The Output Gap: Method Choice, Data Revisions, and Policy Implications

Labesh Kumar
University of Wisconsin-Milwaukee *

March 25, 2026

Abstract

This paper compares eight widely used methods for estimating the output gap, ranging from simple deterministic trends to state space models, using both revised and real time U.S. quarterly data from 1980 onward. The resulting measures differ heavily across approaches. Average gap estimates vary by nearly four percentage points, volatility differs by an order of magnitude, and correlations across methods span from strongly positive to negative. Stability across data vintages also varies substantially. Hamilton type filters show relatively strong agreement between real time and final estimates, while simpler trend based methods are considerably less stable. These differences matter for empirical inference. The choice of output gap measure has important implications for Phillips curve estimates and for forecasting performance. Beveridge Nelson decompositions display strong predictive power for inflation when estimated using revised data but perform less well in real time, whereas refined Beveridge Nelson and modified Hamilton filters deliver more consistent results across vintages. Time varying analysis shows that the relationship between economic slack and inflation strengthens during periods of macroeconomic stress, including the early 1990s recession, the global financial crisis, and the post-pandemic period, rather than declining monotonically. For output growth forecasting, HP filter gaps reduce forecast errors using revised data, while unobserved components models perform best in real time. Although Beveridge Nelson based measures are informative for inflation, they tend to worsen growth forecasts. Combining forecasts across gap measures, particularly using Bates Granger weights, yields more reliable performance by offsetting weaknesses of individual methods. Overall, the findings highlight that methodological uncertainty in measuring slack translates directly into policy uncertainty, cautioning against exclusive reliance on any single output gap estimate.

Keywords: Output gap, trend-cycle decomposition, real-time data, Phillips curve, forecast combination

JEL classification: C52, E32, E37

*University of Wisconsin - Milwaukee, Bolton Hall 8th floor, 3210 N Maryland Ave, Milwaukee, WI 53211, USA.

1 Introduction

What is the true state of the U.S. economy relative to its potential? This deceptively simple question has fueled decades of macroeconomic debate and remains a persistent challenge for policymakers. The output gap, a measure of economic slack defined as the percentage deviation of actual output from its potential level, is a cornerstone concept in policy analysis. It underpins inflation forecasts via the Phillips curve, informs estimates of structural budget balances, and guides the timing of monetary tightening or easing. Yet this critical gauge of the business cycle is unobservable and must be extracted from the data. Economists have proposed a wide range of statistical filters and model-based decompositions for this task, ranging from simple deterministic trends to sophisticated state-space models. Each method makes different assumptions about the nature of economic growth and fluctuations, often yielding starkly different real-time assessments of economic slack. This uncertainty is not just academic as it translates directly into uncertainty about the appropriate policy stance. To complicate matters further, policymakers operate in real time, making decisions based on preliminary data that are subject to significant revision. In this landscape of methodological diversity and data revision, which approach to measuring the output gap proves most reliable for real-time analysis and forecasting?

To answer this question, this paper offers a systematic evaluation of alternative output gap estimation methods, examining their statistical properties, their predictive content for inflation, and their utility for forecasting economic activity. I compare eight distinct decomposition approaches ranging from simple deterministic trends to sophisticated state space models, analyzing both the latest revised data and the real-time information actually available to policymakers. The main finding is that different methods produce dramatically different estimates of economic slack, and no single approach dominates across all dimensions of performance. Mean output gap estimates span nearly four percentage points across methods, while the predictive relationship between output gaps and inflation is both method-dependent and highly time-varying. Rather than a steady Phillips curve decline, slack-inflation relationships strengthen during specific periods of macroeconomic stress. For forecasting output growth, the HP filter reduces errors by over 10% with revised data, while the unobserved components model performs best in real time. Notably, refined Beveridge-Nelson method shows strong inflation prediction but worsens growth forecasts. Forecast combinations, particularly the

Bates-Granger weighted average, provide consistently reliable performance across contexts. This fundamental uncertainty about measuring slack has important implications for monetary policy design, inflation forecasting, and our understanding of business cycle dynamics.

These findings contribute to three distinct literatures and carry important implications for both research and policy. First, I engage with methodological research on trend-cycle decomposition techniques. The Hodrick and Prescott (1997) filter became the workhorse method for extracting cyclical components, despite well-known limitations including endpoint bias and spurious cycle generation Canova (1998). Harvey (1985, 1990) and P. K. Clark (1987) developed unobserved components models that place economic structure on the decomposition through state space representations. More recently, Hamilton (2018) proposed a regression-based alternative that avoids the two-sided filtering problems inherent in the HP approach, though Quast and Wolters (2022) show this method exhibits sensitivity to horizon choice. Beveridge and Nelson (1981) offered a fundamentally different perspective by defining the trend as the long-run forecast of the series, an approach refined by Kamber, Morley, and Wong (2018, 2025); Morley (2002) to address finite-sample biases. Each method embeds different assumptions about the nature of trend growth and the persistence of shocks, yielding potentially divergent estimates of the output gap.

Despite this theoretical development, few studies comprehensively compare how these methods perform empirically across multiple dimensions. Orphanides and Norden (2002) demonstrated that real-time output gap estimates are highly unreliable and subject to substantial revision, raising serious questions about their utility for policy. Building on this work, Orphanides (2001) showed that mismeasurement of potential output and the natural rate of unemployment contributed significantly to the Great Inflation of the 1970s. Orphanides (2008) further documented how real-time misperceptions about the output gap led to policy errors during critical historical episodes. Koenig, Dolmas, and Piger (2003) emphasized that real-time data constraints fundamentally constrain the reliability of output gap estimates, showing that the use and abuse of preliminary data can lead to misleading inferences. The Federal Reserve Bank of Philadelphia's real-time data set, developed by Croushore and Stark (2001) and expanded in subsequent work by Croushore (2006), has become an essential resource for examining these issues. Croushore (2011) provides a comprehensive survey of real-time data methods and their implications for economic forecasting and policy evaluation. Marcellino and Musso (2011) compared various univariate and multivariate approaches using euro area data, find-

ing considerable disagreement across methods but limited guidance on selecting among them. Edge and Rudd (2016) challenged the entire enterprise, arguing that standard gap estimates fail basic validation tests and may be essentially uninformative about inflationary pressures. My analysis extends this literature by providing a broad comparison, examining eight methods across real-time and revised data with explicit evaluation of both inflation prediction and output forecasting performance.

Second, I contribute to the long-standing debate about the Phillips curve and its empirical relevance. The relationship between economic slack and inflation has been a central concern since Phillips (1958) documented an inverse association between unemployment and wage growth in the UK. Theoretical advances by Friedman (1968) and Phelps (1968) introduced expectations, arguing that only unexpected inflation affects real activity. Subsequent models, including the triangle specification of Gordon (1981); Gordon et al. (1977) and the New Keynesian derivation by Blanchard and Galí (2007), formalized the role of the output gap in inflation dynamics.inflationary processes.

However, empirical support for a stable Phillips curve has weakened in recent decades. Studies by Stock and Watson (1999, 2007) show substantial instability in its coefficients over time, while Cogley and Sargent (2001) attributes changing inflation dynamics to shifts in monetary policy regimes and anchored expectations. Some researchers question the curve's practical usefulness as Atkeson, Ohanian, et al. (2001) find that simple benchmarks often outperform Phillips-curve forecasts, and Stock and Watson (2008) conclude that output-gap measures offer limited predictive power in recent data. Ball and Mazumder (2011) found that standard Phillips curves dramatically underpredicted the absence of deflation during the Great Recession's large negative output gaps. Explanations for this "missing disinflation" range from anchored expectations (Coibion & Gorodnichenko, 2015) to financial frictions (Del Negro, Giannoni, & Schorfheide, 2015). Watson (2014) notes the elusive nature of traditional Phillips-curve relationships during the period, and Faust and Svensson (2001) highlights how policy-rule robustness depends critically on how slack is measured. Recent work by Hazell, Herreno, Nakamura, and Steinsson (2022) using regional variation suggests the aggregate relationship may be steeper than time-series estimates imply, renewing interest in the slack-inflation link.

