Commentary

Dean Croushore

It is a pleasure to discuss Richard Anderson and Charles Gascon's (2009) article on their attempt to develop a state-space model to measure potential output growth in the face of data revisions. They use the methodology of Cunningham et al. (2007) applied to real output, to see if they can develop a better measure of potential output than other researchers. Such an approach seems promising, and they develop a unique method to study the data.

This approach holds promise because many practical approaches based on standard statistical models or production functions have not proven reliable indicators of potential output. One reason these methods may fail could be that the data are revised and the methods used do not account for such revisions. By accounting for data revisions in a systematic way, the authors hope to develop an improved calculation of potential output.

However, if the potential output series is subject to breaks not easily detected for many years, this approach may not be fruitful—you simply must wait many years to determine what potential output is. The state-space method may be ideal for calculating latent variables that correspond to an observable variable subject to large data revisions, but it is not helpful for early detection of breaks in series like potential output. There is simply no getting around the fundamental fact that potential output inherently requires the use of a two-sided filter and will be tremendously imprecise at the end of the sample when only a one-sided filter can be used.

CONGRESSIONAL BUDGET OFFICE MEASURES OF POTENTIAL OUTPUT

Many economic models rely on the concept of potential output, yet it is not observable. As new data arrive over time, practitioners who need a measure of potential output for their models use various statistical procedures to revise their view of potential output. One such practitioner is the Congressional Budget Office (CBO), which has produced a measure of potential output since 1991. An examination of some of the changes in their measure of potential output over time helps illustrate some of the difficulties of using the concept.

Figure 1 shows the CBO January 1991 and January 1996 versions of potential output growth. The vertical bars indicate the dates the series were created. In the 1991 version, potential output growth rises in discrete steps over time; in the 1996 version, growth rates evolve more smoothly. In the 1996 version, there is substantial volatility in potential output growth in the 1970s and early 1980s; in the 1991 version it is smoother.

Figure 2 compares the CBO 1996 and 2001 versions. Differences in the series' volatility in the 1970s and early 1980s and growth rates in the 1990s and 2000s are substantial. For example, in 1996, the CBO thought potential output growth for 1996 was about 2 percent per year; but in 2001, they thought it was about 3 percent.

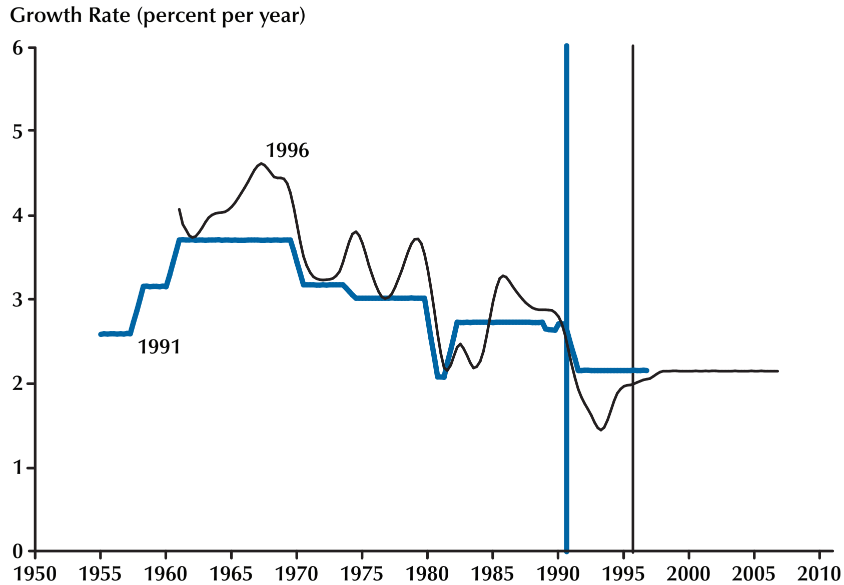
The CBO 2001 and 2008 versions of potential output growth (Figure 3) show even greater volatil-

Dean Croushore is a professor of economics and Rigsby Fellow at the University of Richmond.

Federal Reserve Bank of St. Louis *Review*, July/August 2009, *91*(4), pp. 371-81.

