

The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time

Athanasios Orphanides*
Board of Governors of the Federal Reserve System

Simon van Norden
HEC Montréal, CIRANO and CIREQ

July 24, 2004

Abstract

A stable predictive relationship between inflation and the output gap, often referred to as a Phillips curve, provides the basis for countercyclical monetary policy in many models. In this paper, we evaluate the usefulness of alternative univariate and multivariate estimates of the output gap for predicting inflation. Many of the ex post output gap measures we examine appear to be quite useful for predicting inflation. However, forecasts using real-time estimates of the same measures do not perform nearly as well. The relative usefulness of real-time output gap estimates diminishes further when compared to simple bivariate forecasting models which use past inflation and output growth. Forecast performance also appears to be unstable over time, with models often performing differently over periods of high and low inflation. These results call into question the practical usefulness of the output gap concept for forecasting inflation.

KEYWORDS: Phillips curve, output gap, inflation forecasts, real-time data.

JEL Classification System: E37, C53.

Correspondence: Orphanides: Division of Monetary Affairs, Board of Governors of the Federal Reserve System, Washington, D.C. 20551, USA. Tel.: (202) 452-2654, e-mail: Athanasios.Orphanides@frb.gov. van Norden: H.E.C. Montréal, 3000 Chemin de la Côte Sainte Catherine, Montréal QC, Canada H3T 2A7. e-mail: simon.van-norden@hec.ca.

* We benefited from presentations of earlier drafts at the European Central Bank, CIRANO, the Federal Reserve Bank of Philadelphia Conference on Real Time Data Analysis, the Centre for Growth and Business Cycle Research, as well as at the annual meetings of the American Economics Association, the European Economics Association and the Canadian Economics Association. We would also like to thank Sharon Kozicki, Tim Cogley, Jeremy Piger, Todd Clark, Ken West and two anonymous referees for useful comments and discussions. Athanasios Orphanides wishes to thank the Sveriges Riksbank and European Central Bank for their hospitality during September 2001 when part of this work was completed. Simon van Norden wishes to thank the SSHRC and the HEC Montréal for their financial support. The opinions expressed are those of the authors and do not necessarily reflect views of the Board of Governors of the Federal Reserve System.

1 Introduction

A stable predictive relationship between inflation and a measure of deviations of aggregate demand from the economy's potential supply—the “output gap”—provides the basis for many formulations of activist countercyclical stabilization policy. Such a relationship, referred to as a Phillips curve, is often seen as a helpful guide for policymakers aiming to maintain low inflation and stable economic growth. According to this paradigm, when aggregate demand exceeds potential output, the economy is subject to inflationary pressures and inflation should be expected to rise. Under these circumstances, policymakers might wish to adopt policies restricting aggregate demand aiming to contain the acceleration in prices. Similarly, when aggregate demand falls short of potential supply, inflation should be expected to fall, prompting policymakers to consider adoption of expansionary policies to restore stability.¹

Even assuming that the theoretical motivation for a relationship between the output gap and inflation is fundamentally correct, a number of issues may complicate its use for forecasting in practice. First, the appropriate empirical definition of “potential output”—and the accompanying “output gap”—that might be useful in practice is far from clear. For any given empirical definition of the output gap, the exact form of its empirical relationship with inflation is not known a priori and would need to be determined from the data. Second, even if the proper concept and empirical relationship were identified, the operational usefulness of the output gap will be limited by the availability of timely and reliable estimates of the identified concept. As is well known, empirical estimates of the output gap are generally subject to significant and highly persistent revisions. The subsequent evolution of the economy provides useful information about the state of the business cycle which leads to improved historical estimates of the gap. As a result, considerable uncertainty regarding

¹The appeal of this paradigm is evidenced by the fact that many estimated models employed for monetary policy analysis, including at numerous central banks, feature estimated “Phillips curves” of various forms. See Bryant, Hooper and Mann (1993) and Taylor (1999) for collections of monetary policy evaluation studies that feature such estimated models.

the value of the gap remains even long after it would be needed for forecasting inflation.² This in turn implies that although the output gap may be quite useful for historical analysis, its practical usefulness for forecasting inflation in real time may be quite limited.

In this paper we assess the usefulness of alternative methods for estimation of the output gap for predicting inflation, paying particular attention to the distinction between *suggested usefulness*—based on *ex post* analysis using revised output gaps and *operational usefulness*—based on simulated real-time out-of-sample analysis.³ First, using out-of-sample analysis based on *ex post* estimates of the output gap, we confirm that many concepts appear to be useful for predicting inflation. This is as would be expected since the implicit Phillips curve relationships recovered in this manner are similar to the relationships commonly found in empirical macroeconomic models. To assess their operational usefulness, we generate out-of-sample forecasts based on *real-time* output gap measures; those constructed using only data (and parameter estimates) available at the time forecasts are generated.⁴ We compare the resulting forecasts against two benchmarks, univariate forecasts of inflation as well as bivariate forecasts that employ information from output growth, in addition to past inflation.

Our findings based on this real-time analysis show that forecasts based on *ex post* estimates of the output gap severely overstate the gap's usefulness for predicting inflation. Further, real-time forecasts using the output gap are often less accurate than forecasts that abstract from the output gap concept altogether. And the relative usefulness of real-time output gap estimates diminishes further when compared to simple bivariate forecasting models which use past inflation and output growth. In some cases, we find certain measures of the output gap produce superior forecasts of inflation. However, relative performance seems

²Orphanides and van Norden (2002) document the extent of this unreliability.

³Our analysis is related to investigations of the usefulness of the unemployment gap for forecasting inflation, such as Stock and Watson (1999), Atkeson and Ohanian (2001), and Fisher, Liu and Zhou (2001). In some macroeconomic models, unemployment gaps and output gaps are related though Okun's law.

⁴For this exercise, we rely on the real-time dataset for macroeconomists which was created and is maintained by the Federal Reserve Bank of Philadelphia. See Croushore and Stark (2001) for background information regarding this database.

to vary considerably over time, with models which perform relatively well in some periods performing relatively poorly in others. Thus, past forecast performance may provide little guidance in selecting an operationally useful definition of the output gap going forward.

The remainder of this paper is organized as follows. In sections 2 and 3 we define the output gap concepts used for our forecasting exercise and detail the methodology of the exercise. The main results of the forecasting exercise are presented in section 4. Section 5 concludes.

2 Trends and Cycles Ex Post and in Real Time

One way to define the output gap is as the difference between actual output and an underlying unobserved trend towards which output would revert in the absence of business cycle fluctuations. Let q_t denote the (natural logarithm of) actual output during quarter t , and μ_t its trend. Then, the output gap, y_t can be defined as the cyclic component resulting from the decomposition of output into a trend and cycle component:

$$q_t = \mu_t + y_t$$

Since the underlying trend is unobserved, its measurement, and the resulting measurement of the output gap, very much depends on the choice of estimation method, underlying assumptions and available data that are brought to bear on the measurement problem. For any given method, simple changes in historical data and the availability of additional data can change, sometimes drastically, the resulting estimates of the cycle for a given quarter.

Evidence of the difference between historical and real-time estimates of output gaps has been presented by Orphanides and van Norden (2002). In Table 1, we present some of the summary reliability indicators they examine for twelve alternative measures of the output gap which we employ in our analysis.⁵ These results mirror those of Orphanides

⁵Brief descriptions of the various methods appear in Appendix A. Further details, as well as the output gaps used in this study, including the data and programs used to create them, are freely available from the authors at <http://www.hec.ca/pages/simon.van-norden>.

and van Norden (2002). We find that revisions in real-time estimates are often of the same magnitude as the historical estimates themselves and that, for many of the alternative methods, historical and real-time estimates frequently have opposite signs.

The importance of *ex post* revisions to output gap estimates suggests that the presence of a predictive relationship between inflation and *ex post* estimated output gap measures does not guarantee that the output gap will be useful for forecasting inflation in practice. Simply, the *ex post* estimates of output gaps at a point in time may differ substantially from estimates which could be made without the benefit of hindsight. As well, these differences may hinder the real-time estimation of the presumed predictive relationship, further complicating the real-time forecasting problem. In the remainder of this section we describe the data and the measurement of output gap we use in our forecasting experiment.

2.1 Data Sources and Vintages

We use the term *vintage* to describe the values for data series as published at a particular point in time. Most of our data is taken from the real-time data set compiled by Croushore and Stark (2001); we use the quarterly vintages from 1965Q1 to 2003Q3 for real output. Construction of the output series and its revision over time is further described in Orphanides and van Norden (1999, 2002.) We use 2003Q3 data as “final data” recognizing, of course, that “final” is very much an ephemeral concept in the measurement of output.