Notably less explored in this debate is systematic analysis of how the choice of output gap measurement affects inferences about the Phillips curve relationship. If different decomposition methods produce fundamentally different estimates of output gap, do they also yield

different conclusions about whether slack predicts inflation? I address this question directly by conducting Granger causality tests for each estimation method across both real-time and revised data.

Third, I draw on the literature concerning forecast combination as a response to model uncertainty. The insight that combined forecasts often outperform individual models dates to Bates and Granger (1969). Subsequent research has explored various weighting schemes, from simple averaging (Newbold & Granger, 1974) to performance-based methods (Stock & Watson, 2004; Timmermann, 2006). Theoretical justifications emphasize diversification benefits and robustness to structural breaks (Elliott & Timmermann, 2013). Applications in macroeconomics show that combinations improve accuracy for inflation and output growth, especially in real time (Aiolfi & Timmermann, 2006; T. E. Clark & McCracken, 2009).

Yet despite this evidence, forecast combination remains underutilized for measuring unobserved variables like the output gap. Policy institutions typically rely on a single preferred method rather than combining multiple approaches. Recent work by Kishor and Nguyen (2025) demonstrate that combining multiple credit-gap estimators yields more reliable measures and forecasts. Given the substantial disagreement across output-gap methods documented, combination offers a natural strategy for integrating information and guarding against specification error. This paper implements three combination approaches - simple average, trimmed mean, and Bates–Granger weighting - and evaluates their performance for both inflation prediction and growth forecasting.

I construct output gap estimates using eight decomposition methods, applying each to both revised and real-time data vintages. This dual-vintage approach follows Orphanides and Norden (2002) and Koenig et al. (2003), who showed that real-time data limitations fundamentally constrain gap reliability for policy. The eight methods span the range of approaches employed in practice, from simple deterministic trends to sophisticated state space models. I also construct three combination measures to synthesize information across methods. The empirical analysis addresses three questions: How much do gap estimates differ across methods and vintages? Do different slack measures predict inflation? Do gaps improve GDP growth forecasts? The forecasting exercise uses recursive pseudo-out-of-sample evaluation to assess predictive performance under realistic information constraints.

The results matter for policy and research. First, the choice of decomposition method matters substantially yet remains fundamentally uncertain, suggesting that exclusive reliance on

any single measure is not ideal. Facing a similar challenge in constructing state-level Taylor rules, Andriantomanga, Kishor, and Kumar (2025) counter it with a selection of output gap measures affecting both the implied optimal interest rates and the estimated effects of policy deviations. Second, the time varying nature of the Phillips curve relationship complicates policy rules that rely on slack measures. Third, real-time measurement challenges underscore the difficulty of conditioning policy on variables that are both unobserved and subject to substantial revision. More broadly, the persistent disagreement across methods suggests limits to how precisely potential output can be measured.

The remainder of the paper proceeds as follows. Section 2 describes the data and presents the output gap estimation methods along with the forecast combination approaches. Section 3 documents the substantial disagreement across methods through descriptive statistics and comparisons of real-time versus revised estimates. Section 4 presents the empirical analysis examining the predictive content of output gaps for inflation and evaluates forecasting performance for output growth. Section 5 concludes the paper.

2 Data and Methodology

2.1 Data

This study uses quarterly U.S. real output and CPI inflation data. Real output data cover the period from 1965:Q3 to the present and are obtained from the Federal Reserve Bank of Philadelphia's real-time dataset. This dataset preserves the historical sequence of quarterly data vintages exactly as they were available to policymakers and analysts at each point in time. The series transitions from real GNP (through 1991:Q4) to real GDP (from 1992:Q1 onward), following standard national accounting conventions.

A key feature of this analysis is the distinction between two data perspectives: The latest vintage and the real-time data, constructed through real time vintages. The latest vintage series uses the most recent, fully revised data, representing the best available ex post measure of economic activity. The real-time series is constructed by identifying, for each calendar quarter, the earliest subsequent data vintage that contained the initial estimate of output for that quarter. This approach replicates the information set available to contemporary analysts and is essential for assessing true real-time predictive performance.

Quarterly CPI inflation data are sourced from the Bureau of Labor Statistics via the Federal

Reserve Bank of St. Louis FRED database. The monthly CPI series is converted to a quarterly frequency by taking the average of the three monthly values within each quarter. Inflation is measured as the year-over-year percentage change in the quarterly CPI, calculated as:

$$\pi_t = 100 \times \left(\frac{CPI_t}{CPI_{t-4}} - 1 \right) \tag{1}$$

Where CPI_t denotes the quarterly CPI index. This year-over-year measure is commonly used in policy analysis and helps reduce high-frequency volatility.

Output growth rates are constructed separately for each vintage to maintain consistency with the output gap estimates. For the latest vintage, the annualized quarter-over-quarter growth rate is computed as:

$$g_t^{\text{latest}} = 100 \times \left[\left(\frac{y_t}{y_{t-1}} \right)^4 - 1 \right] \tag{2}$$

Where y_t represents the level of real output in quarter t.

For the real-time output series, the construction follows a strict real-time principle. For each quarter t, I identify the vintage V_{t+1} that first contains an estimate for quarter t. The output level for quarter t is taken from V_{t+1}, while the output level for quarter $t-1$ is taken from the same vintage V_{t+1} (which contains the most recent revision of the previous quarter available at that time). The growth rate is then computed using the same annualization formula:

$$g_t^{\text{real-time}} = 100 \times \left[\left(\frac{y_t^{V_{t+1}}}{y_{t-1}^{V_{t+1}}} \right)^4 - 1 \right] \tag{3}$$

This procedure ensures that the growth series reflects exactly the information that would have been available to an analyst in real time, avoiding any look-ahead bias.

All output gap estimation methods are initialized using data from 1965:Q3 to allow for sufficient burn-in for model estimation. Several methods, particularly those requiring longer historical series for identification or smoothing, produce missing values at the beginning of the sample. To ensure comparability across methods and between real-time and latest vintage estimates, the analysis is restricted to a balanced sample that begins in 1980:Q4. This common sample period ensures that all output gap series are available for every quarter and facilitates meaningful comparisons across estimation approaches and data vintages.

2.2 Output Gap Estimation Methods

I estimate output gaps using eight trend–cycle decomposition methods that differ in how they separate permanent and transitory movements in output. In all cases, the output gap is defined as the percentage deviation of actual output from its estimated trend. For each method, two versions are computed: one using the latest vintage data (representing ex post best estimates) and one using real-time data (replicating contemporaneous information sets). For the latest vintage, each decomposition method is applied once to the complete, final revised series. For the real-time, the estimation is recursive and information-constrained. At each historical quarter t, I identify the earliest data vintage that contains the output estimate for quarter t, then apply each decomposition method using only the data available in that vintage (which includes revised estimates for all previous quarters). The eight methods, described in detail below (and summarized in Table 1), encompass a range of approaches from simple deterministic trends to sophisticated model-based and filter-based techniques.