Figure 1

CBO Potential Output Growth, 1991 and 1996

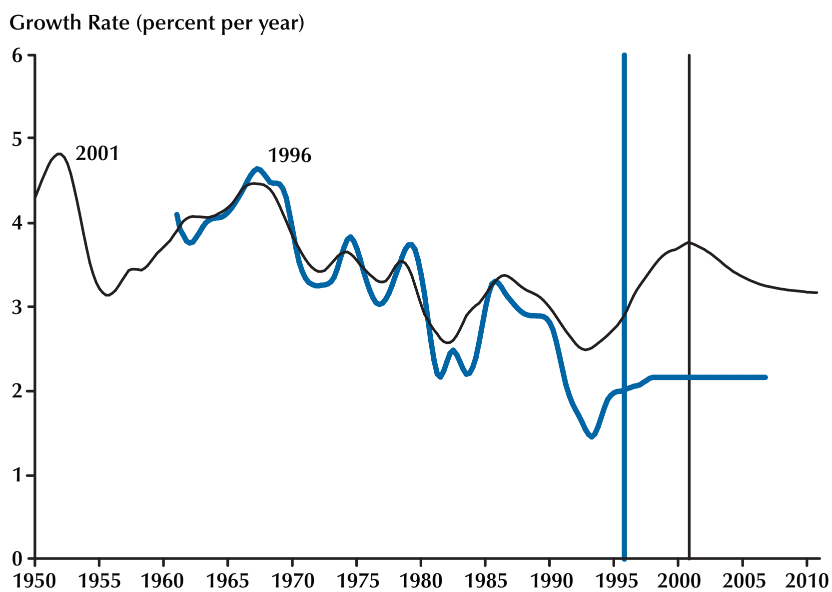


NOTE: The vertical bars indicate the dates the data series were created.

SOURCE: Federal Reserve Bank of St. Louis ArchivaL Federal Reserve Economic Data (ALFRED) database (series ID: GDPPOT).

Figure 2

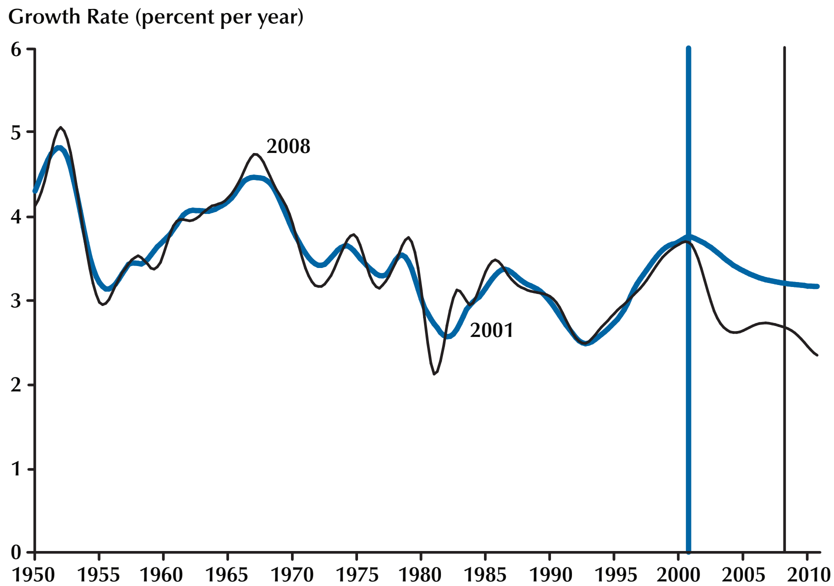
CBO Potential Output Growth, 1996 and 2001



SOURCE: Federal Reserve Bank of St. Louis ALFRED database (series ID: GDPPOT).

Figure 3

CBO Potential Output Growth, 2001 and 2008



SOURCE: Federal Reserve Bank of St. Louis ALFRED database (series ID: GDPPOT).

ity in the 1970s and early 1980s than earlier published series (see Figures 1 and 2) and a large difference in growth rates in the 2000s.

Thus, in the CBO's view, the period with the greatest revisions to potential output growth over time is the 1970s and early 1980s. In addition, in both the 1996 and 2001 versions, the potential growth rates at the end points of the sample changed substantially over time. This end-point problem is the major challenge to constructing a better measure of potential output.

