. The p-value for the test is only .02, so we reject the hypothesis at the 5 percent level.

Further, when we update the sample to include data through 1997, we reject the hypothesis even more convincingly. Again, the coefficient estimates change dramatically, and the F-statistic rises to 8.1, with a p-value of less than 0.005.

Further investigation shows that, beginning with Hall's vintage data, as we use data from later and later vintages, the p-value of the F-test declines (not changing the sample dates, just using later vintages of data). But the p-value remains above .05 until the shift to chain weighting occurs. What should we make of the lack of robustness under chain weighting? It may be that chain weighting makes a fundamental change in the statistical properties of the consumption data that leads us to reject the hypothesis. But this lack of robustness needs then to be handled by theory as well as empirical work. In particular, theory has not dealt with the issue of how best to deal with changes in relative prices, which government data agencies must handle. Perhaps they

are doing so in a manner that is inconsistent with theory, but that was also true before chain weighting was used.

This result again raises the issue of the best vintage of the data to use in empirical applications. There may be a reason to think that chain weighting is not ideal for this application, since chain weighting is based on annual weights, which give the data the feature of a two-sided filter. As a result of this filter, Hall's orthogonality conditions may fail to hold. However, this was also true under fixed weighting, because seasonal factors are also based on a two-sided filter. Hence, the "best" vintage to use in testing Hall's theory is not at all clear.

These results mean that Hall's original hypothesis—that only the first lag of consumption matters in determining contemporaneous consumption—is not well supported by the data. Hall's test was legitimate, but his empirical result does not stand the test of time, either in terms of revisions to the data or in terms of additional data.

Blanchard and Quah (1989)

Blanchard and Quah use a structural VAR in output and unemployment to define supply disturbances as shocks that have a permanent effect on output, and demand disturbances as shocks that have a temporary effect on output. They examine U.S. data from 1950 to 1987, calculating impulse responses and variance decompositions based on a VAR model in output and unemployment. We examine how changes in the vintage of the data affect the decomposition of shocks into supply disturbances versus demand disturbances, how the impulse responses change across data vintages, and how the cumulative effects of demand and supply shocks vary with the data vintage.

We compare Blanchard and Quah's results to ours using the February 1988 version of our data set, then comparing those results in turn to our November 1993 data set and our February 1998 data set. First, using our February 1988 data set, we are able to replicate the results of Blanchard and Quah fairly precisely. The impulse responses to supply and demand shocks (not shown) are quite similar to those found by Blanchard and Quah, both qualitatively and quantitatively.[7]

When we look at the decomposition of shocks into demand and supply shocks for the three different vintages of the data (Figure 5), we notice there are substantial differences across data vintages. The differences are particularly noticeable for demand shocks, as many of the local peaks and troughs are largest in magnitude when using the 1988 vintage data and smallest in magnitude when using the 1998 vintage data. However, demand shocks are temporary, so these differences in magnitude do not matter as much for the cumulative effect of the shocks. But even the fairly small differences across vintages in the measured supply shocks may have a large impact on the cumulative effect on output and unemployment.

The other way in which the method of Blanchard and Quah is often used is to establish stylized facts about how economic variables respond to shocks. These are generally shown in figures that illustrate the impulse responses to a shock. Using the Blanchard and Quah method,

[7] To measure the unemployment rate, Blanchard and Quah use the seasonally adjusted rate for males, age 20 and over. Because this rate does not appear in our data set, we substitute the total civilian rate of unemployment for the Blanchard/Quah measure. On the basis of our replication using the February '88 vintage, this substitution has little effect on the results.

and the same three vintages of data used above, we calculate the impulse responses for demand and supply shocks (Figure 6). Note that the impulse responses show the same general shape across vintages, but the magnitudes are very sensitive to vintage, especially for demand shocks. The response of output or unemployment to a demand shock is sometimes as much as five times as large, using 1998 data, than when using 1988 data. So the vintage of the data set seems to matter quite significantly for impulse responses. Why this is so is difficult to determine, but the estimated variance-covariance matrix shows a much different variance of the structural shocks, along with a substantially different parameter estimate of the coefficient on output in the unemployment equation. This occurs despite the fact that differences in the data do not seem large. This suggests that there may be something about the procedure for estimating a structural VAR that makes it very sensitive to small changes in the data.

Can we be more precise? As noted above, in examining the estimated coefficients of the structural VAR representation, we notice particularly large differences in the estimated coefficient on contemporaneous output growth in the structural unemployment equation as we move from vintages February 1988 and November 1993 to February 1998. The coefficient estimate is 4.62 in the February 1988 data, 2.45 in the November 1993 data, and 0.63 in the February 1998 data, with output growth measured in log first differences and the unemployment rate expressed as a percent, rather than in percentage points.