Linear and Quadratic Deterministic Trends serve as simple benchmarks. The trend component is modeled as a deterministic function of time. For the linear trend, the specification is:

$$\ln y_t = \alpha + \beta t + \varepsilon_t, \tag{4}$$

where $\ln y_t$ is the natural log of real output, t is a linear time index, and ε_t is the stationary cyclical component. For the quadratic trend, an additional squared term is included to capture potential curvature:

$$\ln y_t = \alpha + \beta_1 t + \beta_2 t^2 + \varepsilon_t. \tag{5}$$

The output gap in both cases is $100 \times \varepsilon_t$.

The **Hodrick–Prescott (HP) filter** (Hodrick & Prescott, 1997) is a commonly used smoothing technique that decomposes the log of output, $\ln y_t$, into a trend component τ_t and a cyclical component c_t by solving the following minimization problem:

$$\min_{\{\tau_t\}} \sum_{t=1}^{T} (\ln y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} \left[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}) \right]^2, \tag{6}$$

where λ controls the smoothness of the trend. For quarterly data, λ is set to 1600. The output gap is then $100 \times c_t$, with $c_t = \ln y_t - \tau_t$.

The **Unobserved Components (UC) model** (P. K. Clark, 1987; Harvey, 1985, 1990) specifies the

log of output as the sum of a stochastic trend (μ_t), a stochastic cycle (ψ_t), and irregular noise (ϵ_t):

$$\ln y_t = \mu_t + \psi_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2), \tag{7}$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2), \tag{8}$$

$$\beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim N(0, \sigma_\zeta^2), \tag{9}$$

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}, \quad \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \sim N\left(0, \sigma_\kappa^2 I\right), \tag{10}$$

where μ_t is a local linear trend with slope β_t, ψ_t is a stationary cycle with period $2\pi/\lambda_c$ and damping factor $\rho \in (0,1)$, and ψ_t^* is an auxiliary cycle component. The model is estimated by maximum likelihood using the Kalman filter, with initial variance priors obtained from a preliminary HP filter. The output gap is the smoothed estimate of ψ_t multiplied by 100.

The **Hamilton (2018) filter** proposes a one-sided, regression-based alternative to two-sided smoothing filters. The trend is defined as the forecast of the series h periods ahead conditional on its own lags:

$$\ln y_{t+h} = \beta_0 + \beta_1 \ln y_t + \beta_2 \ln y_{t-1} + \cdots + \beta_p \ln y_{t-p+1} + u_{t+h}. \tag{11}$$

Consistent with Hamilton's recommendation for quarterly data, I set the forecast horizon $h = 8$ and the number of lags $p = 4$. The output gap is $100 \times (\ln y_t - \widehat{\ln y}_{t+h|t})$, where $\widehat{\ln y}_{t+h|t}$ is the fitted value from equation (11). The **Modified Hamilton filter** (Quast & Wolters, 2022) addresses sensitivity to the choice of horizon by averaging cycle estimates obtained from projection horizons $h = 4, 5, \ldots, 12$, keeping $p = 4$ fixed.

The **Beveridge–Nelson (BN) decomposition** (Beveridge & Nelson, 1981) defines the trend as the long-horizon forecast of the series given its current and past behavior. For a series $\ln y_t$ following an ARIMA(1,1,0) process ($\Delta \ln y_t = \phi \Delta \ln y_{t-1} + \varepsilon_t$), the BN trend is

$$\tau_t^{\text{BN}} = \ln y_t + \lim_{j \to \infty} \sum_{k=1}^{j} E_t[\Delta \ln y_{t+k}]. \tag{12}$$

The cyclical component is simplified as

$$c_t = \ln y_t - \tau_t^{\text{BN}} = -\frac{\phi}{1-\phi}\varepsilon_t, \tag{13}$$

where ϕ is the autoregressive parameter from the ARIMA(1,1,0) model and ε_t are the model innovations. The output gap is $100 \times c_t$.

The **Refined Beveridge–Nelson decomposition** builds on the work of Kamber et al. (2018, 2025). This method implements few refinements to improve the finite-sample properties of the standard BN decomposition. First, an iterative dynamic demeaning procedure estimates a time-varying drift, reducing spurious cycles caused by low-frequency mean shifts. Second, the persistence parameter δ is chosen to minimize the estimated variance of trend shocks, yielding a smoother trend. Third, the implementation uses a Bayesian approach with a Minnesota prior for the AR coefficients, shrinking higher-order lags. I implement their algorithm with AR(12) specification and the iterative dynamic demeaning procedure.

Table 1: Summary of Output Gap Estimation Methods

Method	Brief Description	Key Parameter(s)
Linear Trend	Deterministic linear time trend	None
Quadratic Trend	Deterministic quadratic time trend	None
HP Filter	Hodrick-Prescott smoothing filter	$\lambda = 1600$
Unobserved Components	State-space model with stochastic trend and cycle	Estimated variances
Hamilton Filter	Regression-based trend projection	$h = 8, p = 4$
Modified Hamilton	Average of multiple projection horizons	$h = 4$ to 12, $p = 4$
Beveridge-Nelson	ARIMA-based long-horizon forecast	ARIMA(1,1,0)
Refined Beveridge-Nelson	Modified BN with dynamic demeaning and persistence selection	AR(12), δ optimization

Notes: Brief description and key parameter(s) for each output gap estimation method. All methods are applied separately to latest vintage and real-time data.

In addition to these eight individual methods, I construct three aggregate measures to summarize the information across estimators. The first is a simple **Average** of all eight output gap series. The second is a **20% Trimmed Mean**, which discards the highest and lowest 20% of the eight estimates in each period before averaging, thereby reducing the influence of outliers. The third is a **Bates–Granger weighted average** (Bates & Granger, 1969), which weights each method inversely by its recent forecast error variance, assigning greater weight to methods with stronger historical forecasting performance. These aggregates provide a more robust central measure of economic slack and are included in the subsequent analysis.

2.3 Granger Causality

To test the fundamental Phillips curve proposition, I examine whether various output gap measures contain predictive information for future inflation using Granger causality tests. This

analysis addresses whether different methods for measuring economic slack yield different inferences about inflation dynamics, both in real time and with the benefit of hindsight.

2.3.1 Full-Sample Granger Causality Tests

For each output gap method m and vintage v, I test the null hypothesis that the output gap does not Granger-cause inflation. This corresponds to the basic Phillips curve proposition that economic slack should predict future inflation movements. The analysis uses a bivariate VAR framework:

$$\pi_t = \alpha_1 + \sum_{j=1}^{k} \beta_{1j} \pi_{t-j} + \sum_{j=1}^{k} \gamma_{1j} \tilde{y}^m_{t-j} + \epsilon_{1t} \tag{14}$$

$$\tilde{y}^m_t = \alpha_2 + \sum_{j=1}^{k} \beta_{2j} \pi_{t-j} + \sum_{j=1}^{k} \gamma_{2j} \tilde{y}^m_{t-j} + \epsilon_{2t} \tag{15}$$

where π_t is quarterly CPI inflation and $\tilde{y}^m_t$ is the output gap from method m. The test examines whether the coefficients γ_{1j} are jointly zero. Tests are performed with both a fixed lag length of one quarter and with optimal lag length selected by BIC up to a maximum of six lags.

2.3.2 Time-Varying Granger Causality

Given swell-documented instability in the Phillips curve over recent decades, I implement rolling-window Granger causality tests to examine potential changes in the predictive relationship. For each quarter t in the sample (starting from the 40th observation), I estimate equation (**??**) using a 40-quarter rolling window ending at quarter t, with a fixed lag length of one quarter for consistency across windows. The resulting F-statistics are plotted over time to visualize how the strength of the inflation-slack relationship evolves.