KEY ASPECTS OF THE ANDERSON AND GASCON APPROACH

Key aspects of the approach taken by Anderson and Gascon include using a state-space approach (a very reasonable method) and exploiting the forecastability of data revisions following Cunningham et al. (2007). However, the real-time research literature, as described in detail in Croushore (2008a), includes few examples of macroeconomic variables whose revisions are forecastable in real time. Forecastable variables include U.S. retail sales (see Conrad and Corrado, 1979), Mexican industrial production (see Guerrero, 1993), gross domestic product (GDP) in Japan and the United Kingdom (see Faust, Roger, and Wright, 2005), and U.S. core personal consumption expenditures (PCE) inflation (see Croushore, 2008b). U.K. GDP is the focus of Cunningham et al. (2007). For U.S. GDP, revisions are not likely forecastable at all. And if this indeed is the case, the major feature of the Anderson and Gascon article could be a false trail.

A simulated out-of-sample exercise using real-time data must be performed to determine whether revisions are forecastable. Simply running a regression using an entire sample of data is not sufficient because finding a significant coefficient using the whole sample does not mean revisions are forecastable in real time.

The proper procedure to determine whether revisions are forecastable is described in Croushore

(2008b) regarding forecasting revisions to core PCE inflation. Suppose you think the initial release of data is not a good forecast of data to be released in the annual July revision of the national income and product accounts. Specifically, suppose you are standing in the second quarter of 1985 and have just received the initial release of the PCE inflation rate for 1985:Q1. You need to run a regression using as the dependent variable all the data on revisions from the initial release through the government's annual release in the current period, so the sample period is 1965:Q3–1983:Q4. So, you regress the revisions to the initial release for each date and a constant term:

$$Revision(t) = \alpha + \beta \; initial(t) + \varepsilon(t). \tag{1}$$

Next, use the estimates of α and β to make a forecast of the August revision that will occur in 1986:

$$\hat{r}(1985\text{:Q1}) = \hat{\alpha} + \hat{\beta} \cdot initial(1985\text{:Q1}).$$

Repeat this procedure for releases from 1985:Q2 to 2006:Q4. Finally, forecast the value of the annual revision for each date from 1985:Q1 to 2006:Q4 based on the formula

$$\hat{A}(t) = initial(t) + \hat{r}(t). \tag{2}$$

At the end of this process, examine the root mean squared forecast errors (RMSEs) as follows: Take the annual release value as the realization and compare the RMSE of the forecast of that value (given by equation (2)) with the RMSE of the forecast of that value assuming that the initial release is an optimal forecast. In such a case, the results show that it *is* possible to forecast the annual revision. Indeed, had the Federal Reserve used this procedure, it would have forecast an upward revision to core PCE inflation in 2002 and might not have worried so much about the unwelcome fall in inflation that was a major concern in this period. However, following such a method does not appear to work for U.S. real GDP. Cunningham et al. (2007) found that it worked for U.K. real GDP, but Anderson and Gascon's attempt to use it for U.S. real GDP is less likely to be fruitful. This is not to say that the initial release of U.S. real GDP data is an optimal forecast of the latest data, only that no one has successfully forecasted the revisions in the manner described above. You could argue that we should always assume that real GDP will be revised upward because the statistical agencies will always fall behind innovative processes, so GDP will be higher than initially reported. But the major reasons for upward revisions to GDP in the past include the reclassification of government spending on capital goods as investment, the change in the treatment of business software, and similar innovations that raised the entire level of real GDP. Whether similar upward revisions will occur in the future is uncertain.

THE STRUCTURE OF REAL-TIME DATA

Researchers of real-time data begin by developing a vintage matrix, consisting of the data as reported by the government statistical agency at various dates. An example is given in Table 1.

In the vintage matrix, each column represents a vintage, that is, the date on which a data series is published. For example, the first column reports the dates from 1947:Q1 to 1965:Q3 for data that would have been observable in November 1965. Each row in the matrix represents an activity date, that is, the date for which economic activity is measured. For example, the first row shows various measures for 1947:Q1. Moving across rows shows how data for a particular activity date are revised over time. The main diagonal of the matrix shows initial releases of the data for each activity date, which moves across vintages. Huge jumps in numbers indicate benchmark revisions with base-year changes. For example, in the first row, for 1947:Q1 the value rises from 306.4 in early vintages to 1570.5 in the most recent vintages.