In a recent paper, Sarte (1997) shows that standard structural VAR instrumental variables (IV) techniques—which use structural shock estimates as instruments—can fail over certain ranges of the parameter space. The key condition for such a failure is a low pairwise correlation between the instrument/structural shock and the variable instrumented. In estimating the model,

we employ the standard IV approach and use the estimated structural shock attached to the output equation as an instrument for contemporaneous output growth in the unemployment equation. We then checked Sarte's key condition for IV failure by computing for each vintage the correlation coefficient between the output-equation structural shock and output growth. For vintages February 1988 and November 1993, those correlations border on zero: 0.04 and 0.08, respectively. Such low correlations call into question the usefulness of structural shocks as instruments and, by implication, the just-identified structural VAR methodology. Indeed, a reasonable conclusion is that the SVAR is unidentified empirically in the first two vintages. In contrast, the pairwise correlation in the February 1998 data rises significantly, to 0.23, suggesting a higher possibility that the model is identified empirically.

We view these results as an extension of Sarte's. Sarte showed that alternative identification schemes, holding constant the data vintage, may fail empirically. Our results indicate that a given identification scheme may fail empirically in some vintages but not in others. On the basis of these results, structural VAR users may wish to check their results for robustness along the lines suggested by Sarte and across different vintages of data.

IV. CONCLUSIONS

This paper reports on the nature of data revisions and on how such revisions can lead to somewhat different results for major studies in macroeconomics. It is somewhat reassuring that for many of the studies we examined, including some not discussed in this version of the paper, the results are generally robust, at least qualitatively, for different vintages of the data. But in some cases, the empirical results are quite sensitive to the exact vintage of the data.

What can we conclude from these results? In practice, economists run thousands of empirical exercises each day, some of which get reported in academic journals and influence economists' thoughts about the structure of the economy. Our exercise is really one in the spirit of checking such results for robustness and can thus be used to confirm some results in the literature, such as those of Kydland and Prescott. But when empirical results are sensitive to the vintage of the data, economists should be more cautious about accepting those particular results or perhaps about accepting the empirical methods that led to those results. If an empirical method is robust to data vintage, as in the case of Kydland and Prescott, an empirical researcher can have more confidence that the method itself is sound and not overly sensitive to variations in the data. But if the empirical method is one that leads to very different results for variations in the data, a researcher should be skeptical. Or, certainly, further research is needed to establish the validity of the research method.

A researcher might argue that empirical work should be based only on the most recent data available to the researcher. But we would argue that without checking the empirical results against alternative vintages, researchers cannot be sure that their results are not simply a special case that applies to a particular data set. Is a research result obtained only for national income and product account data that count software as investment (year 2000 and after), or is it independent of the accounting treatment for software? Is an empirical result dependent on chain-weighting rather than fixed-weighting of the national income and product account data? Would empirical outcomes be changed if the base year for real variables was different? We argue that if empirical results do not hold up across alternative vintages of the data, then those results are of limited value. A true empirical regularity should be evident in data that are constructed

somewhat differently, yet measure the same basic economic concept. Thus we would argue that not only should researchers investigate older research results using newer data, but they should also investigate newer research results using older data.

Table 1
Standard Deviations of Growth Rates
(In Percentage Points)

1965Q3 to 1995Q3

Data Set	C	Y	PY	BFI	RFI	C-D	C-S
Initial Release	3.40	3.56	3.63	9.20	22.11	14.34	1.43
1-Year Later	3.57	3.79	3.98	10.32	23.45	14.73	1.79
3-Years Later	3.17	3.76	4.07	10.04	23.08	13.55	1.74
Latest	3.10	3.89	4.32	9.75	22.41	14.13	1.88

Variables:
C = real consumption
Y = real output
PY = nominal output
BFI = real business fixed investment
RFI = real residential fixed investment
C-D = real consumption spending on durables
C-S = real consumption spending on services

Table 2
Correlations of Revisions with Growth Rates
1965Q3 to 1995Q3

A. Real Consumption Data

Revisions/Data Set	Initial	1-Year	3-Year	Final
Initial to 1-Year	-0.02	0.30*	0.27*	0.20
	(0.13)	(2.23)	(2.21)	(1.80)
Initial to 3-Year	-0.37†	-0.17	0.08	0.04
	(3.91)	(1.49)	(0.58)	(0.29)
Initial to Final	-0.42†	-0.28?	-0.08	0.10
	(4.68)	(3.01)	(0.78)	(0.86)
1-Year to 3-Year	-0.44†	-0.48†	-0.15	-0.14
	(5.52)	(5.58)	(1.61)	(1.44)
1-Year to Final	-0.43†	-0.50†	-0.26?	-0.04
	(5.42)	(5.65)	(3.04)	(0.38)
3-Year to Final	-0.16	-0.22†	-0.23†	0.11
	(1.96)	(2.37)	(2.32)	(1.32)

Absolute values of adjusted t-statistics are in parentheses below each correlation coefficient.
An asterisk (*) means there is a significant (at the 5% level) correlation between the revision and the later data, implying "news."
A dagger (†) means there is a significant (at the 5% level) correlation between the revision and the earlier data, implying "noise."
A question mark (?) means there is a significant correlation that does not fit easily into the news/noise dichotomy.