For each window, the F-statistic for testing $\gamma_{11} = 0$ is computed as:

$$F = \frac{(RSS_r - RSS_u)/1}{RSS_u/(T - 2k - 1)} \tag{16}$$

where RSS_r is the restricted sum of squares (omitting the output gap), RSS_u is the unrestricted sum of squares, T is the window size (40 quarters), and k is the number of lags.

To summarize the time-varying evidence across methods and periods, I compute for each

output gap series the proportion of rolling windows in which the Granger causality test rejects the null hypothesis at the 5% significance level. These proportions are presented as heatmaps, providing a visual summary of which methods show consistent predictive power over different historical periods and whether the Phillips curve relationship has strengthened or weakened over time.

This Granger causality methodology allows for comprehensive assessment of whether different measures of economic slack yield different inferences about inflation dynamics, and how these relationships have evolved in the context of the changing Phillips curve.

2.4 Forecasting Methodology

2.4.1 Direct Forecasting Framework

I assess the forecasting value of output gaps using a direct, pseudo-out-of-sample framework. The forecasting exercise adopts a direct h-step ahead approach (Marcellino, Stock, & Watson, 2006), where separate models are estimated for each forecast horizon $h \in \{1, 2, 3, 4\}$ quarters. For a given output gap method m and vintage v (latest or real-time), the forecasting model takes the form:

$$g_{t+h} = \alpha_h + \sum_{j=1}^{p} \beta_{h,j} g_{t-j+1} + \gamma_h \tilde{y}_t^m + \epsilon_{t+h}, \tag{17}$$

where g_{t+h} is the h-period ahead annualized output growth rate, g_{t-j+1} are lagged growth rates, and $\tilde{y}_t^m$ is the output gap estimate from method m. The autoregressive order p is selected recursively using the Bayesian Information Criterion (BIC), with a maximum of four lags considered. A univariate ARMA model, also with order selected by BIC, serves as a benchmark against which gap-augmented models are compared.

2.4.2 Recursive Estimation and Evaluation

The forecasting exercise follows a pseudo-out-of-sample design with recursive estimation. The initial estimation period ends in 1985:Q4, providing approximately five years of data after the balanced sample begins in 1980:Q4. For each subsequent quarter through the end of the sample, the following steps are performed:

1. Estimate model (17) using data available up to quarter t

2. Generate h-step ahead forecasts for horizons 1 through 4

3. Move to quarter $t + 1$ and repeat

This procedure is implemented separately for each of the eight output gap methods and three aggregate measures, using both real-time and latest vintage data. The recursive design mimics a realistic forecasting environment where parameters are re-estimated as new data become available.

2.4.3 Forecast Combination Methods

Given the pronounced disagreement across decomposition methods documented in section 3, I implement three forecast combination strategies to consolidate different approaches:

(a) **Simple Average**: The mean forecast across all eight individual method forecasts at each horizon and point in time:

$$\hat{g}_{t+h}^{\text{avg}} = \frac{1}{8} \sum_{m=1}^{8} \hat{g}_{t+h}^{m} \tag{18}$$

(b) **Trimmed Mean**: The average after excluding the highest and lowest 20% of individual forecasts (effectively dropping the extreme two forecasts in each period):

$$\hat{g}_{t+h}^{\text{trim}} = \frac{1}{6} \sum_{m \in \mathcal{M}_{\text{middle}}} \hat{g}_{t+h}^{m} \tag{19}$$

Where $\mathcal{M}_{\text{middle}}$ contains the six middle forecasts after sorting.

(c) **Bates-Granger Weighted Average**: A performance-based combination that weights each method's forecast inversely by its recent forecast error variance:

$$\hat{g}_{t+h}^{\text{BG}} = \sum_{m=1}^{8} w_{t,m} \hat{g}_{t+h}^{m}, \quad w_{t,m} = \frac{1/\hat{\sigma}_{t,m}^2}{\sum_{j=1}^{8} 1/\hat{\sigma}_{t,j}^2} \tag{20}$$

Where $\hat{\sigma}_{t,m}^2$ is the recursive mean squared forecast error for method m computed over a 40-quarter rolling window. Weights are updated each period, making this a time-varying, adaptive combination.

2.4.4 Evaluation Metrics

Forecast accuracy is evaluated using the Root Mean Squared Forecast Error (RMSE):

$$\text{RMSE}_{m,h} = \sqrt{\frac{1}{T_h}\sum_{t=1}^{T_h}(g_{t+h} - \hat{g}_{t+h}^{m})^2}, \tag{21}$$

where T_h is the number of h-step ahead forecasts available. Results are presented as relative RMSEs (ratio to the ARMA benchmark RMSE), with values below 1 indicating superior performance.

For pairwise comparison of forecast accuracy between the Bates-Granger combination and individual methods, I employ the Clark-West test (T. E. Clark & West, 2007). This test is specifically designed for comparing nested models in a forecasting context and accounts for parameter estimation uncertainty. The test examines whether the mean squared prediction error of the Bates-Granger combination is significantly smaller than that of individual method forecasts. The test statistic uses Newey-West standard errors to correct for serial correlation in multi-step ahead forecast errors.

I also report mean absolute errors and directional accuracy, though RMSE remains the primary evaluation criterion as it penalizes large forecast errors more heavily.

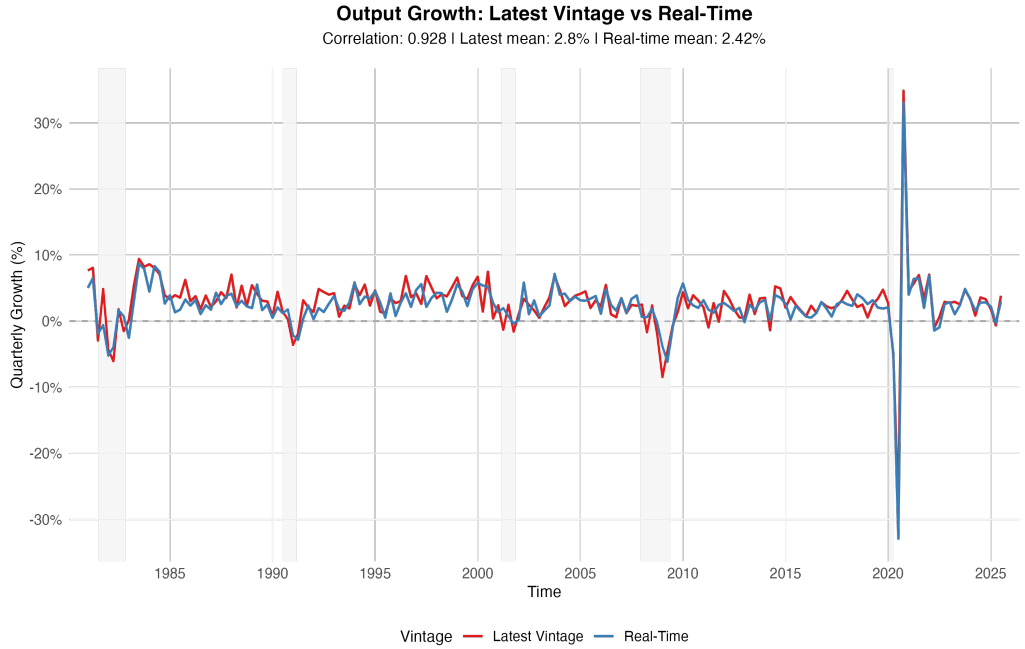
3 Descriptive Statistics

3.1 Output Growth: Latest Vintage vs Real-Time

Figure 1 compares output growth rates from latest vintage and real-time data. The series correlate at 0.93 suggesting data revisions change growth magnitudes but preserve the timing and direction of business cycle fluctuations.