Until about 1999, researchers studying monetary policy or forecasters building models ignored the vintage matrix and simply used the last column of the matrix available at the time—the latest data. If data revisions are small and white noise, this is a reasonable procedure. But in 1999, the Federal Reserve Bank of Philadelphia put together a large real-time dataset for macroeconomists, and it became possible for researchers and fore-

Table 1

The Vintage Matrix

	Real Output					
	Vintage (*v*)					
Activity date	11/65	02/66	05/66	…	11/07	02/08
1947:Q1	306.4	306.4	306.4	…	1,570.5	1,570.5
1947:Q2	309.0	309.0	309.0	…	1,568.7	1,568.7
1947:Q3	309.6	309.6	309.6	…	1,568.0	1,568.0
⋮	⋮	⋮	⋮		⋮	⋮
1965:Q3	609.1	613.0	613.0	…	3,214.1	3,214.1
1965:Q4	NA	621.7	624.4	…	3,291.8	3,291.8
1966:Q1	NA	NA	633.8	…	3,372.3	3,372.3
⋮	⋮	⋮	⋮		⋮	⋮
2007:Q1	NA	NA	NA	…	11,412.6	11,412.6
2007:Q2	NA	NA	NA	…	11,520.1	11,520.1
2007:Q3	NA	NA	NA	…	11,630.7	11,658.9
2007:Q4	NA	NA	NA	…	NA	11,677.4

SOURCE: Federal Reserve Bank of Philadelphia Real-Time Dataset for Macroeconomists (RTDSM; series ID: ROUTPUT).

casters to use the entire vintage matrix (see Croushore and Stark, 2001). Subsequent work at the Federal Reserve Bank of St. Louis expanded the Philadelphia Fed's work to create the vintage matrix for a much larger set of variables. The availability of such data has allowed researchers of real-time data to study data revisions and how they affect monetary policy and forecasting. The data revisions turn out to be neither small nor white noise, so accounting for data revisions is paramount.

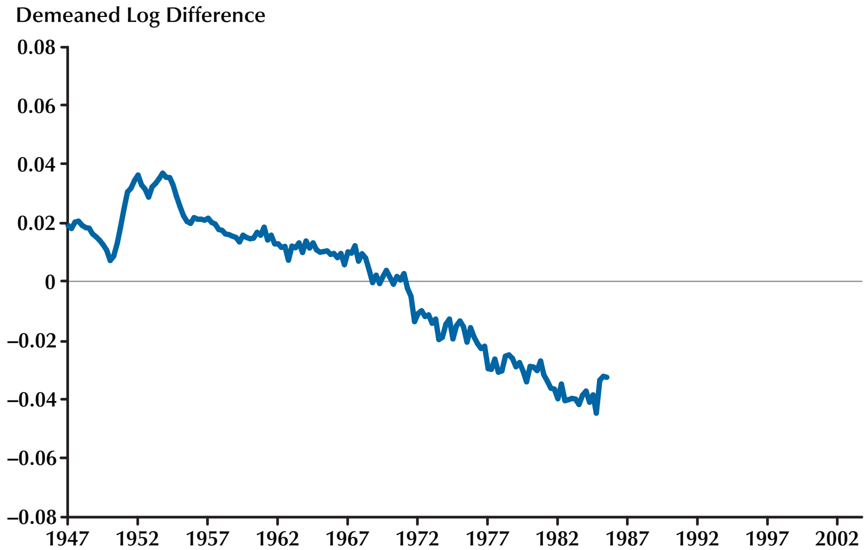
Researchers of real-time data have explored a number of ways to study what happens in the vintage matrix. One of the main distinctions in the literature that is crucial to econometric evaluation of data revisions is the distinction between "news" and "noise." Data revisions contain *news* if the initial release of the data is an optimal forecast of the later data. If so, then data revisions are not predictable. On the other hand, if data revisions reduce *noise*, then each data release equals the truth plus a measurement error; but because the data release is not an optimal forecast, it is predictable.