B. Real Output Data

Revisions/Data Set	Initial	1-Year	3-Year	Final
Initial to 1-Year	0.02	0.35*	0.37*	0.32*
	(0.18)	(3.67)	(3.97)	(4.36)
Initial to 3-Year	-0.13	0.13	0.34*	0.24*
	(1.57)	(1.45)	(2.86)	(3.14)
Initial to Final	-0.16	0.02	0.14	0.43*
	(1.15)	(0.17)	(0.88)	(3.01)
1-Year to 3-Year	-0.22†	-0.18†	0.12	0.02
	(2.60)	(2.10)	(1.34)	(0.21)
1-Year to Final	-0.20	-0.20	-0.08	0.30*
	(1.74)	(1.61)	(0.61)	(2.13)
3-Year to Final	-0.07	-0.10	-0.16	0.31*
	(0.64)	(0.81)	(1.91)	(2.33)

C. Nominal Output Data

Revisions/Data Set	Initial	1-Year	3-Year	Final
Initial to 1-Year	0.10	0.42*	0.43*	0.32*
	(1.00)	(7.01)	(6.70)	(5.53)
Initial to 3-Year	0.01	0.26?	0.45*	0.32*
	(0.12)	(2.70)	(4.93)	(3.54)
Initial to Final	0.08	0.24	0.36?	0.54*
	(0.52)	(1.78)	(3.51)	(7.69)
1-Year to 3-Year	-0.10	-0.07	0.22*	0.13
	(1.06)	(0.71)	(2.74)	(1.59)
1-Year to Final	0.02	-0.03	0.09	0.39*
	(0.14)	(0.24)	(0.74)	(4.24)
3-Year to Final	0.09	0.01	-0.05	0.34*
	(0.76)	(0.07)	(0.37)	(3.77)

D. Real Business Fixed Investment Data

Revisions/Data Set	Initial	1-Year	3-Year	Final
Initial to 1-Year	-0.03	0.45*	0.35*	0.38*
	(0.26)	(6.26)	(4.29)	(4.95)
Initial to 3-Year	-0.14	0.20?	0.42*	0.33*
	(1.40)	(2.23)	(4.92)	(4.10)
Initial to Final	-0.25†	0.08	0.17	0.40*
	(2.37)	(0.79)	(1.66)	(3.70)
1-Year to 3-Year	-0.15	-0.27†	0.15	-0.01
	(1.55)	(2.55)	(1.71)	(0.06)
1-Year to Final	-0.29†	-0.35†	-0.13	0.14
	(2.86)	(3.17)	(1.17)	(1.14)
3-Year to Final	-0.18	-0.14	-0.29†	0.16
	(1.83)	(1.37)	(3.03)	(1.41)

E. Real Residential Fixed Investment Data

Revisions/Data Set	Initial	1-Year	3-Year	Final
Initial to 1-Year	-0.04	0.33*	0.28*	0.16*
	(0.48)	(3.36)	(3.31)	(1.98)
Initial to 3-Year	-0.17	0.07	0.33*	0.20*
	(1.95)	(0.70)	(2.96)	(2.04)
Initial to Final	-0.23†	-0.08	0.13	0.28*
	(2.53)	(0.77)	(1.36)	(2.76)
1-Year to 3-Year	-0.18†	-0.23†	0.15	0.09
	(2.54)	(2.26)	(1.61)	(1.12)
1-Year to Final	-0.21†	-0.33†	-0.08	0.16
	(2.05)	(3.39)	(0.85)	(1.84)
3-Year to Final	-0.08	-0.20†	-0.27†	0.12
	(0.74)	(2.04)	(2.54)	(1.45)

F. Real Durable Consumption Data

Revisions/Data Set	Initial	1-Year	3-Year	Final
Initial to 1-Year	-0.01	0.23*	0.19*	0.16
	(0.06)	(2.78)	(2.22)	(1.76)
Initial to 3-Year	-0.33†	-0.19	0.00	-0.05
	(3.15)	(1.75)	(0.01)	(0.51)
Initial to Final	-0.22†	-0.11	0.04	0.14
	(3.09)	(1.49)	(0.58)	(1.75)
1-Year to 3-Year	-0.38†	-0.42†	-0.17	-0.20*
	(3.81)	(3.85)	(1.80)	(2.12)
1-Year to Final	-0.23†	-0.28†	-0.10	0.03
	(3.26)	(4.00)	(1.21)	(0.37)
3-Year to Final	0.13	0.10	0.06	0.29*
	(1.05)	(0.80)	(0.47)	(2.14)