Three patterns stand out in this comparison. The latest vintage series averages 2.8% growth versus 2.42% for real-time data, showing a systematic upward revision of 0.38 percentage points. This positive mean revision aligns with the well-documented tendency for output growth estimates to be revised upward as more complete source data become available (Croushore & Stark, 2001; Orphanides, 2001). Differences peak around recessions, particularly during the early 1980s, early 1990s, and 2007-2009. These episodes involve substantial revisions as statistical agencies incorporate delayed source data, adjust seasonal factors, or implement method-

Figure 1: Output Growth: Latest Vintage vs Real-Time



Notes: The figure displays annualized quarter-over-quarter output growth rates. Latest vintage series (red) uses fully revised data, while real-time series (blue) replicates information available contemporaneously. Gray shaded areas indicate NBER recessions.

ological changes. The real-time series excludes such ex post information, providing a clearer picture of what policymakers actually saw. The standard deviation of revisions is 1.96 percentage points, indicating considerable quarter-to-quarter variation. This revision uncertainty underscores why distinguishing between ex post and real-time data matters for evaluating historical policy decisions or testing forecasting models. I maintain this distinction throughout, presenting results separately for each vintage.

3.2 Output Gaps: Methods Comparison

The eight decomposition methods and three aggregate measures produce output gap series that differ dramatically. Table 2 reports summary statistics for both real-time and latest-vintage estimates. Mean estimates span a range of nearly 4 percentage points; volatility varies by a factor of ten.

Average estimated gaps vary substantially across methods. Using latest-vintage data, means range from -0.20% (Beveridge-Nelson) to 0.83% (linear trend). Real-time averages diverge even more, from -3.76% (linear trend) to -0.01% (HP filter). These differences stem from contrasting assumptions about trend growth and the nature of economic fluctuations. Deterministic trend

Table 2: Summary Statistics of Output Gap Estimates

Method	N	Mean	Std. Dev.	Min	Max
Latest Vintage					
Linear Trend	179	0.832	5.170	-15.010	10.327
Quadratic Trend	179	-0.203	3.314	-10.610	6.514
HP Filter	179	-0.058	1.381	-8.905	2.371
Unobserved Components	179	-0.043	0.989	-2.764	1.842
Hamilton	179	0.041	3.033	-9.803	8.472
Modified Hamilton	179	0.008	2.691	-9.613	4.598
Beveridge-Nelson	179	-0.204	0.499	-4.395	3.434
Refined BN	179	-0.052	1.350	-8.310	3.291
Average	179	0.040	1.593	-5.369	3.532
Trimmed Mean	179	0.038	1.365	-4.808	2.978
Bates-Granger	179	-0.161	0.442	-2.614	2.638
Real-Time					
Linear Trend	179	-3.756	4.852	-20.322	4.636
Quadratic Trend	179	-1.425	3.079	-13.026	4.035
HP Filter	179	-0.001	1.454	-9.716	3.841
Unobserved Components	179	-0.121	1.226	-3.282	2.015
Hamilton	179	-0.051	2.756	-11.084	7.614
Modified Hamilton	179	-0.072	2.274	-10.802	4.082
Beveridge-Nelson	179	-0.322	0.840	-3.293	3.453
Refined BN	179	-0.045	1.445	-11.686	3.115
Average	179	-0.724	1.393	-5.937	1.922
Trimmed Mean	179	-0.464	1.262	-5.070	1.778
Bates-Granger	179	-0.494	1.060	-8.776	1.131

Notes: Sample period: 1980:Q4 to 2025:Q2 (179 quarters). Mean, standard deviation, minimum, and maximum values are reported in percentage points. The *Latest Vintage* panel uses fully revised data; the *Real-Time* panel replicates information available to policymakers in real time.

methods show the most negative means in real time, systematically overestimating economic slack when using preliminary data.

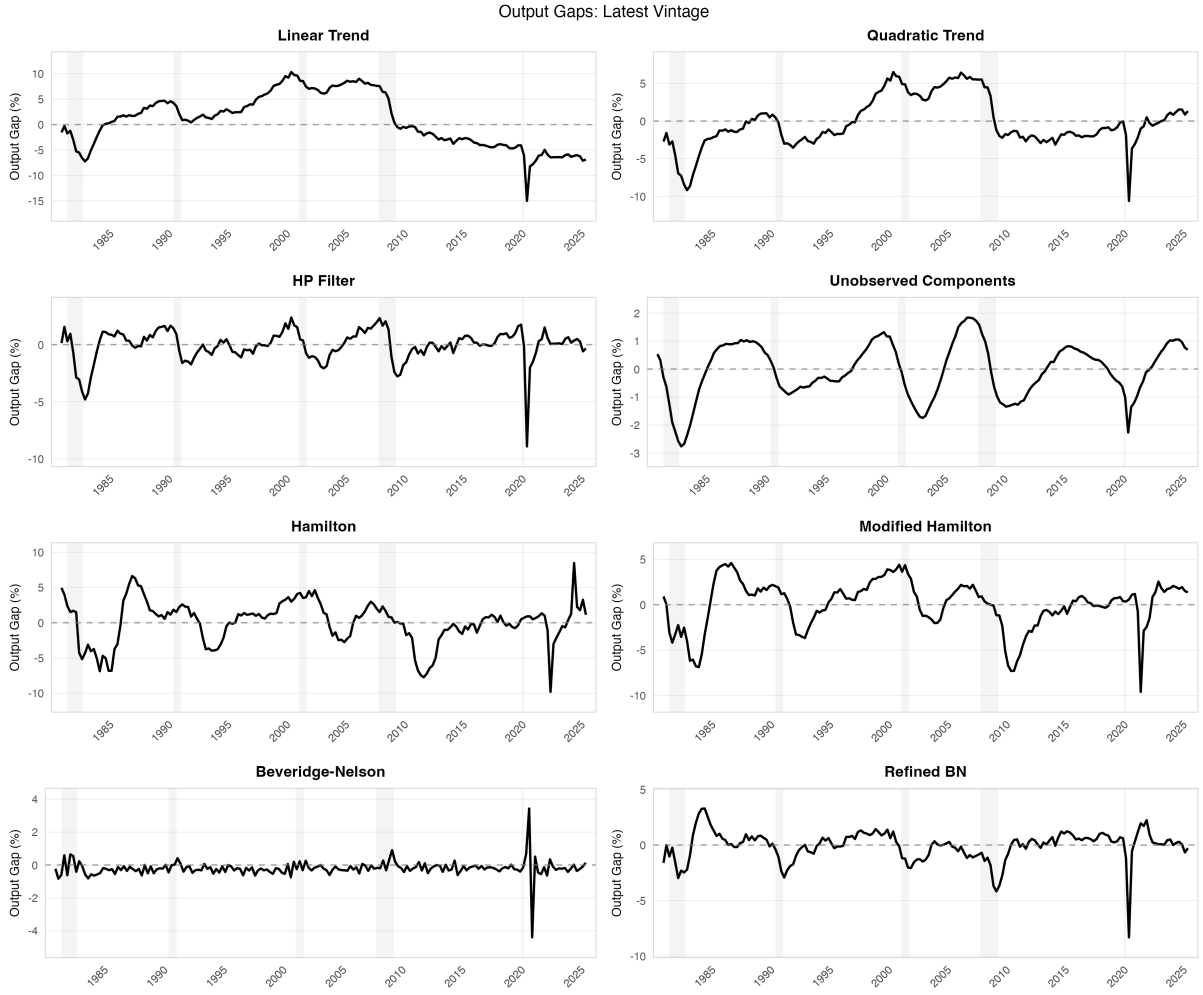
Volatility differs even more sharply. Beveridge-Nelson method yield the smoothest series with standard deviations of 0.499 for latest vintage (slightly above Bates-Granger's 0.442) and 0.840 for real time (lowest). Linear and quadratic trend methods produce high volatility with standard deviations exceeding 3.0 percentage points. Real-time series run somewhat more volatile than their latest-vintage counterparts, reflecting additional uncertainty in preliminary data.