To measure inflation, we use the change in the log of the consumer price index (CPI). We use this both for our forecasting experiments and also to estimate measures of the output gap based on multivariate models that include inflation. For all of our analysis, we rely on the consumer price index (CPI) as available in 2003Q3. CPI data do not generally undergo a similar revision process as the output data. The major source of revision is changes in seasonal factors most noticeable at a monthly frequency. We therefore use the 2003Q3 vintage of CPI data for all the analysis which allows us to focus our attention on the effects of revisions in the output data and the estimation of the output gap in our analysis. One

of our models (Structural VAR) also uses data on interest rates, which are never revised.

2.2 Measuring Output Gaps

We construct output gap estimates using a variety of different models, as listed in Table 1. Some, such as the linear or the quadratic trend, are based on purely deterministic detrending methods. Some, such as the Hodrick-Prescott filter, do not directly rely on statistical model-fitting. Five are estimated unobserved components models, of which three (Watson, Harvey-Clark and Harvey-Jaeger) are univariate models and two (Kuttner and Gerlach-Smets) are bivariate models, using data on both output and prices. The remaining models are all univariate with the exception of the Structural VAR method, which uses a trivariate VAR with long-run restrictions as proposed by Blanchard and Quah (1989).

Each of the output gap models was used to produce gap estimates of varying vintages. Each output gap vintage uses precisely one vintage of the output data. An estimated output gap is called a final estimate if it uses the final data vintage. Note that all the output gap estimation techniques (aside from the Hodrick-Prescott filter) require that one or more parameters be estimated to fit the data. Such estimation was repeated for every combination of technique and vintage. This means, for example, that in constructing output gap vintages from an unobserved components model spanning the period 1969Q1-2002Q2 (134 quarters), we reestimate the model's parameters 134 times, and then store 134 series of smoothed estimates.

3 A Forecasting Experiment

We are interested in quantifying the extent to which the output gap concept provides a practical means of improving forecasts of inflation. The answer will clearly depend on a large number of factors, such as the time period of interest, the way in which forecasts are constructed, the benchmark against which such forecasts are compared, and the loss function used to evaluate the quality of different forecasts. We restrict our attention to

US CPI inflation since 1969 and use the mean-squared forecast error (MSFE) to compare forecast quality.

3.1 Forecasting Inflation and Benchmarks

Let $\pi_t^h = \log(P_t) - \log(P_{t-h})$ denote inflation over h quarters ending in quarter t . We examine forecasts of inflation at various horizons but use one year ($h=4$) as our baseline. Note that because of reporting lags, information for quarter t is not available before quarter $t+1$. Thus, a four-quarter ahead forecast is a forecast five quarters ahead of the last quarter for which actual data are available. Given data for quarter $t-1$ and earlier periods, our objective therefore is to forecast π_{t+4}^4 .

We examine simple linear forecasting models of the form:

$$\pi_{t+h}^h = \alpha + \sum_{i=1}^n \beta_i \cdot \pi_{t-i}^1 + \sum_{i=1}^m \gamma_i \cdot y_{t-i} + e_{t+h} \quad (1)$$

where n and m denote the number of lags of inflation and the output gap in the equation. We estimate the unknown coefficients $\{\alpha, \beta_i, \gamma_i\}$ by ordinary least squares. We set n and m using a variety of different methods; in the base case they are set using the Bayes Information Criterion (BIC).⁶

To provide a benchmark for comparison, we estimate a univariate forecasting model of inflation based on equation (1) but omitting the output gaps. We refer to this model as the autoregressive (AR) benchmark. Of course, the problem faced by forecasters in practice is more complex than the one we consider. One obvious and important difference is that the information set available to policymakers is much richer. It is therefore possible that output gaps might improve on simple univariate forecasts of inflation but not on forecasts using a broader range of inputs. For this reason, tests against an autoregressive forecast benchmark should be considered to be weak tests of the utility of empirical output gap models.

To provide a slightly stronger test, we also consider benchmark forecasts which replace

⁶Results which used AIC and other lag selection methods are available on request. These were found to give similar conclusions.

the output gap in (1) with the first difference of the log of real output. As St-Amant and van Norden (1998) argue, using output growth in this way can be interpreted as *implicitly* defining an estimated output gap as a one-sided filter of output growth with weights based on the estimated coefficients of equation (1). van Norden (1995) refers to such estimates as TOFU gaps.⁷ We therefore refer to this as the TF benchmark forecast. We interpret it as a simple reduced-form inflation forecast that uses a slightly larger information set than the AR benchmark, one which contains historical information on both prices and output. Since the forecasts based on our output gap measures include the same information set (past prices and the current vintage of output) as these unrestricted forecasting equations, comparing these forecast to the TF benchmark aids in isolating the usefulness (or lack thereof) of the economic structure and other restrictions embedded in the construction of the output gaps.

3.2 Forecasting and Output Gap Revisions

Several practical issues complicate the use of (1) for inflation forecasting. Since the most suitable number of lags of inflation and the output gap n and m , and the parameters of the equation are not known a priori, these need to be estimated with available data. As our sample increases and additional data become available, these estimates would change. In addition, output gap estimates (like output data) are revised over time. This in turn, can influence the selected number of lags and the parameters of equation (1) estimated in any given sample. In addition, given the parameters of the equation, revisions in the output gap will directly change the forecast value of inflation.

We therefore use (1) to construct 3 to 4 different kinds of forecasts for each output gap model. These forecasts differ in the way lag lengths are determined and in the way the output gap model is used.

Let $y_t^{i,j}$ be an estimate of the output gap at time t formed using data of vintage i ,

⁷Trivial Optimal Filter–Unrestricted.

where $i > t$ and $j = t$ or $i - 1$. For non-UC models (i.e. all except WT, CL, HJ, KT and GS) the index j is irrelevant; $y_t^{i,t} = y_t^{i,i-1}$. For UC models, $j = t$ denotes a *filtered* output gap estimate; although the model parameters are estimated from using data up to $i - 1$, the Kalman filter recursions to estimate the gap do not use data beyond t . For these same models, $j = i - 1$ denotes a *smoothed* estimate; although $y_t^{i,t}$ and $y_t^{i,i-1}$ use the same parameter estimates to calculate the output gap, the latter also uses the data after t to recursively update its estimate of y_t . When $T = 2003Q3$, the terminology of Orphanides and van Norden (2002) refers to the time series $\{y_t^{T,T-1}\}$ as *Final* estimates of the gap and to $\{y_t^{T,t}\}$ as *Quasi-Final* estimates. We will commonly refer to these as FL and QF estimates.

These different kinds of output gap estimates are used to construct different kinds of forecasts. The first of these uses fixed lag lengths with final estimates of the output gap to recursively estimate the forecasting equation

$$\pi_{t+h}^h = \hat{\alpha}^{t-1} + \sum_{i=1}^{\hat{n}} \hat{\beta}_i^{t-1} \cdot \pi_{t-i}^1 + \sum_{i=1}^{\hat{m}} \hat{\gamma}_i^{t-1} \cdot y_t^{T,T-1} + e_{t+h} \quad (2)$$

where T refers to 2003Q2. This replicates the kind of recursively-estimated, out-of-sample forecasting experiments which are commonly performed but which ignore output gap revision. These forecasts are infeasible because they require information (Final estimates of output gaps) which is not available at the time the forecast is made. They also estimate the optimal lag lengths \hat{m}, \hat{n} ex post. We refer to this Fixed-Lag Final-estimate forecast as FL-FL.

In the case of UC models, we can construct similar forecasts using Quasi-Final rather than Final estimates of the output gap

$$\pi_{t+h}^h = \hat{\alpha}^{t-1} + \sum_{i=1}^{\hat{n}} \hat{\beta}_i^{t-1} \cdot \pi_{t-i}^1 + \sum_{i=1}^{\hat{m}} \hat{\gamma}_i^{t-1} \cdot y_t^{T,t} + e_{t+h} \quad (3)$$

Orphanides and van Norden (2002) note that the difference between the Final and Quasi-Final estimates of the output accounts for the bulk of the revisions in the output gaps

they examine. The difference between the accuracy of these and the Final gap forecasts above helps us to understand the relative importance of errors in gap estimation for forecast accuracy. Like the Final gap forecasts, these forecasts are infeasible. We refer to these as FL-QF forecasts.