Empirical findings concerning news and noise are mixed. Money-supply data contain noise, according to Mankiw, Runkle, and Shapiro (1984), but GDP releases represent news, according to Mankiw and Shapiro (1986). Different releases of the same variable can vary in their news and noise content, as Mork (1987) found. For U.K. data, releases of most components of GDP contain noise, according to Patterson and Heravi (1991). The distinction between news and noise is vital to some state-space models, such as the one developed by Jacobs and van Norden (2007).

Anderson and Gascon ignore the distinction between news and noise because they develop a new and unique way to slice up the vintage matrix. Rather than focus on the vintage date, their analysis is a function of the "maturity" of data—that is, how long a piece of data for a given activity date has matured. They then track that piece of data over a length of time that they call the "revision horizon," which they can vary to discover different properties in the data of the revisions. This is a clever procedure and has the potential to lead to interesting results.

Figure 4

Stark Plot: December 1985–November 1991



SOURCE: Author's calculations using data from the Federal Reserve Bank of Philadelphia RTDSM (series ID: ROUTPUT).

The statistical model used by Anderson and Gascon is based on the following equation:

$$y_t^j = y_t + c^j + v_t^j.$$

A measured piece of data of some maturity j for activity date t is equal to the true value of the variable at activity date t, plus a bias term that is a function of maturity (but not vintage or activity date), plus a measurement error that is a function of both maturity and the activity date. This is the same method used by Cunningham et al. (2007).

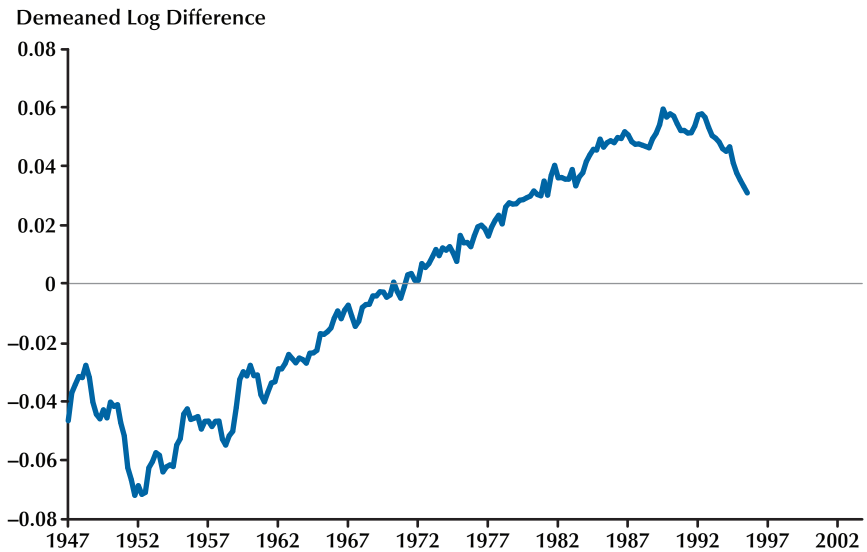
The Problem of Benchmark Revisions

Unfortunately, the Anderson and Gascon method may not work well if there are large and significant benchmark revisions to the data, because then the relationships in question would be a function of not only the activity date and maturity, but also a function of vintage, because benchmark revisions hit only one vintage of data every five years or so. But when they do hit, they affect the values of a different maturity for every activity date. So, if benchmark revisions are significant, then the Anderson and Gascon procedure could face problems.

Are benchmark revisions significant? I like to investigate the size of benchmark revisions using Stark plots, which I named after my frequent coauthor Tom Stark, who invented the plot (see Croushore and Stark, 2001). Let $X(t,s)$ represent the level of a variable that has been revised between vintages a and b, where vintage b is farther to the right in the vintage matrix and thus later in time than vintage a. Let m = the mean of $\log[X(\tau,b)/X(\tau,a)]$ for all the activity dates τ that are common to both vintages. The Stark plot is a plot of $\log[X(t,b)/X(t,a)] - m$. Such a plot would be a flat line if the new vintage were just a scaled-up version of the old one, that is, if $X(t,b) = \lambda X(t,a)$. If the plot shows an upward trend, then later data have more upward revisions than earlier data. Spikes in the plot show idiosyncratic data revisions. More important to analysis of data revisions would be any persistent deviation of the Stark plot from the zero line, which would imply a cor-