The aggregate measures fall between the extremes. The Bates-Granger weighted average produces the least volatile series (standard deviation of 0.44% for latest vintage, 1.06% for real time), while the simple average shows greater variability. This suggests that combining methods mitigates the extreme properties of individual estimators.

Figures 2 and 3 present latest-vintage and real-time output gap series. The figures reveal visible differences in amplitude, smoothness, and sometimes even the sign of estimated cycles.

For example, during the recovery from the global financial crisis, some methods indicated a large and persistent negative output gap while others suggested a much quicker return to potential. Beveridge Nelson method being the exception and theoretically consistent shows counterintuitive and small amplitudes. Notably, no single statistical filter provides a definitive measure of economic slack.

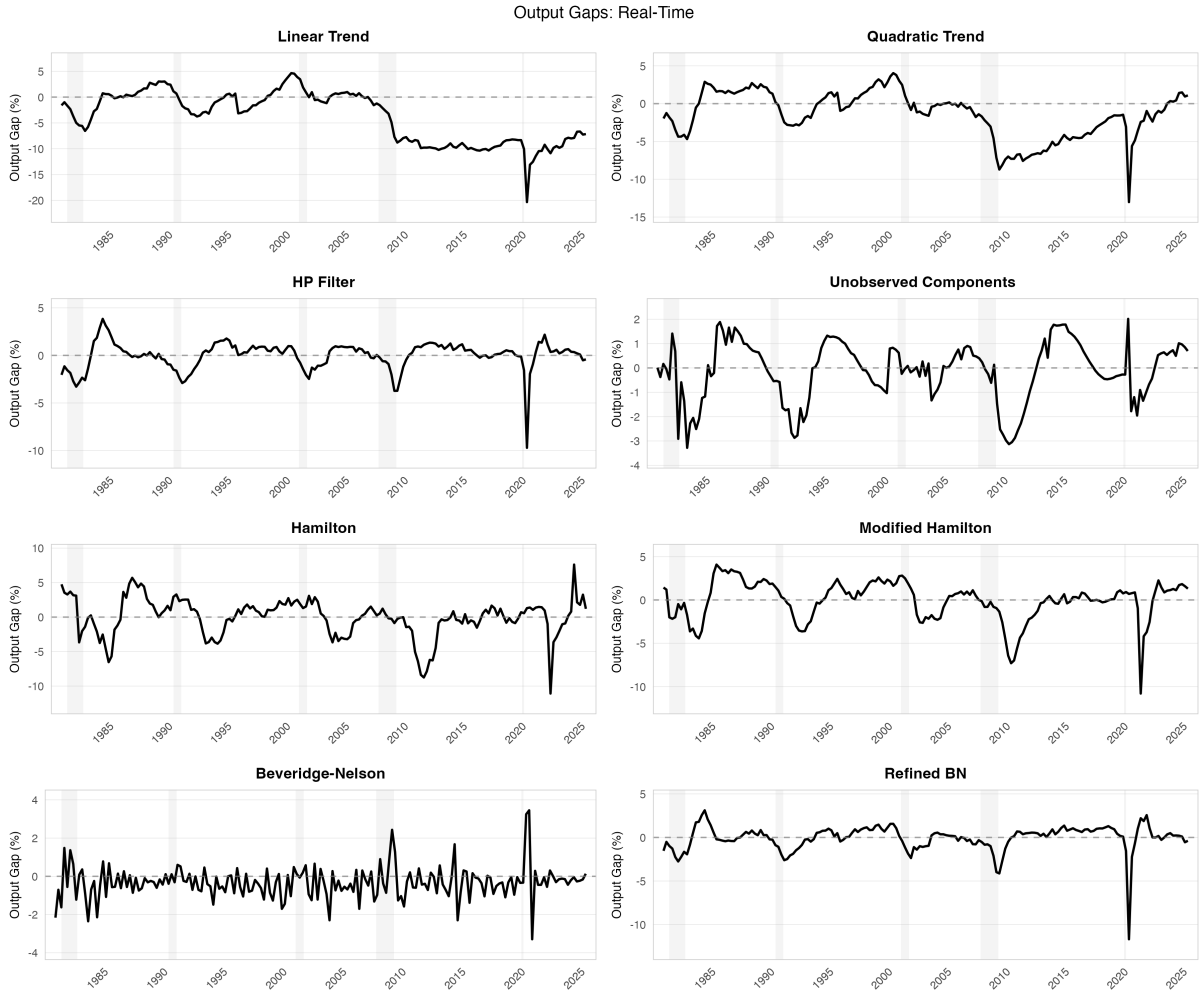
Figure 2: Latest-Vintage Output Gap Estimates from Eight Decomposition Methods



Notes: Each panel displays the output gap for one decomposition method using fully revised data. Gray shaded areas indicate NBER recessions.

The correlation matrices in Tables 3 and 4 reveal how methods relate within each series. These correlations vary sharply across methods and vintages. In latest vintage, the HP filter and unobserved component correlate most strongly (0.756), along with linear-quadratic methods (0.759), capturing similar cyclical dynamics. While Beveridge Nelson method shows negative, near-zero or both with most other methods. In real time data too the Beveridge Nelson shows same pattern. However correlation between HP filter and unobserved component

Figure 3: Real-Time Output Gap Estimates from Eight Decomposition Methods



Notes: Each panel displays for one decomposition method using real-time data. Gray shaded areas indicate NBER recessions.

Table 3: Latest Vintage Output Gaps - Correlation

	Linear Trend	Quadratic Trend	HP Filter	Unobserved Components	Hamilton	Modified Hamilton	Beveridge-Nelson	Refined BN
Linear Trend	1.000	0.759	0.377	0.327	0.308	0.367	-0.075	0.016
Quadratic Trend	0.759	1.000	0.591	0.584	0.391	0.499	-0.070	0.117
HP Filter	0.377	0.591	1.000	0.756	0.167	0.399	-0.190	0.606
Unobserved Components	0.327	0.584	0.756	1.000	0.428	0.689	-0.077	0.299
Hamilton	0.308	0.391	0.167	0.428	1.000	0.548	0.022	-0.221
Modified Hamilton	0.367	0.499	0.399	0.689	0.548	1.000	0.018	-0.020
Beveridge-Nelson	-0.075	-0.070	-0.190	-0.077	0.022	0.018	1.000	-0.321
Refined BN	0.016	0.117	0.606	0.299	-0.221	-0.020	-0.321	1.000

Notes: Pairwise correlation coefficients between output gap estimates from eight decomposition methods using latest-vintage (fully revised) data.

method is flipped from strongest in latest vintage to near zero in real time. HP filter's correlation with refined BN strengthen from 0.606 in latest to 0.922 (the strongest) in real time.