We also construct feasible forecasts which attempt to mirror closely the kinds of forecasts which practitioners would construct using such output gap models. Specifically:

- lag lengths for both explanatory variables vary over time and are estimated recursively.
- every time the parameters of the forecasting equation are re-estimated, the output gap series is updated by its latest available vintage.

The resulting Variable-Lag Real-Time output gap (VL-RT) forecasting equation takes the form⁸

$$\pi_{t+h}^h = \hat{\alpha}^{t-1} + \sum_{i=1}^{\hat{n}^{t-1}} \hat{\beta}_i^{t-1} \cdot \pi_{t-i}^1 + \sum_{i=1}^{\hat{m}^{t-1}} \hat{\gamma}_i^{t-1} \cdot y_{t-i}^{t,t-1} + e_{t+h} \quad (4)$$

where the superscripts on (\hat{m}, \hat{n}) indicate the information set used to estimate the lag lengths. While these are the most realistic forecasts we examine, they are also the most difficult to compute. Among other things, they require more than just the real-time gap estimates presented in Orphanides and van Norden (2002); they require *all vintages* of the complete estimated output gap series.

To summarize, we can construct two to three series of forecasts for each output gap model we analyze.

- (2) uses recursive estimation, fixed lag lengths and final output gap estimates.
- (3) uses recursive estimation, fixed lag lengths and quasi-final output gap estimates (which are only available for the 5 UC models we examine.)

⁸Note that in equation (4) we use *smoothed* estimates of the output gap ($y_{t-i}^{t,t-1}$) rather than *filtered* estimates ($y_{t-i}^{t,t-i}$). This reflects the common practice of practitioners, which is to use the most accurate possible estimate of the gap in estimating their forecast equations. Limited experiments which replaced these smoothed estimates with filtered estimates suggest that this does not have a major impact on forecast performance. Koenig, Dolmas and Piger (2003) discuss how the use of data of varying vintage affects forecast accuracy.

- (4) uses recursive estimation, variable lag lengths and all available vintages of smoothed output gap estimates.

We also examine one other type of forecast, one which uses variable lag lengths and final output gaps and which we refer to as VL-FL. Like the FL-QF forecast, this helps to isolate the contribution of output gap revision to forecast accuracy. As we will see below, however, these methods differ in the appropriate ways one should conduct inference.

3.3 Forecast Evaluation

We wish to evaluate the quality of the resulting forecasts by testing the null hypothesis that a given pair of models have equal MSFEs. Various tests of equal forecast accuracy have been proposed in recent years, notably by Diebold and Mariano (1995) for forecasting models without estimated parameters and by West (1996) for models with estimated parameters. While such tests have been popular, the assumptions they require are unfortunately violated for some of the hypotheses of interest here.

First, the use of Diebold-Mariano statistics with standard normal critical values for asymptotic inference is justified only if the two models being compared are not nested. However, when using suitable lag lengths, the output gap models nest the AR benchmark model. Clark and McCracken (2001) suggest alternative tests for the case of nested models, while Clark and McCracken (2003) find that the limiting distribution of these statistics is non-pivotal for forecast horizons greater than one period. To compare these models, we therefore use the MSE-F statistic proposed by McCracken (2000), which takes the form

$$MSE-F = P \cdot \frac{(MSFE_1 - MSFE_2)}{MSFE_2} \quad (5)$$

where P is the number of forecasts, $MSFE_1$ is the MSFE of the restricted model and $MSFE_2$ is the MSFE of the unrestricted model. The distribution of the statistic under the null hypothesis of equal MSFE is estimated via a bootstrap experiment with 2000 repli-

cations.⁹ Because these distributions are non-pivotal, the test statistics are bootstrapped anew for every different choice of (P, h, y, m, n) . This means that every p-value we report for the AR benchmark is based on its own set of 2000 bootstrap experiments.¹⁰

Second, while the available asymptotic theory underlying all such tests allows for the coefficients in an equation like (1) to be re-estimated over time, it assumes:

1. That lag lengths are fixed during the recursive estimation.
2. That the data remain fixed during the recursive estimation.
3. That the data are not estimated.

All these assumptions are violated for the VL-RT forecasts we construct. While we can compare MSFEs, there is therefore little which we can say with certainty about the significance of their apparent differences. However, as an approximate guide, we calculated the usual MSE-F statistics for these models and constructed approximate p-values for them. These p-values were estimated using the empirical distribution of bootstrapped MSE-F statistics generated from an experiment similar to that used above, but for the fact that:

1. Every time the forecasting equation is re-estimated, the information criterion is now used to re-determine the appropriate number of lags and adjust the equation accordingly.¹¹
2. Due to the difficulty in designing a bootstrap for *revisions* of the output gap, we simply ignored output gap revisions and performed the bootstrap using the quasi-final $(y_t^{T,t})$ estimates of the output gap.¹²

⁹See Appendix B for details on the bootstrap experiment.

¹⁰The single exception to this rule was that variations in the starting date of the OLS estimation of the forecasting equation were ignored; all bootstrap simulation used the base case starting date of 1955Q1.

¹¹This is computationally intensive. Given 12 different output gap measures, 2000 trials, approximately 135 recursive updates and a maximum lag length of 12, this requires $12 \cdot 2000 \cdot 135 \cdot 12^2 = 466,560,000$ OLS regressions for each set of reported p-values.

¹²This also considerably reduces the computational burden, as it allows for fast recursive OLS estimation via updating of the cross-moment matrices.

Inference in the case of the TF benchmark is more straightforward as the models of interest are no longer nested. Accordingly, we base our inference on the test statistics proposed by Diebold and Mariano (1996) and West (1996). Specifically, letting $d_t \equiv e_{it}^2 - e_{jt}^2$ be the difference in squared forecast errors between model i and model j at time t , $\bar{d} \equiv T^{-1} \cdot \sum_{t=1}^T (d_t)$ the mean difference, and $\rho_\tau \equiv T^{-1} \cdot \sum_{t=\tau+1}^T (d_t - \bar{d}) \cdot (d_{t-\tau} - \bar{d})$ the estimated autocovariance of d_t at lag τ , we compute the test statistic:

$$z = \frac{\bar{d}}{\sqrt{\Omega/T}} \quad (6)$$

where $\Omega \equiv \sum_{l=3}^3 (1 - |l|/4) \cdot \rho_l$ is the Newey-West Heteroscedasticity and Autocorrelation (HAC) robust estimator of the long-run variance of d_t for a lag truncation length of 3 (Newey and West, 1986).

West (1996) shows that under conventional assumptions this statistic is asymptotically normally distributed under the null hypothesis of equal forecast accuracy when the parameters of the forecast model are estimated by ordinary least squares. We therefore calculate and report 2-sided p-values for the TF benchmark using the standard normal distribution. However, as noted above in our discussion of the MSE-F test statistics, this asymptotic theory is based upon assumptions which seem untenable in the case of the VL-RT forecasts we construct; the p-values reported in this case should therefore be viewed as approximate and interpreted with caution.

4 Does the Output Gap Improve Forecasts of Inflation?

We now examine the results of the experiments described above. We begin in this section by examining the extent to which output gaps improve forecasts of inflation, comparing the conclusions reached using ex post and real time estimates of the gap.

4.1 Are Improvements in Forecast Accuracy Significant?

Table 2 shows the results of formal tests for differences in MSFE between the two benchmark models and the twelve output gap models. The upper panel of the table compares forecasts

constructed using final output data, final estimates of the output gap, and constant lag lengths in the forecasting equation (FL-FL). The middle panel of the table shows the comparable results when using quasi-final rather than final (i.e. filtered rather than smoothed) estimates of the output gap (FL-QF). Since such estimates can only be constructed from UC models of the output gap, only results for the five UC models are presented. In both cases, we see the MSFE of the benchmark models, the fractional improvement in MSFE relative to the benchmark models ($(MSFE_{Benchmark} - MSFE_{Gap})/MSFE_{Gap}$) and the p-value for the test of the null hypothesis that the MSFEs of the benchmark and the gap model are equal. Differences between these two panels are entirely due to the effects of *ex post* revisions of output gaps.

The first thing apparent from the top panel of the table is that all the gap models forecast better than the autoregressive benchmark model when using final output gaps. In all but one case the differences in MSFE are greater than 10 per cent, and in four of the twelve cases they are greater than 30 per cent. All the differences are statistically significant from zero at the 10 per cent level, all but the SVAR model are significant at the 5 per cent level, and seven of the twelve are significant at the one per cent level. This confirms the conventional wisdom that *ex post* output gaps appear to help forecast inflation. It also shows that out-of-sample tests have sufficient power to detect relevant differences in MSFE.