G. Real Services Consumption Data

Revisions/Data Set	Initial	1-Year	3-Year	Final
Initial to 1-Year	-0.12	0.61*	0.47*	0.30*
	(1.51)	(8.06)	(5.94)	(2.87)
Initial to 3-Year	-0.32†	0.19	0.63*	0.38*
	(3.61)	(1.94)	(9.53)	(3.03)
Initial to Final	-0.29†	0.08	0.39?	0.68*
	(2.44)	(0.78)	(4.19)	(12.9)
1-Year to 3-Year	-0.28†	-0.38†	0.31*	0.18
	(3.15)	(3.97)	(4.14)	(1.76)
1-Year to Final	-0.20	-0.39†	0.04	0.48*
	(1.81)	(4.37)	(0.51)	(6.49)
3-Year to Final	0.02	-0.13	-0.25†	0.44*
	(0.13)	(1.28)	(2.51)	(3.65)

Table 3
Kydland-Prescott Cross-Correlations

		Cross Correlation of Real GNP/GDP With										
Vintage	**Variable x**	**x(t-5)**	**x(t-4)**	**x(t-3)**	**x(t-2)**	**x(t-1)**	**x(t)**	**x(t+1)**	**x(t+2)**	**x(t+3)**	**x(t+4)**	**x(t+5)**
KP 1990	Real GNP/GDP	-0.03	0.15	0.38	0.63	0.85						
Feb. 1990		-0.03	0.15	0.37	0.62	0.85						
Feb. 1994		-0.02	0.15	0.36	0.61	0.84						
Feb. 1998		-0.09	0.11	0.34	0.60	0.84						
KP 1990	GNP/GDP deflator	-0.50	-0.61	-0.68	-0.69	-0.64	-0.55	-0.43	-0.31	-0.17	-0.04	0.09
Feb. 1990		-0.49	-0.60	-0.67	-0.69	-0.64	-0.56	-0.43	-0.31	-0.18	-0.05	0.08
Feb. 1994		-0.51	-0.60	-0.66	-0.66	-0.59	-0.48	-0.36	-0.26	-0.15	-0.05	0.07
Feb. 1998		-0.35	-0.49	-0.60	-0.68	-0.70	-0.66	-0.55	-0.40	-0.22	-0.04	0.12
KP 1990	Real Consumption	0.25	0.41	0.56	0.71	0.81	0.82	0.66	0.45	0.21	-0.02	-0.21
Feb. 1990		0.24	0.40	0.55	0.70	0.80	0.82	0.65	0.44	0.21	-0.02	-0.21
Feb. 1994		0.18	0.35	0.53	0.71	0.84	0.87	0.70	0.48	0.25	0.02	-0.17
Feb. 1998		0.12	0.31	0.50	0.69	0.84	0.88	0.71	0.48	0.23	-0.02	-0.20
KP 1990	M2	0.48	0.60	0.67	0.68	0.61	0.46	0.26	0.05	-0.15	-0.33	-0.46
Feb. 1990		0.46	0.57	0.64	0.66	0.60	0.46	0.25	0.05	-0.14	-0.31	-0.42
Feb. 1994		0.44	0.57	0.65	0.69	0.64	0.50	0.29	0.08	-0.10	-0.28	-0.41
Feb. 1998		0.44	0.58	0.66	0.69	0.63	0.48	0.27	0.07	-0.12	-0.28	-0.40

Table 4
Hall's Tests on Consumption

Regression Equation: $c_t = \hat{a}_0 + \hat{a}_1 c_{t-1} + \hat{a}_2 c_{t-2} + \hat{a}_3 c_{t-3} + \hat{a}_4 c_{t-4} + e_t$

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	R^2	s	DW	F
Sample 1948Q1 to 1977Q1									
Hall's results	8.2	1.130	-0.040	0.030	-0.113	.9988	14.5	1.96	1.7
	(8.3)	(0.092)	(0.142)	(0.142)	(0.093)				(0.17)
Replication	-8.122	1.130	-0.024	-0.004	-0.095	.9988	14.7	1.97	1.7
vintage May 1977	(8.489)	(0.092)	(0.142)	(0.143)	(0.094)				(0.17)
Replication	-9.859	1.102	0.166	-0.256	-0.007	.9988	57.5	2.00	3.5
vintage Feb. 1998	(27.498)	(0.093)	(0.138)	(0.137)	(0.094)				(0.02)
Sample 1948Q1 to 1997Q4									
Vintage Feb. 1998	15.589	1.153	0.163	-0.011	-0.157	.9997	57.0	1.97	8.1
	(14.296)	(0.070)	(0.108)	(0.108)	(0.070)				(0.00)

REFERENCES

Amato, Jeffery D., and Norman R. Swanson. "The Real Time Predictive Content of Money for Output," *Journal of Monetary Economics* 48 (2001), pp. 3–24.