Table 4: Real-Time Output Gaps - Correlation

	Linear Trend	Quadratic Trend	HP Filter	Unobserved Components	Hamilton	Modified Hamilton	Beveridge-Nelson	Refined BN
Linear Trend	1.000	0.799	0.148	0.105	0.258	0.373	-0.192	0.140
Quadratic Trend	0.799	1.000	0.381	0.267	0.308	0.557	-0.223	0.414
HP Filter	0.148	0.381	1.000	0.003	-0.384	-0.016	-0.403	0.922
Unobserved Components	0.105	0.267	0.003	1.000	0.358	0.667	0.129	-0.010
Hamilton	0.258	0.308	-0.384	0.358	1.000	0.367	0.035	-0.204
Modified Hamilton	0.373	0.557	-0.016	0.667	0.367	1.000	0.045	-0.017
Beveridge-Nelson	-0.192	-0.223	-0.403	0.129	0.035	0.045	1.000	-0.451
Refined BN	0.140	0.414	0.922	-0.010	-0.204	-0.017	-0.451	1.000

Notes: Pairwise correlation coefficients between output gap estimates from eight decomposition methods using real-time data.

The extent of revision between real-time and latest-vintage estimates varies considerably across methods. Table 5 reports correlation and revision-error metrics between the series for each method. The Hamilton and Modified Hamilton filters display the highest correlations (0.935 and 0.939) and modest revision errors. The refined Beveridge-Nelson decomposition also shows strong correlation (0.933) and relatively low revision error (MAE 0.37). At the other extreme, the quadratic trend method has the lowest correlation (0.503) and larger errors, with mean absolute revision of 2.76 percentage points.

Table 5: Output Gap Comparison - Latest Vintage vs Real-Time

Method	N	Correlation	MAE	MaxAE	RMSD
Linear Trend	179	0.845	4.662	9.302	5.373
Quadratic Trend	179	0.503	2.762	7.316	3.410
HP Filter	179	0.597	1.012	3.254	1.272
Unobserved Components	179	0.561	0.822	4.284	1.058
Hamilton	179	0.935	0.765	4.295	1.078
Modified Hamilton	179	0.939	0.736	2.752	0.960
Beveridge-Nelson	179	0.696	0.446	2.571	0.619
Refined BN	179	0.933	0.366	3.377	0.519
Average	179	0.801	1.025	2.843	1.226
Trimmed Mean	179	0.760	0.798	2.475	1.041
Bates-Granger	179	0.570	0.632	7.091	0.945

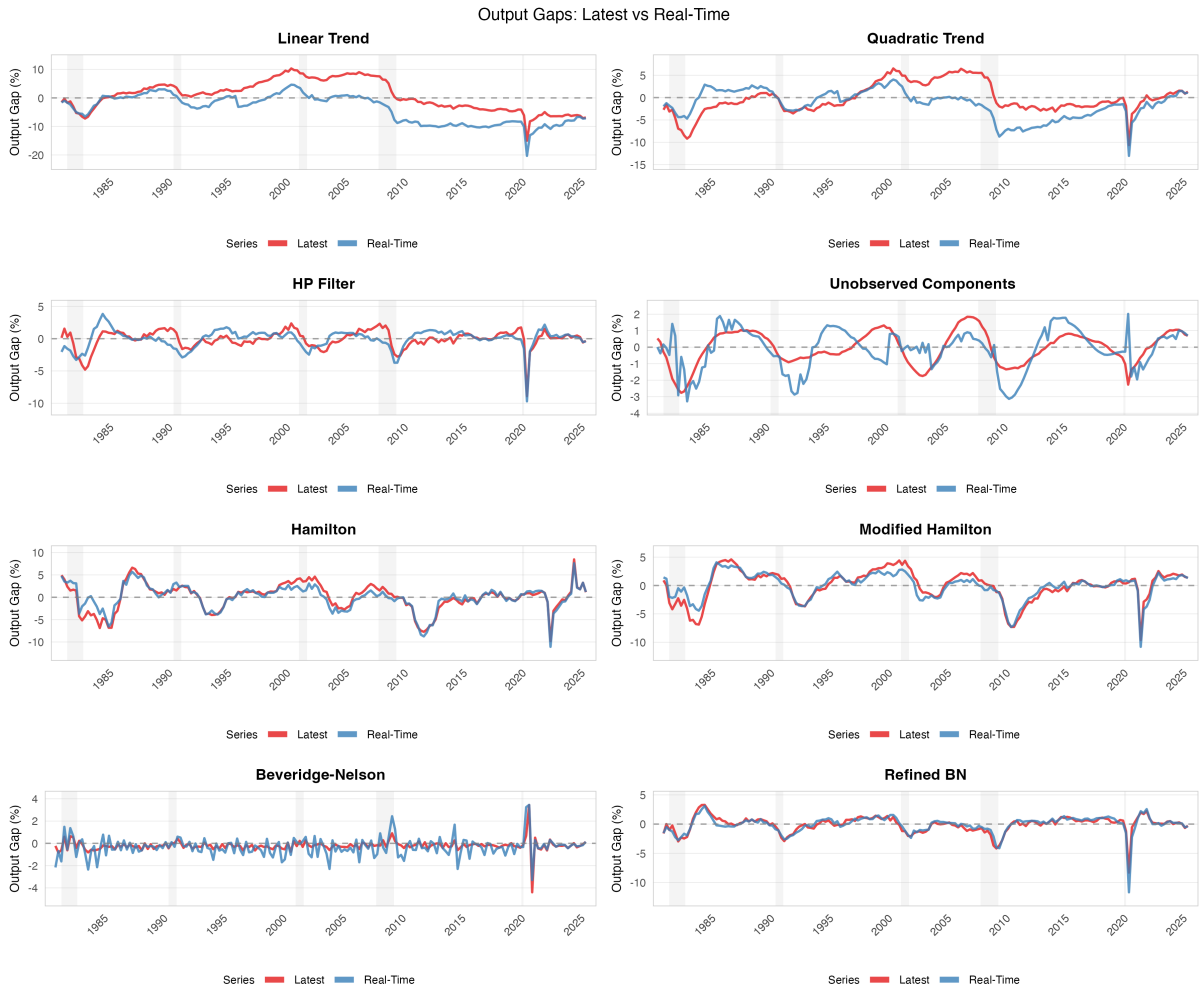
Notes: MAE = Mean Absolute Error; MaxAE = Maximum Absolute Error; RMSD = Root Mean Square Deviation. All metrics are computed as differences between real-time and latest-vintage estimates for each method and are expressed in percentage points.

Figure 4 shows these revision patterns. For linear and quadratic trends, real-time estimates run substantially more negative than later revisions, particularly during and after recessions. In contrast, the refined Beveridge-Nelson series lie almost on top of one another as revisions barely move them. Hamilton and modified Hamilton filters also track closely, though some divergence appears around cyclical turning points.