This evidence is weakened when the benchmark model is changed by adding real output growth to the forecasting equation (the TF model). As can be seen on the right side of the top panel, three of the twelve gap models now have larger MSFEs than the benchmark, and only five of the twelve show an improvement of more than 10 per cent. The differences in MSFE are significant at the 10 per cent level in only one case.¹³

¹³Comparison of the *significance* of the differences in MSFE across the two benchmarks is complicated by differences in the tests used for nested and non-nested models, as explained in section 3.3. Note, in particular, that the reported p-values for nested models (the AR benchmark) are based on *one-sided* tests, while those for non-nested models (the TF benchmark) are based on *two-sided* tests. In addition, Clark and McCracken (2001, 2003) suggest that the MSE-F statistic, which is used for the AR benchmark, is more powerful than the *z* statistic used for the TF benchmark.

The apparent superiority of output-gap based forecasts is also weakened by the use of quasi-final rather than final estimates of the gap, as the second panel of the Table shows. Improvements over the AR benchmark are now lower in every case, falling 10 to 20 per cent, and in one case output-gap-based forecasts are less accurate than the benchmark. However, improvements in forecast accuracy are still significant at or near the 5 per cent significance level in the four remaining cases. The situation changes still further if we instead use the TF benchmark. Four of the five models now forecast less accurately than the benchmark model and none of the differences in forecast accuracy are significant at even the 20 per cent level. This implies that ignoring the effects of output gap revisions will tend to overstate the importance and significance of the output gap's forecasting power for inflation.

The bottom panel of Table 2 shows the results of tests for differences in MSFE between the two benchmark models and the twelve output gap models. The forecasts are now constructed with time-varying lag lengths and real-time output gap estimates (VL-RT). This variation in data structure invalidates the usual asymptotic theory used to justify the distribution of the test statistics under the null. The p-values presented should therefore be interpreted with caution. This change also increases the MSFE of the benchmark AR model by a little over 10 per cent.

The relative accuracy of these real-time forecasts is almost always lower than that of the ex post forecasts analysed in the top panel of the Table. Drops in relative MSFE are over twenty per cent in 9 of the 24 cases, with all of the output gap models deteriorating relative to the TF benchmark. The increase relative to the AR benchmark shown in two cases is due to the AR benchmark's increase in MSFE. Taken at face value, the p-values show considerable changes and an overall weakening of the evidence of superior forecast performance. Only 4 of the 12 models now appear to have MSFEs which are significantly smaller than those of the AR model. Seven of the eleven models which appeared to forecast significantly more accurately than the AR benchmark in the top panel now show no

significant difference in accuracy. Even more striking is the reversal in the performance of the output gap models relative to the output growth (TF) benchmark, as can be seen by comparing the the top and bottom panels on the right-hand side of the table. In real time, *none* of the output gap models forecasts better than the TF benchmark and three (LT, QT and BN) appear to forecast significantly worse.

4.2 The Effect of Output Gap Revisions on Relative Forecast Accuracy

To better understand the causes for the changes in MSFE noted above, Table 3 compares the MSFEs of three different forecasting experiments. The first is identical to that documented in the upper panel of the previous table, using final output data and gap estimates as well as constant lag lengths in the forecasting equation (FL-FL). The second experiment uses the same output data and gap estimates, but now updates the lag lengths each time the forecast coefficients are recursively re-estimated (VL-FL). The third experiment is identical to that documented in the bottom panel of the previous table, using time-varying lag lengths and real-time output gap estimates (VL-RT). Differences in outcomes between the first two experiments isolate the effects of variations in lag length. Differences between the second two experiments similarly isolate the effects of output gap revision.

The table shows that the introduction of time-varying lag lengths has important effects on forecast accuracy. A priori, such time-variation may improve forecasts if the underlying relationship is unstable over time. On the other hand, it may introduce another source of estimation error, which could reduce forecast accuracy. The table shows that all forecasts see a reduction in accuracy, averaging 15 per cent. The benchmarks forecasts see changes in MSFEs which are very close to the average.

Moving from Final to real-time output gap estimates has no effect on the AR benchmark forecast, but tends to make forecasts less accurate. While the average effects of this change are smaller than those of changes in lag length, the impact varies much more across models. Four models see their accuracy improve while three see their MSFE rise by more than 20 per

cent. Note that the TF benchmark sees the greatest improvement in accuracy. Evidently, revisions in output growth contain useful information about future inflation.

The net effect of the changes in lag length determination and data vintage worsens forecast accuracy in all but one case. The net effect on the AR benchmark is somewhat less than average, while the TF benchmark improves more than any other model.

The results above suggest that:

- some output gap models forecast inflation more accurately than an autoregressive model, even when using real-time output gap estimates.
- none of the output gap models we examine forecasts inflation as well as simple models which use both past inflation and output growth.
- the relative performance of different models is greatly affected by the use of real-time rather than ex post output gap estimates. However, uncertainty about the lag structure also adds considerably to MSFEs.

4.3 The Robustness of Changes in Forecast Accuracy

To investigate the robustness of the results presented in Table 2, we now examine how these conclusions are altered as we change various feature of the forecasting experiment. Table 4 examines the effects of changing the period over which forecasts are evaluated. The full 1969-2002 sample is split into two roughly equal halves, with the 1969-83 portion characterized by relatively high and volatile inflation, whereas prices were more stable over the 1984-2002 period. The greater volatility of inflation in the former period imply that least-squares methods tend to emphasize the fit of the model over the former period. Perhaps as a consequence, the full-sample results presented in Table 2 largely reflect forecast performance over the first half of the sample. Results for the low-inflation period after 1983 may be more a relevant guide for contemporary decision-making, but they differ from the full-sample results in several ways.

First, looking at forecasts with final output gaps, we see that the AR benchmark has become harder to beat. Nine of the 12 models see their relative MSFEs decline, and only four can reject the null of equal forecast accuracy at the 5 per cent level (compared to 10 in the earlier portion of the sample.) This decline in the predictability of inflation has been noted previously in other studies.¹⁴ The picture for the TF benchmark is less clear; while the relative performance of the output gap models improves somewhat in the latter sample, there is little evidence of significantly different forecast accuracy.

Second, looking at forecasts with real-time output gaps, it appears that it has become increasingly difficult to forecast as well as the benchmarks. Out of 12 models 11 (10) have larger MSFEs than the AR (TF) benchmark. The Band-pass filter is the only model to forecast inflation better than either benchmark in the recent period, giving over a 20 per cent reduction in MSFE. It is also interesting to note that, consistent with the reported decline in the predictability of inflation, the AR benchmark now forecasts slightly better in real time than the TF benchmark.

One possible explanation for the difference in results across the two sample periods is parameter instability, a feature which has been noted by other research on inflation forecasts.¹⁵ Table 5 provides more evidence of such instability by showing the results of changes in the period over which the forecasting model is estimated. The first four columns start the estimation in 1947 rather than 1955. The next four columns start the estimation in 1965; to allow for a sufficiently long estimation sample, forecast results in this case are shown only for the 1984-2002 sample.

The first set of columns show a marked deterioration in relative forecast performance using ex post data, regardless of the benchmark model we use. This occurs despite an important increase in MSFE of both benchmark models, implying that the inclusion of this early period causes important problems for the estimation of the forecasting relationship.

¹⁴For example, see Atkeson and Ohanian (2001) or Fisher, Liu and Zhou (2002).

¹⁵See, in particular, Stock and Watson (1996, 1999).

Using real-time data, the AR benchmark is now much more difficult to beat, while the TF benchmark continues to perform better than most models. Starting the estimation later, we obtain results for the 2nd half of the sample which closely resemble those from Table 4 for the same period, suggesting that parameter instability may have become less serious over time. The BN model appears to do particularly well using the longer estimation period, while the BP model again appears to do well over the more recent forecast period.

Table 6 shows how the results change when the forecast horizon is increased to 8 quarters from 4. The relative performance of the output gap models improve over the full sample, regardless of whether we use the AR or the TF benchmark, and regardless of whether we use final or real-time output gaps. Improvements in MSFE in real-time data are particularly striking, exceeding 25 per cent for 7 of the 12 output gap models when using the AR benchmark. However, forecasts from the latter part of the sample show very different results. Only 4 of the 12 models forecast more accurately than the benchmark using final data (2 of 5 using quasi-final gaps.) Results are similar using real-time data, where output gap models are more accurate in 5 of 12 cases using the AR benchmark (3 of 12 using the TF benchmark.) There is no evidence to suggest that any of the improvements in real-time forecast performance are statistically significant. The good real-time performance of the BN and BP models noted above vanishes, with both models performing consistently worse than either benchmarks, perhaps significantly so.