Bernanke, Ben S., and Jean Boivin. "Monetary Policy in a Data-Rich Environment," manuscript, 2001.

Blanchard, Olivier Jean, and Danny Quah. "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review* 79 (September 1989), pp. 655-73.

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 (November 2001), pp. 111–30.

Deaton, Angus S. "Life-Cycle Models of Consumption: Is the Evidence Consistent with the Theory?" in T.F. Bewley, ed., *Advances in Econometrics: Fifth World Congress, Vol. II* (Cambridge: Cambridge University Press, 1987), pp. 121-48.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. "Replication in Empirical Economics: The *Journal of Money, Credit, and Banking* Project," *American Economic Review* 76 (September 1986), pp. 587–603.

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.

Flavin, Marjorie A. "The Adjustment of Consumption to Changing Expectations about Future Income," *Journal of Political Economy* 89 (1981), pp. 974-1009.

Ghysels, Eric, Norman R. Swanson, and Myles Callan. "Monetary Policy Rules and Data Uncertainty," Manuscript, Texas A&M University, 1999.

Hall, Robert E. "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy* 86 (December 1978), pp. 971-87.

Hamilton, James D. *Time Series Analysis.* Princeton: Princeton University Press, 1994.

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

Koenig, Evan, Sheila Dolmas, and Jeremy Piger. "The Use and Abuse of 'Real-Time' Data in Economic Forecasting," Federal Reserve Bank of Dallas working paper 2001.

Kydland, Finn E., and Edward C. Prescott. "Business Cycles: Real Facts and a Monetary Myth," Federal Reserve Bank of Minneapolis *Quarterly Review* (Spring 1990), pp. 3-18.

Mankiw, N. Gregory, David E. Runkle, and Matthew D. Shapiro. "Are Preliminary Announcements of the Money Stock Rational Forecasts?" *Journal of Monetary Economics* 14 (July 1984), pp. 15-27.

Mankiw, N. Gregory, and Matthew D. Shapiro. "News or Noise: An Analysis of GNP Revisions," *Survey of Current Business* (May 1986), pp. 20-5.

Maravall, Augustin, and David A. Pierce. "The Transmission of Data Noise into Policy Noise in U.S. Monetary Control," *Econometrica* 54 (July 1986), pp. 961-79.

Mork, Knut A. "Ain't Behavin': Forecast Errors and Measurement Errors in Early GNP Estimates," *Journal of Business and Economic Statistics* 5 (April 1987), pp. 165-75.

Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55 (May 1987), pp. 703-8.

Orphanides, Athanasios. "Monetary Rules Based on Real-Time Data," *American Economic Review* 94 (September 2001), pp. 964–85.

Rudebusch, Glenn D. "Is the Fed Too Timid? Monetary Policy in an Uncertain World," *Review of Economics and Statistics* 83 (May 2001), pp. 203–17.

Runkle, David E. "Revisionist History: How Data Revisions Distort Economic Policy Research," Federal Reserve Bank of Minneapolis *Quarterly Review* (Fall 1998), pp. 3-12.

Sargent, Thomas J. *Macroeconomic Theory*, 2nd ed. Boston: Academic Press, 1987.

Sarte, Pierre-Daniel G. "On the Identification of Structural Vector Autoregressions," Federal Reserve Bank of Richmond *Economic Quarterly* (Summer 1997), pp. 45-67.

Stark, Tom, and Dean Croushore. "Forecasting with a Real-Time Data Set for Macroeconomists," Federal Reserve Bank of Philadelphia working paper no. 01-10, July 2001.

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., Eric Ghysels, and Myles Callan. "A Multivariate Time Series Analysis of the Data Revision Process for Industrial Production and the Composite Leading Indicator," in Robert F. Engle and Halbert White, eds., *Cointegration, Causality, and Forecasting.* Oxford: Oxford University Press, 1999.