Figure 5

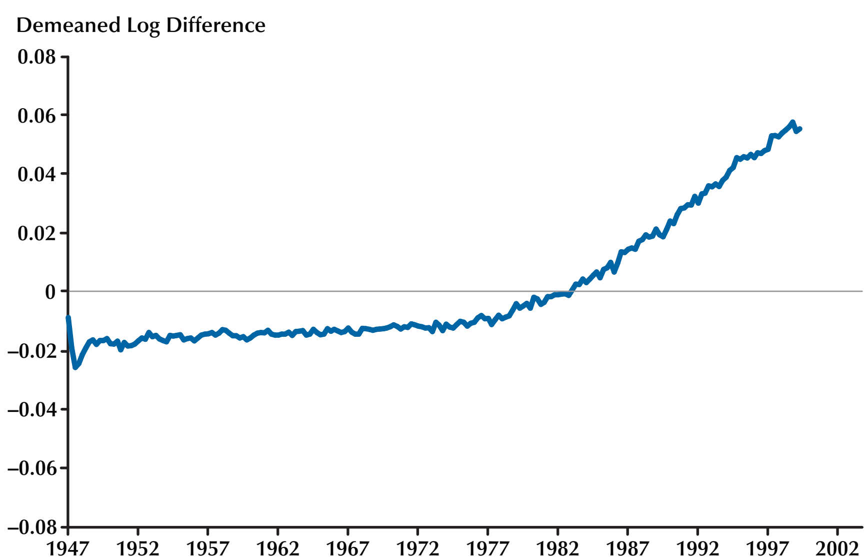
Stark Plot: December 1995–October 1999, Chain Weighting; Government Purchases Reclassified as Investment



SOURCE: Author's calculations using data from the Federal Reserve Bank of Philadelphia RTDSM (series ID: ROUTPUT).

Figure 6

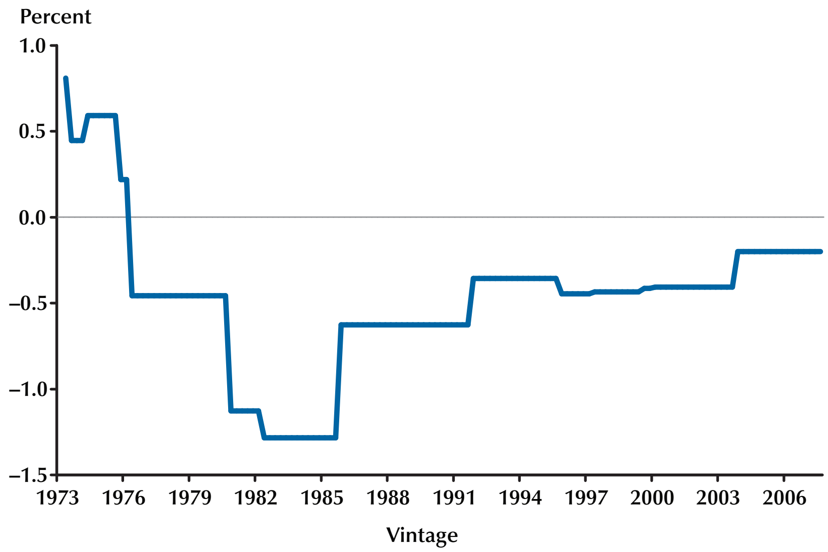
Stark Plot: October 1999–November 2003; Software Reclassified as Investment



SOURCE: Author's calculations using data from the Federal Reserve Bank of Philadelphia RTDSM (series ID: ROUTPUT).

Figure 7

Real Consumption Growth, 1973:Q2



SOURCE: Author's calculations using data from the Federal Reserve Bank of Philadelphia RTDSM (series ID: RCON).

relation of revisions arising from the benchmark revision.

In Figures 4, 5, and 6, we examine Stark plots that span particular benchmark revisions. Figure 4 shows how vintage data were revised from December 1985 to November 1991 for activity dates from 1947:Q1 to 1985:Q3. The data early in the sample period show upward revisions and those later in the sample period show downward revisions. There is a clear pattern in the data, which is mainly driven by the benchmark revision to the data that was released in late December 1985 (the December 1985 vintage date corresponds to the data as it existed in the middle of the month).