A number of other variants were analysed, but these gave similar conclusions, particularly for the relative performance with real-time data. In addition to examining more changes in forecast horizons and estimation periods, we also examined sensitivity to forecasting changes rather than the level of inflation, the use of different lag lengths and other benchmarks. (Detailed results on these sensitivity checks are available on request.) Based on a review of these results, it appears that the results shown in Table 2 are among the *best* results that can be obtained for inflation forecasts from simple linear forecasting models

using output gaps.

Having considered the above results, one might also ask which of the output gap models examined here a practitioner should use to forecast inflation (if forced to do so.) This task is complicated by difficulty of drawing inferences about real-time forecast accuracy, which prevent us from reaching firm conclusions. It would appear that the deterministic trend models (LT, QT and BT) were often among the worst-performing in real-time, and should probably be avoided for that reason. UC models which estimated Phillips Curves (KT and GS) had some of the largest differences in performance when used with real-time rather than final estimates. The Band-Pass (and to a lesser extent, the Beveridge-Nelson) performed reasonably well in our simulated real-time experiments, despite showing little promise on the basis of ex post analysis. However, their success appears to be sensitive to the forecast horizon used. Rather than rely on any of these output gap models, our analysis suggests that a practitioner could do well by simply taking into account the information contained in real output without attempting to measure the level of the output gap—the TOFU model. This model was consistently among the best performers, particularly over the post-1983 forecast sample.

5 Conclusion

Forecasting inflation is a difficult but essential task for the successful implementation of monetary policy. The hypothesis that a stable predictive relationship between inflation and the output gap—a Phillips curve—is present in the data, suggests that output gap measures could be useful for forecasting inflation. This has served as the basis for empirical formulations of countercyclical monetary policy in many models. We find that many alternative measures of the output gap *appear* to be quite useful for forecasting inflation, on the basis of ex post analysis. That is, a historical Phillips curve is suggested by the data, and final (constructed ex post) estimates of the output gap are useful for understanding subsequent

movements in inflation.

However, this historical usefulness does not imply a similar operational usefulness. Our simulated real-time forecasting experiment suggests, instead, that the predictive ability of many different output gap measures may be illusory. Output gaps typically can not forecast inflation as well out of sample as simple linear models of inflation and output growth (although the differences are rarely statistically significant.) This is particularly true if we restrict our attention to the post-1983 period. These rather pessimistic findings regarding the output gap mirror earlier investigations regarding the predictive power for forecasting inflation of “unemployment gaps,” that is the difference between the rate of unemployment and estimates of the NAIRU. As demonstrated by Staiger, Stock and Watson (1997a,b) and Stock and Watson (1999), estimates of the NAIRU are inherently unreliable, and simulated out-of-sample forecasting exercises do not indicate a robust improvement in inflation forecasts from using information about unemployment. Stock and Watson (1999) also show that better inflation forecasts may be obtained by indicators other than the unemployment gap. Our analysis suggests this appears to be the case with the output gap as well. Instead of output gaps, forecasts of inflation which simply incorporate information from the growth rate of output appear to forecast inflation as well or better.

Finally, we note that these negative findings regarding the usefulness of real-time measures of the output gap do not necessarily invalidate the potential usefulness of the theoretical Phillips curve framework per se, nor that of ex post constructed output gaps for historical analysis. That said, the dubious contribution of real-time measures of the output gap for forecasting inflation brings into question their role in the formulation of reliable real-time policy analysis.

References

- Atkeson, Andrew and Lee E. Ohanian, "Are Phillips Curves Useful for Forecasting Inflation," *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1), 2-11, Winter 2001.
- Baxter, Marianne; King, Robert G., "Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series" National Bureau of Economic Research Working Paper: 5022, 1995.
- Beveridge, S and C. R. Nelson, "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle'," *Journal of Monetary Economics*, 7, 151-174, 1981.
- Blanchard. Olivier and Danny Quah, "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79(4), 655-673, September, 1989.
- Bryant, Ralph C., Peter Hooper and Catherine Mann eds. *Evaluating Policy Regimes: New Research in Empirical Macroeconomics*, Brookings: Washington DC, 1993.
- Cayen, Jean-Philippe "Fiabilité des estimations de l' écart de production au canada.? École des Hautes Études Commerciales (Montréal), memoire, 2001.
- Clark, Peter K., "The Cyclical Component of U.S. Economic Activity," *Quarterly Journal of Economics* 102(4), 1987, 797-814.
- Clark, Todd E. and Michael W. McCracken, "Tests of Equal Forecast Accuracy and Encompassing for Nested Models" *Journal of Econometrics*, 105, 85-110, 2001.
- Clark, Todd E. and Michael W. McCracken, "Evaluating Long-Horizon Forecasts," Federal Reserve Bank of Kansas City, mimeo, 2002.
- Croushore, Dean and Tom Stark, "A Real-Time Data Set for Macroeconomists," *Journal of Econometrics*, 105, 111-130, November, 2001.
- Diebold, Francis X. and Roberto S. Mariano, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 1995, 253-265.
- Fisher, Jonas D. M., Chin Te Liu, and Ruilin Zhou, "When Can we Forecast Inflation?" Federal Reserve Bank of Chicago *Economic Perspectives*, 1Q/2002, 30-42, 2002.
- Gerlach, Stefan and Frank Smets, "Output Gaps and Inflation: Unobservable-Components Estimates for the G-7 Countries." Bank for International Settlements mimeo, Basel 1997.
- Harvey, Andrew C., "Trends and Cycles in Macroeconomic Time Series," *Journal of Business and Economic Statistics*, 3, 216-227, 1985.
- Hodrick, R, and E. Prescott, "Post-war Business Cycles: An Empirical Investigation," *Journal of Money, Credit, and Banking*, 29, 1997, 1-16.
- Koenig, Evan F., Sheila Dolmas and Jeremy Piger, "The Use and Abuse of 'Real-Time' Data in Economic Forecasting," forthcoming, *Review of Economics and Statistics*, 2003.

- Kuttner, Kenneth N., "Estimating Potential Output as a Latent Variable," *Journal of Business and Economic Statistics*, 12(3), 1994, 361-68.
- McCracken, Michael W., "Asymptotics for Out-of-Sample Causality" *University of Missouri mimeo* 2000.
- Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55(3), 703-08, May 1987.
- Orphanides, Athanasios and Simon van Norden, "The Reliability of Output Gap Estimates in Real Time," Finance and Economics Discussion Series 1999-38, August 1999.
- Orphanides, Athanasios and Simon van Norden, "The Unreliability of Output Gap Estimates in Real Time," *Review of Economics and Statistics*, 84(4), 569-583, November 2002.
- Orphanides, Athanasios and Simon van Norden, "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time, *CIRANO working paper* 2003s-01, 49 p.
- St-Amant, Pierre and Simon van Norden, "Measurement of the Output Gap: A discussion of recent research at the Bank of Canada," Bank of Canada Technical Report No. 79, 1998.
- Staiger, Douglas, James H. Stock, and Mark W. Watson, "How Precise are Estimates of the Natural Rate of Unemployment?" in Romer, Christina and David Romer, eds. *Reducing Inflation: Motivation and Strategy*, Chicago: University of Chicago Press, 1997a.
- Staiger, Douglas, James H. Stock, and Mark W. Watson, "The NAIRU, Unemployment and Monetary Policy," *Journal of Economic Perspectives* 11(1), Winter 1997b, 33-49.
- Stock, James H. and Mark W. Watson, "Evidence on Structural Instability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics*, 14(1), 11-30, January, 1996.
- Stock and Watson "Business Cycle Fluctuations in U.S. Macroeconomic Time Series." *NBER Working Paper* No. 6528, 1998, 83 p., prepared for *The Handbook of Macroeconomics*, edited by John B. Taylor and Michael Woodford.
- Stock, James H. and Mark W. Watson, "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293-335, 1999.
- Taylor, John B., *Monetary Policy Rules*, Chicago: University of Chicago, 1999.
- van Norden, Simon, "Why is it so hard to measure the current output gap?" Bank of Canada mimeo, 1995.
- West, Kenneth D. "Asymptotic Inference About Predictive Ability." *Econometrica*, 64, 1996, 1067-84.