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

Figure 1. Nonparametric Spectral Density Estimates

Benchmark Vintages, Log Real Consumption Ratios

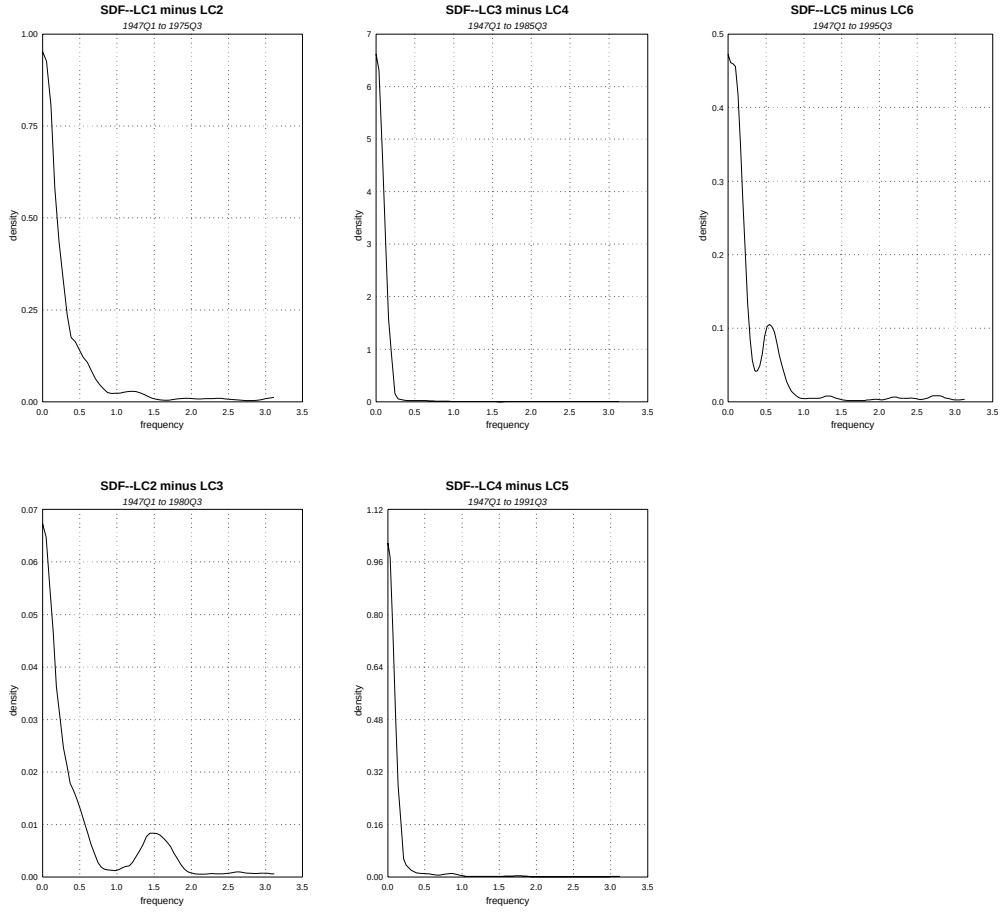

Figure 2. Nonparametric Spectral Density Estimates