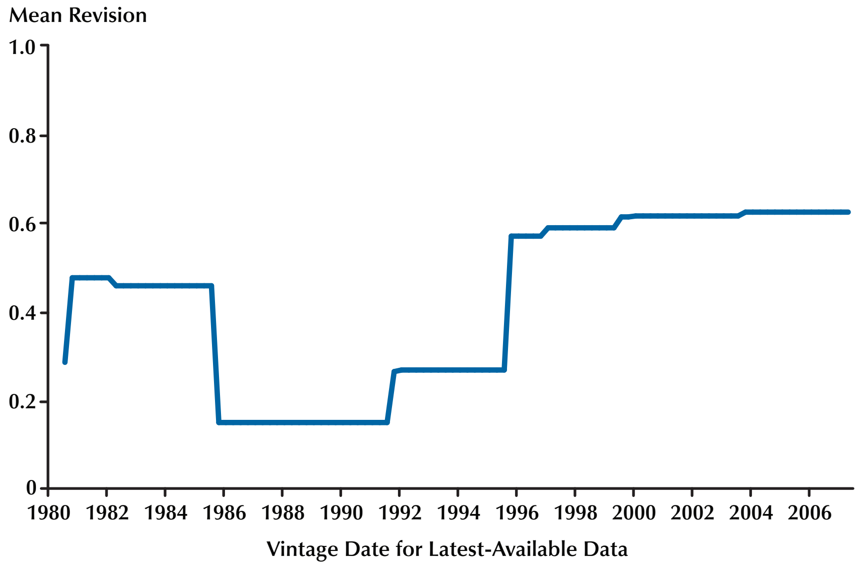
Figure 5 shows the revisions from December 1995 to October 1999, illustrating the impact of the benchmark revision of January 1996, which introduced chain weighting and reclassified government investment expenditures from their previous treatment as an investment expense subject to depreciation. The impact is very large, with data early in the sample showing downward revisions relative to data later in the sample.

Figure 6 illustrates the impact of the November 1999 benchmark revision, in which business software was reclassified as investment; we look at the changes from the October 1999 vintage to the November 2003 vintage. The nonlinear Stark plot suggests little change in growth rates in the early part of the sample, but increasing growth rates later in the sample. The impact of these changes in the benchmarks is considerable. There is clearly a significant change in the entire trajectory of the variable over time, which should be accounted for in any empirical investigation of the variable.

Do revisions ever settle down and stop occurring? In principle, they do under chain weighting, except for redefinitions that occur in benchmark revisions. For example, Figure 7 shows the growth rate of real consumption spending for activity date 1973:Q2. It has been revised by several percentage points over time and changed significantly as recently as 2003, some 30 years after the activity date. Thus, we cannot be confident that data are ever truly final and that there will never be a significant future revision.

Figure 8

Mean Revision, Initial to Latest



SOURCE: Author's calculations using data from the Federal Reserve Bank of Philadelphia RTDSM (series ID: ROUTPUT).

One key idea in the Anderson and Gascon article is to exploit the apparent bias in initial releases of the data. Unfortunately the bias seems to jump at benchmarks, as the Stark plots suggest. To see the jumps more clearly, Figure 8 plots what one would have thought the bias was at different vintage dates for real output growth. That is, it calculates the mean revision from the initial release to the latest available data, where for the sample of data from 1965:Q3 to 1975:Q3 the vintages of the latest available data are from 1980:Q3 to 2007:Q2. If we were standing in 1980:Q3, Figure 8 indicates we would have thought that the bias in the initial release of real output growth was 0.28 percentage points. But someone observing the data in the period from 1980:Q4 to 1982:Q1 would have thought it was 0.45 percentage points. And the apparent bias keeps changing over time, ending in 2007:Q2 at 0.62 percentage points. So, the bias changes depending on the date when you measure the bias. The same is true if you allow the sample period to change, rather than focusing on just one sample period as we did in Figure 8.

The Stark plots provide important information for researchers—that the bias is a function of the benchmark dates, not just maturity. Thus, equation (3) in Anderson and Gascon,

$$c^j - c^1(1+\lambda)^{j-1},$$

which treats the bias solely as maturity dependent, is not likely to work across benchmark revisions.