Table 1

Description of Alternative Output Gap Measures and Summary Reliability Statistics

Method	Data	Method Details	COR	AR	NSR	OPSIGN
Linear Trend	Univariate.		0.88	0.90	1.63	0.58
Quadratic Trend	Univariate.		0.51	0.97	1.06	0.42
Breaking Trend	Univariate.	Trend Break in 1973Q1, starting in 1977Q1.	0.77	0.87	0.81	0.28
Hodrick-Prescott	Univariate.	With $\lambda = 1600$.	0.52	0.93	1.05	0.45
Band Pass	Univariate.	6–32 quarters, series padded with AR forecasts.	0.72	0.77	0.77	0.36
Beveridge-Nelson	Univariate.	Assumes ARIMA(1,1,2).	0.84	0.09	0.63	0.30
Structural VAR	Trivariate.	Imposes long-run restrictions.	0.68	0.85	0.95	0.41
Watson	Univariate.	Local Level and AR(2).	0.88	0.87	1.50	0.55
Harvey-Clark	Univariate.	Local Linear Trend and AR(2).	0.75	0.92	0.91	0.39
Harvey-Jaeger	Univariate.	Local Linear Trend and Cycle.	0.56	0.86	0.92	0.50
Kuttner	Bivariate.	Watson model and inflation equation.	0.87	0.90	1.54	0.61
Gerlach-Smets	Bivariate.	Harvey-Clark model and inflation equation.	0.79	0.82	1.05	0.40

Notes: Univariate methods employ only real GNP/GDP data. Bivariate also employ CPI inflation. Trivariate also employs treasury bill data. The last four columns present summary measures of the reliability of real-time estimates of the output gap. All statistics are for the 1969:1–2003:1 period. **COR** denotes the correlation of the real-time and final estimates of the output gap, **AR** the first order serial correlation of the revision (the difference between the final and real-time series), **NSR** indicates the ratio of the root mean square of the revision and the standard deviation of the final estimate of the gap, and **OPSIGN** indicates the frequency with which the real-time and final gap estimates have opposite signs.

Table 2
Relative Improvement in MSFE

Method	AR	AR p-value	TF	TF p-value
<i>Fixed Lags, Final Gaps</i>				
Benchmark MSFE	0.494		0.436	
Linear	0.302	0.011	0.148	0.215
Quadratic	0.168	0.025	0.030	0.793
Breaking	0.106	0.048	-0.024	0.792
Hodrick-Prescott	0.149	0.002	0.013	0.891
Band-Pass	0.134	0.001	0.000	0.997
Beveridge-Nelson	0.139	0.003	0.004	0.418
SVAR	0.047	0.090	-0.077	0.502
Watson	0.319	0.004	0.163	0.115
Harvey-Clark	0.270	0.004	0.120	0.224
Harvey-Jaeger	0.109	0.003	-0.022	0.789
Kuttner	0.336	0.011	0.178	0.126
Gerlach-Smets	0.362	0.000	0.201	0.072
<i>Fixed Lags, Quasi-Final Gaps</i>				
Watson	0.132	0.056	-0.002	0.979
Harvey-Clark	0.070	0.065	-0.056	0.324
Harvey-Jaeger	-0.032	0.503	-0.146	0.302
Kuttner	0.248	0.030	0.100	0.323
Gerlach-Smets	0.091	0.046	-0.038	0.390
<i>Variable Lags, Real-time Gaps</i>				
Benchmark MSFE	0.559		0.416	
Linear	0.045	0.208	-0.219	0.044
Quadratic	0.021	0.249	-0.237	0.032
Breaking	0.043	0.184	-0.221	0.139
Hodrick-Prescott	0.132	0.016	-0.154	0.370
Band-Pass	0.283	0.002	-0.042	0.763
Beveridge-Nelson	0.211	0.005	-0.095	0.032
SVAR	-0.093	0.806	-0.323	0.057
Watson	0.121	0.072	-0.163	0.131
Harvey-Clark	0.147	0.032	-0.143	0.175
Harvey-Jaeger	0.080	0.087	-0.193	0.282
Kuttner	0.107	0.119	-0.173	0.100
Gerlach-Smets	0.099	0.057	-0.179	0.118

Notes: The AR benchmark is a univariate autoregressive forecast of inflation; the TF benchmark forecasts from a linear regression on lagged inflation and real output growth. Mean squared forecast errors (MSFE) for the two benchmark models are shown multiplied by 1000. The remaining figures in the AR and TF columns denote the relative improvements in MSFE for the output gap models, measured as $(A - B)/B$ where A is the MSFE of the benchmark and B is that of the output gap model. The p-values for the AR benchmark are for the null that $B \geq A$, based on the statistic in equation (5). The p-values shown for the TF benchmark are for two-sided test of the null that $A = B$, based on the statistic in equation (6). See section 3.3 and Appendix B for further discussion of the construction and interpretation of the p-values. The forecast horizon is 4 quarters and forecast performance is evaluated over the period from 1969Q1 to 2002Q2. Forecast equation estimation starts in 1955Q1. Fixed lag lengths are (1,1) while varying lag lengths are reset every quarter using BIC.

Table 3
The Effect of Lag Selection and Data Vintage

Method	MSFE			Change in MSFE (percent)		
	FL-FL	VL-FL	VL-RT	FL to VL	FL to RT	Total
AR benchmark	0.494	0.559	0.559	-13.0	0.0	-13.0
TF benchmark	0.436	0.496	0.416	-13.7	16.0	4.6
Linear	0.380	0.438	0.533	-15.4	-21.7	-40.4
Quadratic	0.423	0.500	0.545	-18.1	-9.0	-28.8
Breaking	0.447	0.494	0.534	-10.6	-8.0	-19.5
Hodrick-Prescott	0.430	0.556	0.492	-29.2	11.5	-14.4
Band-Pass	0.436	0.502	0.434	-15.2	13.5	0.4
Beveridge-Nelson	0.434	0.482	0.460	-11.0	4.5	-6.0
SVAR	0.472	0.502	0.614	-6.4	-22.3	-30.1
Watson	0.375	0.433	0.497	-15.4	-14.9	-32.6
Harvey-Clark	0.389	0.448	0.486	-15.1	-8.4	-24.7
Harvey-Jaeger	0.446	0.577	0.516	-29.5	10.7	-15.7
Kuttner	0.370	0.402	0.503	-8.5	-25.3	-36.0
Gerlach-Smets	0.363	0.426	0.507	-17.3	-19.0	-39.6
Mean				-15.6	-5.2	-21.1
Std Dev				6.7	14.5	14.4

Notes:

MSFE denotes the mean squared forecast error (shown multiplied by 1000.)

FL-FL refers to forecasts using fixed lag lengths and final output gap estimates.

VL-FL refers to forecasts using variable lag lengths and final output gap estimates.

VL-RT refers to forecasts using variable lag lengths and real-time output gap estimates.

FL to VL refers to the change from FL-FL to VL-FL.

FL to RT refers to the change from VL-FL to VL-RT.

Total refers to the change from FL-FL to VL-RT.