Benchmark Vintages, Delta Log Real Consumption Ratios

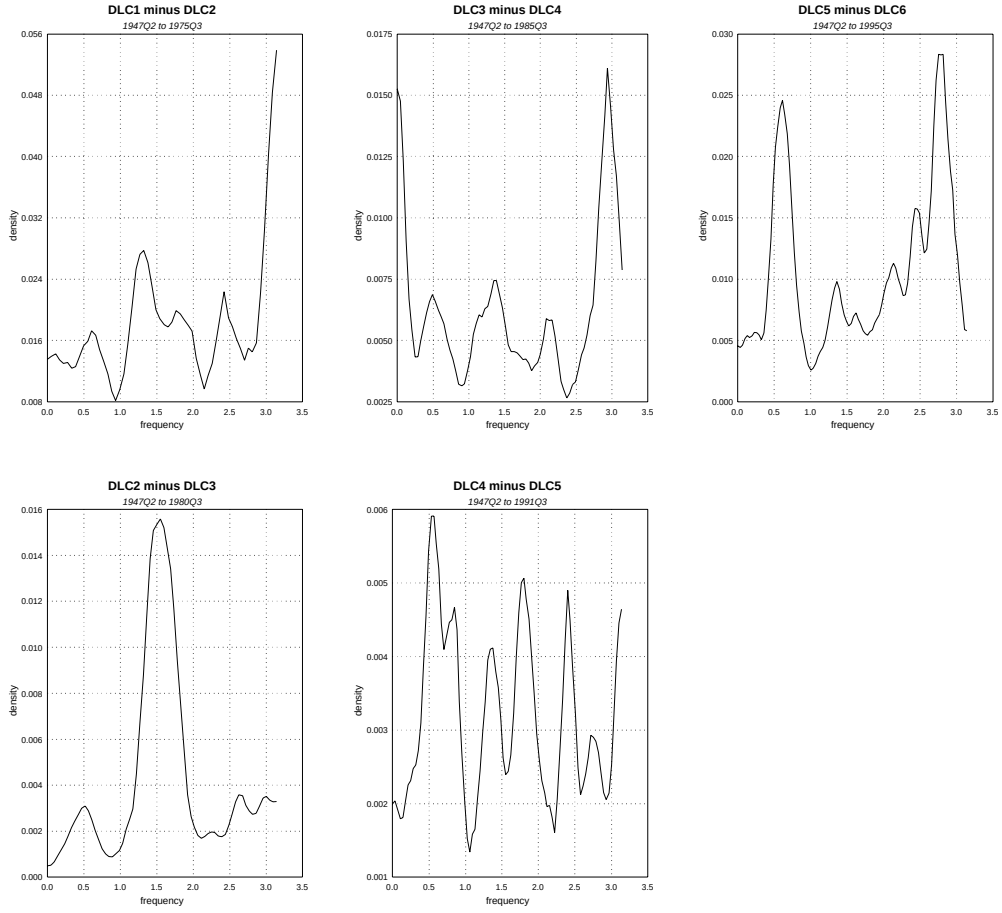

Figure 3. Nonparametric Squared Coherence Estimates

Benchmark Vintages, Delta Log Output & Delta Log Consumption

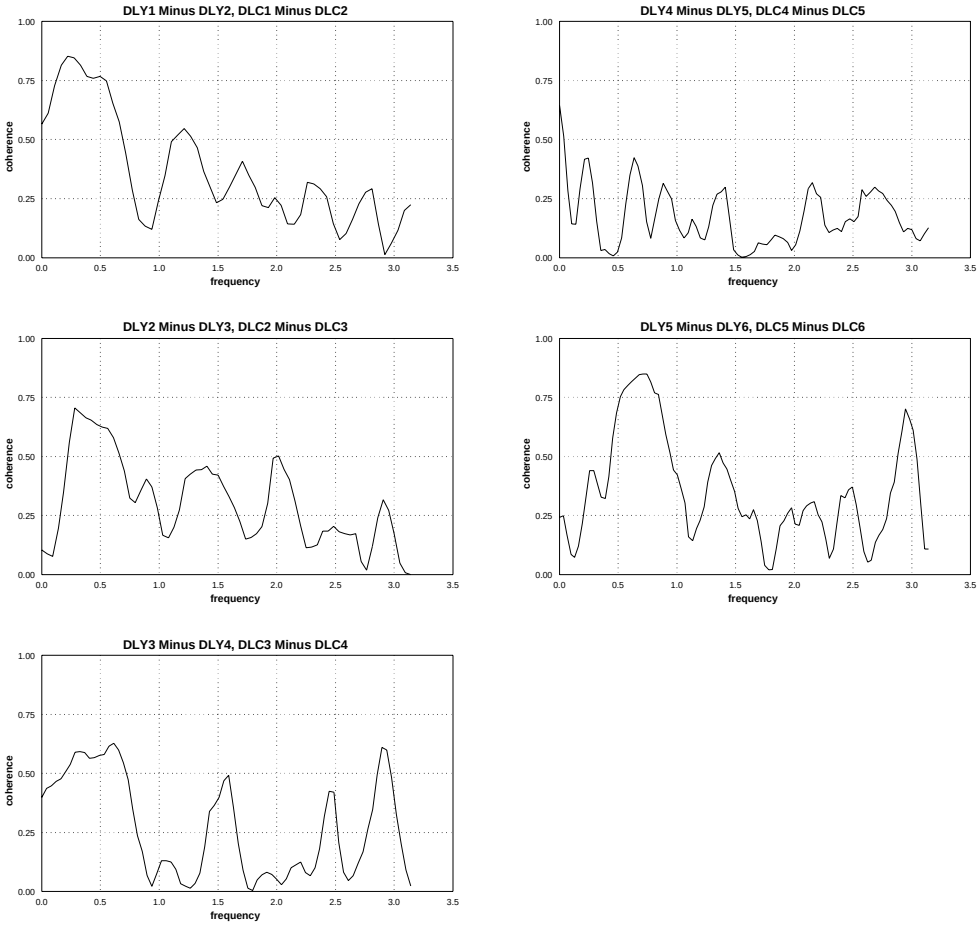