The other key assumption that Anderson and Gascon use in their empirical framework is that the measurement error follows an autoregressive (AR(q)) process. The Stark plots suggest that such an assumption is not well justified, because the process at different benchmark revisions is much more complicated than any AR(q) process can capture.

WHERE NEXT?

Given the issues identified here, how should the authors proceed with their research? I offer five suggestions. First, they should compare their results on the potential output series generated by their method with that of some benchmark, such as some of the series generated in Orphanides and van Norden (2002). Second, they should examine forecasts of revisions that can be generated by the model to see if they match up reasonably well with actual revisions. Third, they should see how stable their model is when it encounters a benchmark revision. That is, if the model were used in real time to generate a series for potential output and then suddenly hit a benchmark revision, what would that do to the potential output series? Fourth, they should attempt to reconcile the Stark plots with their assumptions about the data to see how much damage such assumptions might make. Finally, because they have ignored the distinction between news and noise, they might want to consider the impact the results of Jacobs and van Norden (2007) would have on their empirical model.

CONCLUSION

The research by Anderson and Gascon is an interesting and potentially valuable contribution to estimating potential output. However, practical issues, in particular the existence of benchmark revisions, may derail it. It may be that no new empirical method can handle revisions and produce better estimates of potential output in real time than current methods. If so, then we may have to conclude that potential output cannot be measured accurately enough in real time to be of any value for policymakers.

REFERENCES

Anderson, Richard and Gascon, Charles. "Estimating U.S. Output Growth with Vintage Data in a State-Space Framework." Federal Reserve Bank of St. Louis *Review*, July/August 2009, *91*(4), pp. 349-69.

Conrad, William and Corrado, Carol. "Application of the Kalman Filter to Revisions in Monthly Retail Sales Estimates." *Journal of Economic Dynamics and Control*, May 1979, *1*(2), pp. 177-98.

Croushore, Dean. "Frontiers of Real-Time Data Analysis." Working Paper No. 08-4, Federal Reserve Bank of Philadelphia, March 2008a; www.philadelphiafed.org/research-and-data/publications/working-papers//2008/wp08-4.pdf.

Croushore, Dean. "Revisions to PCE Inflation Measures: Implications for Monetary Policy." Working Paper 08-8, Federal Reserve Bank of Philadelphia, March 2008b.

Croushore, Dean and Stark, Thomas. "A Real-Time Data Set for Macroeconomists." *Journal of Econometrics*, November 2001, *105*(1), pp. 111-30.

Cunningham, Alastair; Eklund, Jana; Jeffery, Christopher; Kapetanios, George and Labhard, Vincent. "A State Space Approach to Extracting the Signal from Uncertain Data." Working Paper No. 336, Bank of England, November 2007.

Faust, Jon; Rogers, John H. and Wright, Jonathan H. "News and Noise in G-7 GDP Announcements." *Journal of Money, Credit, and Banking*, June 2005, *37*(3), pp. 403-19.

Guerrero, Victor M. "Combining Historical and Preliminary Information to Obtain Timely Time Series Data." *International Journal of Forecasting*, December 1993, *9*(4), pp. 477-85.

Jacobs, Jan P.A.M. and van Norden, Simon. "Modeling Data Revisions: Measurement Error and Dynamics of 'True' Values." Working paper, HEC Montreal, June 2007.

Mankiw, N. Gregory; Runkle, David E. and Shapiro, Matthew D. "Are Preliminary Announcements of the Money Stock Rational Forecasts?" *Journal of Monetary Economics*, July 1984, *14*(1), pp. 15-27.

Mankiw, N. Gregory and Shapiro, Matthew D. "News or Noise? An Analysis of GNP Revisions." *Survey of Current Business*, May 1986, *66*(5), pp. 20-25.

Mork, Knut Anton. "Ain't Behavin': Forecast Errors and Measurement Errors in Early GNP Estimates." *Journal of Business and Economic Statistics*, April 1987, *5*(2), pp. 165-75.

Orphanides, Athanasios and van Norden, Simon. "The Unreliability of Output Gap Estimates in Real Time." *Review of Economics and Statistics*, November 2002, *84*(4), pp. 569-83.

Patterson, K.D. and Heravi, S.M. "Data Revisions and the Expenditure Components of GDP." *Economic Journal*, July 1991, *101*(407), pp. 887-901.