Table 4
Relative Improvement in MSFE: Sub-sample Evaluation

Method	1969Q1–1983Q4				1984Q1–2002Q2			
	AR	p-value	TF	p-value	AR	p-value	TF	p-value
<i>Fixed Lags, Final Gaps</i>								
Benchmark MSFE	0.863		0.739		0.191		0.187	
Linear	0.247	0.046	0.068	0.600	0.555	0.007	0.517	0.046
Quadratic	0.194	0.046	0.023	0.849	0.079	0.133	0.054	0.838
Breaking	0.120	0.081	−0.041	0.674	0.060	0.140	0.035	0.865
Hodrick-Prescott	0.178	0.010	0.009	0.929	0.051	0.101	0.025	0.881
Band-Pass	0.172	0.009	0.004	0.968	0.013	0.198	−0.011	0.944
Beveridge-Nelson	0.174	0.012	0.005	0.414	0.025	0.195	0.000	0.952
SVAR	0.013	0.319	−0.133	0.290	0.199	0.012	0.170	0.361
Watson	0.331	0.018	0.140	0.228	0.277	0.032	0.247	0.245
Harvey-Clark	0.320	0.008	0.131	0.223	0.113	0.085	0.086	0.683
Harvey-Jaeger	0.140	0.009	−0.024	0.800	0.006	0.213	−0.018	0.883
Kuttner	0.317	0.027	0.128	0.326	0.411	0.022	0.377	0.099
Gerlach-Smets	0.432	0.003	0.226	0.068	0.154	0.059	0.126	0.546
<i>Fixed Lags, Quasi-Final Gaps</i>								
Watson	0.091	0.158	−0.065	0.404	0.311	0.025	0.280	0.030
Harvey-Clark	0.081	0.107	−0.074	0.232	0.032	0.173	0.007	0.931
Harvey-Jaeger	0.252	0.016	0.072	0.491	−0.474	0.996	−0.487	0.074
Kuttner	0.194	0.082	0.023	0.835	0.494	0.018	0.458	0.042
Gerlach-Smets	0.115	0.073	−0.045	0.339	0.010	0.277	−0.015	0.865
<i>Variable Lags, Real-time Gaps</i>								
Benchmark MSFE	1.010		0.689		0.191		0.196	
Linear	0.225	0.069	−0.165	0.208	−0.357	0.932	−0.341	0.054
Quadratic	0.228	0.047	−0.163	0.144	−0.405	0.977	−0.390	0.085
Breaking	0.172	0.070	−0.201	0.269	−0.289	0.915	−0.272	0.136
Hodrick-Prescott	0.508	0.001	0.028	0.869	−0.451	0.998	−0.438	0.154
Band-Pass	0.301	0.004	−0.113	0.460	0.215	0.016	0.244	0.172
Beveridge-Nelson	0.288	0.009	−0.122	0.017	−0.035	0.423	−0.011	0.880
SVAR	−0.106	0.839	−0.391	0.027	−0.018	0.375	0.006	0.964
Watson	0.209	0.055	−0.176	0.173	−0.144	0.657	−0.123	0.383
Harvey-Clark	0.205	0.039	−0.179	0.112	−0.046	0.439	−0.023	0.912
Harvey-Jaeger	0.445	0.009	−0.015	0.930	−0.480	0.997	−0.468	0.154
Kuttner	0.205	0.088	−0.179	0.159	−0.177	0.704	−0.158	0.231
Gerlach-Smets	0.153	0.051	−0.214	0.083	−0.081	0.582	−0.059	0.766

Notes: The AR benchmark is a univariate autoregressive forecast of inflation; the TF benchmark forecasts from a linear regression on lagged inflation and real output growth. Mean squared forecast errors (MSFE) for the two benchmark models are shown multiplied by 1000. The remaining figures in the AR and TF columns denote the relative improvements in MSFE for the output gap models, measured as $(A - B)/B$ where A is the MSFE of the benchmark and B is that of the output gap model. The p-values for the AR benchmark are for the null that $B \geq A$, based on the statistic in equation (5). The p-values shown for the TF benchmark are for two-sided test of the null that $A = B$, based on the statistic in equation (6). See section 3.3 and Appendix B for further discussion of the construction and interpretation of the p-values. The forecast horizon is 4 quarters and forecast equation estimation starts in 1955Q1. Fixed lag lengths are (1,1) while varying lag lengths are reset every quarter using BIC.

Table 5
Relative Improvement in MSFE: Alternative Estimation Starting Dates

Method	Estimation start in 1947Q1 (Full sample evaluation)				Estimation start in 1965Q1 (1984Q1–2002Q2 evaluation)			
	AR	p-value	TF	p-value	AR	p-value	TF	p-value
<i>Fixed Lags, Final Gaps</i>								
Benchmark MSFE	0.665		0.646		0.293		0.249	
Linear	0.097	0.117	0.064	0.324	1.258	0.000	0.918	0.017
Quadratic	0.014	0.259	−0.016	0.725	0.467	0.007	0.246	0.338
Breaking	0.019	0.208	−0.011	0.378	0.280	0.017	0.087	0.534
Hodrick-Prescott	0.006	0.235	−0.024	0.519	0.060	0.091	−0.099	0.507
Band-Pass	0.012	0.177	−0.018	0.531	0.052	0.093	−0.106	0.445
Beveridge-Nelson	0.578	0.000	0.531	0.073	0.184	0.017	0.006	0.352
SVAR	−0.080	0.722	−0.107	0.406	0.165	0.028	−0.010	0.951
Watson	0.052	0.140	0.021	0.613	0.844	0.000	0.567	0.012
Harvey-Clark	0.045	0.127	0.014	0.625	0.553	0.006	0.320	0.102
Harvey-Jaeger	−0.017	0.536	−0.046	0.363	−0.025	0.458	−0.172	0.210
Kuttner	0.103	0.103	0.070	0.307	1.258	0.001	0.919	0.018
Gerlach-Smets	0.071	0.072	0.039	0.221	0.571	0.003	0.335	0.076
<i>Fixed Lags, Quasi-Final Gaps</i>								
Watson	−0.020	0.450	−0.049	0.197	0.441	0.011	0.225	0.003
Harvey-Clark	−0.048	0.614	−0.076	0.029	0.106	0.084	−0.060	0.377
Harvey-Jaeger	−0.066	0.723	−0.094	0.105	−0.513	0.997	−0.586	0.087
Kuttner	0.129	0.079	0.096	0.418	1.093	0.003	0.779	0.000
Gerlach-Smets	0.054	0.093	0.023	0.703	0.138	0.058	−0.033	0.539
<i>Variable Lags, Real-time Gaps</i>								
Benchmark MSFE	0.610		0.550		0.294		0.287	
Linear	−0.087	0.641	−0.178	0.000	−0.476	0.987	−0.489	0.190
Quadratic	−0.053	0.556	−0.147	0.001	−0.272	0.879	−0.290	0.029
Breaking	0.013	0.280	−0.088	0.049	−0.347	0.955	−0.362	0.023
Hodrick-Prescott	−0.020	0.479	−0.117	0.091	−0.237	0.934	−0.255	0.087
Band-Pass	0.071	0.064	−0.035	0.586	0.275	0.010	0.245	0.035
Beveridge-Nelson	0.327	0.002	0.196	0.316	0.033	0.170	0.009	0.520
SVAR	−0.005	0.346	−0.104	0.651	−0.070	0.561	−0.092	0.531
Watson	−0.126	0.781	−0.212	0.001	−0.022	0.360	−0.045	0.759
Harvey-Clark	−0.059	0.640	−0.153	0.021	0.000	0.299	−0.024	0.869
Harvey-Jaeger	−0.053	0.590	−0.147	0.055	−0.262	0.914	−0.279	0.085
Kuttner	0.109	0.119	−0.001	0.989	−0.388	0.950	−0.402	0.200
Gerlach-Smets	0.183	0.013	0.066	0.438	−0.090	0.552	−0.112	0.731

Notes: The AR benchmark is a univariate autoregressive forecast of inflation; the TF benchmark forecasts from a linear regression on lagged inflation and real output growth. Mean squared forecast errors (MSFE) for the two benchmark models are shown multiplied by 1000. The remaining figures in the AR and TF columns denote the relative improvements in MSFE for the output gap models, measured as $(A - B)/B$ where A is the MSFE of the benchmark and B is that of the output gap model. The p-values for the AR benchmark are for the null that $B \geq A$, based on the statistic in equation (5). The p-values shown for the TF benchmark are for two-sided test of the null that $A = B$, based on the statistic in equation (6). See section 3.3 and Appendix B for further discussion of the construction and interpretation of the p-values. The forecast horizon is 4 quarters and full-sample forecast performance is evaluated over the period from 1969Q1 to 2002Q2. Fixed lag lengths are (1,1) while varying lag lengths are reset every quarter using BIC.