Figure 4. A Comparison of Trend Growth Estimates

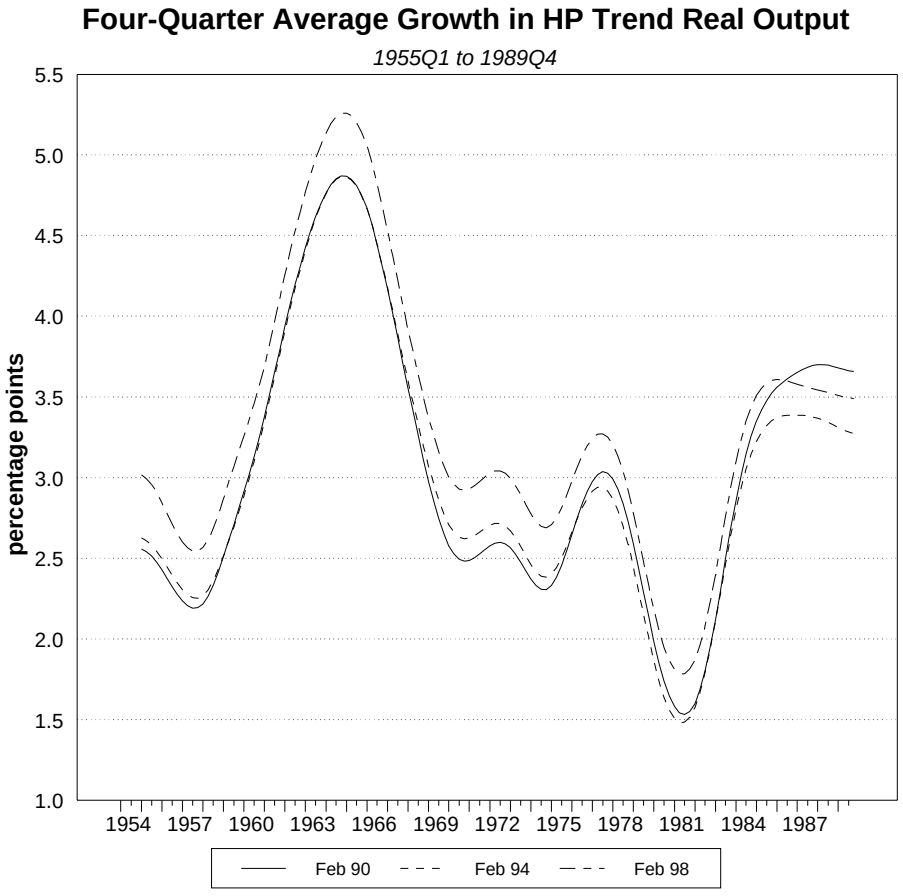

Figure 5. A Comparison of Blanchard/Quah Structural Shocks

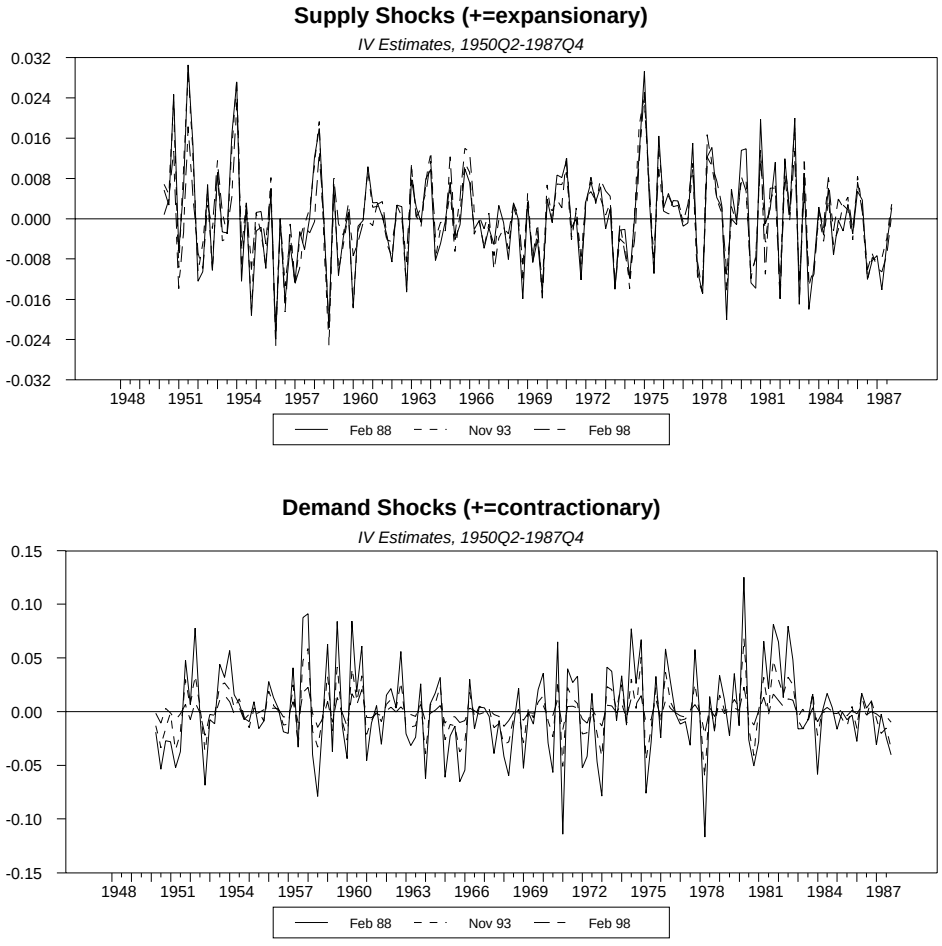

Figure 6. A Comparison of Blanchard/Quah Impulse Responses