Table 6
Relative Improvement in MSFE: 8-Quarter ahead forecasts

Method	Full Sample Evaluation				1984Q1-2002Q2 Evaluation			
	AR	p-value	TF	p-value	AR	p-value	TF	p-value
<i>Fixed Lags, Final Gaps</i>								
Benchmark MSFE	3.070		2.650		0.648		0.626	
Linear	0.609	0.008	0.389	0.034	1.108	0.005	1.036	0.016
Quadratic	0.381	0.010	0.192	0.164	-0.108	0.524	-0.138	0.671
Breaking	0.202	0.027	0.038	0.654	-0.137	0.618	-0.166	0.479
Hodrick-Prescott	0.035	0.105	-0.107	0.554	-0.143	0.705	-0.172	0.235
Band-Pass	0.043	0.082	-0.100	0.544	-0.177	0.801	-0.205	0.130
Beveridge-Nelson	0.160	0.005	0.001	0.870	0.039	0.134	0.003	0.762
SVAR	0.267	0.005	0.094	0.433	-0.107	0.516	-0.137	0.576
Watson	0.515	0.003	0.308	0.008	0.148	0.122	0.109	0.643
Harvey-Clark	0.407	0.004	0.214	0.029	-0.118	0.482	-0.148	0.510
Harvey-Jaeger	-0.006	0.329	-0.142	0.390	-0.171	0.845	-0.199	0.057
Kuttner	0.612	0.005	0.392	0.034	0.457	0.043	0.407	0.079
Gerlach-Smets	0.461	0.003	0.261	0.013	-0.087	0.393	-0.118	0.591
<i>Fixed Lags, Quasi-Final Gaps</i>								
Watson	0.268	0.032	0.094	0.266	0.856	0.005	0.793	0.023
Harvey-Clark	0.095	0.095	-0.055	0.383	-0.004	0.296	-0.038	0.749
Harvey-Jaeger	0.014	0.242	-0.125	0.481	-0.647	0.997	-0.659	0.073
Kuttner	0.492	0.017	0.288	0.057	0.940	0.014	0.874	0.039
Gerlach-Smets	0.124	0.067	-0.030	0.601	-0.111	0.580	-0.141	0.133
<i>Variable Lags, Real-time Gaps</i>								
Benchmark MSFE	4.050		3.210		1.620		1.570	
Linear	0.294	0.067	0.026	0.839	-0.257	0.653	-0.280	0.241
Quadratic	0.083	0.185	-0.142	0.086	-0.398	0.854	-0.416	0.001
Breaking	0.015	0.291	-0.195	0.044	-0.198	0.659	-0.223	0.024
Hodrick-Prescott	-0.094	0.733	-0.282	0.005	-0.100	0.607	-0.128	0.387
Band-Pass	-0.045	0.603	-0.243	0.025	-0.036	0.471	-0.065	0.428
Beveridge-Nelson	0.286	0.008	0.019	0.730	0.025	0.244	-0.006	0.904
SVAR	0.119	0.087	-0.113	0.252	-0.202	0.690	-0.227	0.018
Watson	0.457	0.024	0.155	0.141	0.133	0.186	0.098	0.684
Harvey-Clark	0.355	0.018	0.074	0.388	0.191	0.090	0.154	0.376
Harvey-Jaeger	0.446	0.006	0.146	0.347	-0.080	0.536	-0.108	0.609
Kuttner	0.421	0.034	0.126	0.220	0.066	0.253	0.033	0.890
Gerlach-Smets	0.341	0.022	0.063	0.461	0.165	0.109	0.129	0.469

Notes: The AR benchmark is a univariate autoregressive forecast of inflation; the TF benchmark forecasts from a linear regression on lagged inflation and real output growth. Mean squared forecast errors (MSFE) for the two benchmark models are shown multiplied by 1000. The remaining figures in the AR and TF columns denote the relative improvements in MSFE for the output gap models, measured as $(A - B)/B$ where A is the MSFE of the benchmark and B is that of the output gap model. The p-values for the AR benchmark are for the null that $B \geq A$, based on the statistic in equation (5). The p-values shown for the TF benchmark are for two-sided test of the null that $A = B$, based on the statistic in equation (6). See section 3.3 and Appendix B for further discussion of the construction and interpretation of the p-values. Forecast equation estimation starts in 1955Q1. Fixed lag lengths are (1,1) while varying lag lengths are reset every quarter using BIC.

Appendix A: The Construction of Real Time Output Gaps

The output gaps used in this study, as well as the data and programs used to create them, are freely available from the authors. The estimates examined here include all those examined in Orphanides and van Norden (2002) plus the Band-Pass, Beveridge-Nelson, Harvey-Jaeger and SVAR methods described below; this is identical to the list of models considered in Orphanides and van Norden (2003). The range of available estimates were updated so that the "final" data vintage now corresponds to 2003Q3 (i.e. data available as of mid-August 2003, so data series end in 2003Q2) rather than 2000Q1 as in these two earlier papers. Data for real output were taken from the Philadelphia Federal Reserve Board's Real Time Data Archive in September 2003. Observations span the period from 1947Q1 to 2003Q2. Vintages for output run from Nov. 1965 to August 2003. All CPI data are from the 2003Q3 vintage. The SVAR method also uses data from January 1934 to August 2003 on secondary market yields on 3-month US treasury bills taken from the St. Louis Federal Reserve Board's FRED data base.

All output gap models we consider decompose the logarithm of output into trend and cycle components. The linear trend (LT) and quadratic trend (QT) models are from OLS regressions with linear and quadratic deterministic trends. The breaking trend model is identical to the LT model until 1976Q4. Starting in 1977Q1, it allows for an estimated break in the trend at the end of 1973. The Hodrick-Prescott(HP) method is based on the filter proposed by Hodrick and Prescott (1997) with their recommended smoothing parameter of 1600 for quarterly data. The band-pass method (BP) is based on the Stock and Watson (1998) adaptation of the Baxter and King (1995) approach. Following Stock and Watson (1998), we use a filter 25 observations in width and pad the available observations with forecasts from an AR(4) model. The Beveridge-Nelson follows Beveridge and Nelson (1981) in modelling output as an ARIMA(p,1,Q) series. Based on results for the full sample, we use an ARIMA(1,1,2), with parameters re-estimated by maximum likelihood methods before

each recalculation of the trend.

We examine five unobserved component (UC) models, all of which are estimated by maximum likelihood. Three of the five are univariate models. The Watson (WT) model is based on Watson (1986) and models the output trend as a random walk with drift while the cycle is assumed to follow a stationary AR(2) process. The Harvey-Clark (CL) model follows Harvey (1985) and Clark (1987), replacing the constant drift in the trend of the WT model with a random walk. The Harvey-Jaeger (HJ) model has the same trend as the CL model but replaces the AR(2) component with a stochastic cycle. All three of these univariate models require estimation of five parameters, including variances for the assumed Gaussian shocks. The Kuttner (KT) model appends a Phillips curve, as specified in Kuttner (1994), to the WT model, giving a bivariate model with eight more estimated parameters than its univariate counterpart. The Gerlach-Smets (GS) model similarly adds the Phillips curve specified in Gerlach and Smets (1997) to the CL model, yielding a bivariate model with six more estimated parameters than its univariate counterpart.

The Structural VAR measure of the output gap (BQ) is based on a VAR identified via restrictions on the long-run effects of the structural shocks, as proposed by Blanchard and Quah (1989). Our implementation is identical to that of Cayen and van Norden (2002), who use a trivariate system including output, CPI and yields on 3-month treasury bills. Lag lengths for the VAR are selected using finite-sample corrected LR tests and a general-to-specific testing approach.

Appendix B: Evaluation of Forecast Performance

As noted in section 3.3, our statistical inference for the forecast performance of the output gap models relative to the AR benchmark model is based on the MSE-F statistic proposed by McCracken (2000). This takes the form

$$MSE-F = P \cdot \frac{(MSFE_1 - MSFE_2)}{MSFE_2} \quad (\text{B.1})$$

where P is the number of forecasts, $MSFE_1$ is the mean squared forecast error (MSFE) of the restricted model and $MSFE_2$ is the MSFE of the unrestricted model.

The distribution of the MSE-F statistic under the null hypothesis of equal MSFE is estimated via a bootstrap experiment with 2000 replications. It begins by estimating a constrained VAR(12) in $\pi_{t+h}^h, \pi_{t-i}^1, y_t^{T, T-1}$ or $\pi_{t+h}^h, \pi_{t-i}^1, y_t^{T, t}$ in which we impose the restriction that y does not Granger-Cause π . 2000 simulated realizations of this DGP are created by simulating the estimated model with shocks randomly drawn with replacement from the estimated residuals. For each simulation, the dynamic model is initialized with historical observations starting with $\pi_{k+h}^h, \pi_{k-i}^1, y_k^{T, T-1}$ or $\pi_{k+h}^h, \pi_{k-i}^1, y_k^{T, k}$ for an independently drawn value of k . MSE-F statistics are then calculated for each simulated series and their empirical distribution is used to estimate p-values for the true data's MSE-F statistics. Because these distributions are non-pivotal, the distribution of the test statistics is bootstrapped anew for every different choice of (P, h, y, m, n) so that p-values for every reported MSE-F are based on independent bootstrap experiments. The single exception to this rule was that variations in the starting date of the OLS estimation of the forecasting equation were ignored; all bootstrap simulation used the base case starting date of 1955Q1.