The aggregate measures of output gaps - simple average, trimmed mean, and Bates-Granger

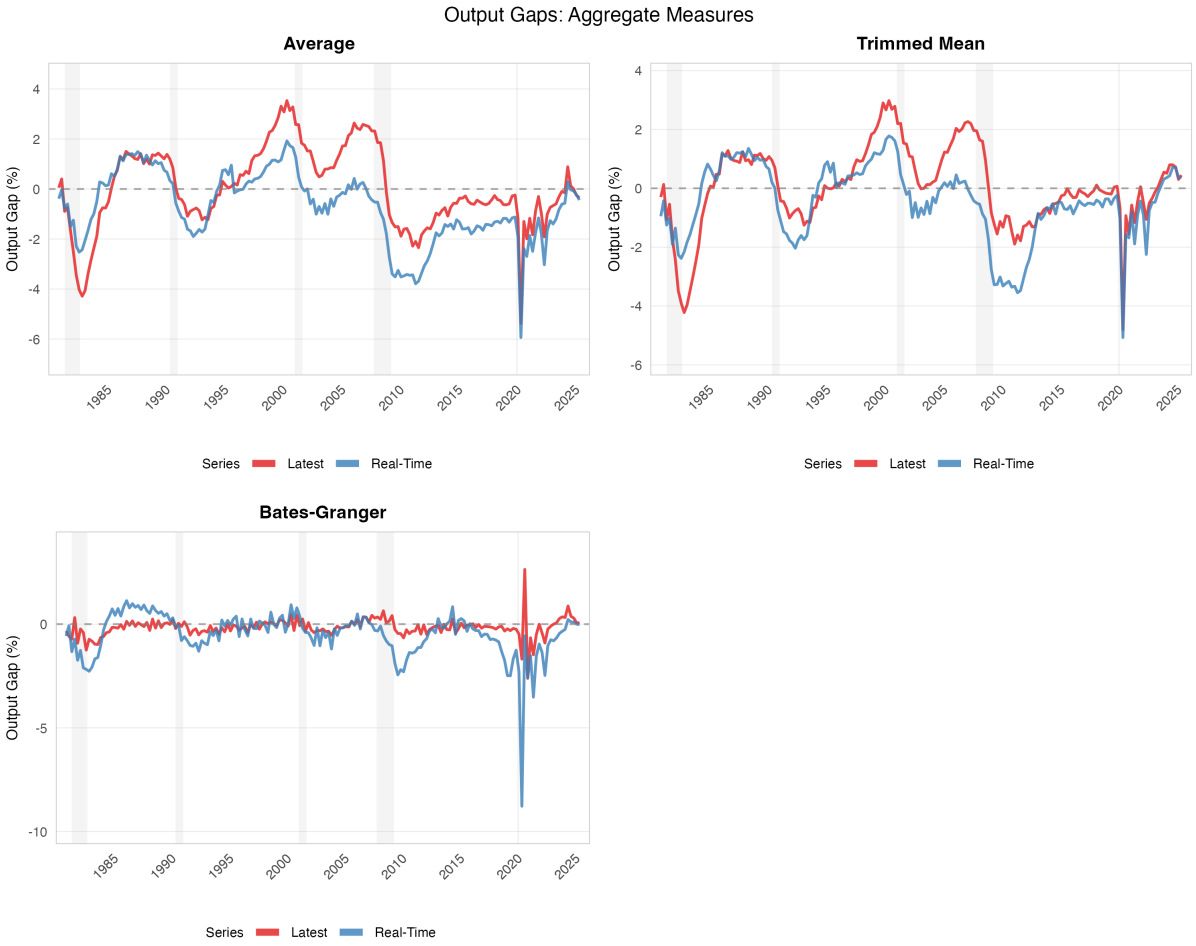
weighted average - synthesize information across methods. Figure 5 plots these aggregates for both vintages. All three aggregates show smaller revisions than the most volatile individual methods. The Bates-Granger aggregate, which weights methods inversely by their historical variance, produces relatively smooth series and exhibits moderate revision sizes, despite some increased volatility in its components. This makes it a potentially useful summary measure for real-time policy analysis.

Figure 4: Real-Time vs Latest-Vintage Output Gap Comparisons



Notes: Each panel compares real-time (blue) and latest-vintage (red) output gap estimates for one method. Gray shaded areas indicate NBER recessions.

Figure 5: Aggregate Output Gap Measures: Latest vs Real-Time



Notes: Comparison of three aggregate output gap measures: simple average, 20% trimmed mean, and Bates-Granger weighted average. Each panel shows both the latest vintage (red) and real-time (blue) estimates for one aggregate measure.

4 Empirical Analysis

4.1 Predictive Content for Inflation

The Phillips curve posits an inverse relationship between economic slack and inflationary pressure, making the output gap a valuable indicator for inflation forecasting. I evaluate whether various output gap estimates contain useful predictive information for future inflation, examining both full-sample relationships and their stability over time.

4.1.1 Stationarity Testing

Prior to causality testing, I verify the stationarity (a necessary condition (Granger, 1969)) of each output gap series and the inflation series using Augmented Dickey-Fuller (ADF) tests. The test specification includes a constant and is implemented with four lags selected by the Bayesian Information Criterion. Table 6 reports ADF test results for all output gap estimates and CPI inflation.

Table 6: Stationarity Tests - ADF Test Results

Series	Vintage	Method	ADF Stat	5% Critical	Stationary	Obs
CPI	All	CPI Inflation	-4.887	-2.880	Yes	179
Output Gap	Latest	Linear Trend	-1.096	-2.880	No	179
		Quadratic Trend	-2.363	-2.880	No	179
		HP Filter	-5.148	-2.880	Yes	179
		Unobserved Components	-6.186	-2.880	Yes	179
		Hamilton	-4.382	-2.880	Yes	179
		Modified Hamilton	-3.825	-2.880	Yes	179
		Beveridge-Nelson	-6.162	-2.880	Yes	179
		Refined BN	-4.553	-2.880	Yes	179
		Average	-3.046	-2.880	Yes	179
		Trimmed Mean	-3.327	-2.880	Yes	179
		Bates-Granger	-2.929	-2.880	Yes	179
	Real-Time	Linear Trend	-1.534	-2.880	No	179
		Quadratic Trend	-2.303	-2.880	No	179
		HP Filter	-4.984	-2.880	Yes	179
		Unobserved Components	-4.338	-2.880	Yes	179
		Hamilton	-4.802	-2.880	Yes	179
		Modified Hamilton	-4.048	-2.880	Yes	179
		Beveridge-Nelson	-6.436	-2.880	Yes	179
		Refined BN	-5.280	-2.880	Yes	179
		Average	-2.756	-2.880	No	179
		Trimmed Mean	-3.183	-2.880	Yes	179
		Bates-Granger	-2.464	-2.880	No	179

Notes: Augmented Dickey-Fuller (ADF) test statistics with a constant included; 4 lags selected by BIC. The critical value at the 5% significance level is -2.880. *Stationary* indicates rejection of the unit root null at the 5% level.

The results indicate that CPI inflation is stationary, rejecting the unit root null at the 5% significance level. For output gaps, most methods produce stationary series in both vintages. However, the linear and quadratic trend methods fail to reject the unit root null for both latest and real-time vintages. Simple deterministic trend removal may not fully extract a stationary cyclical component, complicating inference about their predictive relationships with inflation. In real time, the simple average and Bates-Granger aggregate also fail to reject the unit root, suggesting that combining methods does not guarantee stationarity when components are non-stationary. Despite these limitations, I retain all methods for subsequent analysis to provide a comprehensive comparison, though results for trend-based and non-stationary aggregate methods should be interpreted with caution.

4.1.2 Full-Sample Granger Causality

Table 7 presents Granger causality test results examining whether output gaps contain predictive information for inflation. Tests are performed in two configurations, one with a fixed one-quarter lag and one with optimal lag length selected by the Bayesian Information Criterion (BIC).

Table 7: Granger Causality Tests: Output Gap → CPI Inflation

Method		Optimal Lag			Fixed Lag 1	
Vintage	Method	Lag	F-Stat	p-value	F-Stat	p-value
Latest	Linear Trend	2	0.047	0.954572	0.089	0.76550
	Quadratic Trend	2	0.020	0.980018	0.510	0.47590
	HP Filter	2	0.034	0.966195	1.920	0.16766
	Unobserved Components	5	1.941*	0.090249	0.245	0.62122
	Hamilton	6	0.440	0.851264	0.589	0.44372
	Modified Hamilton	6	3.666***	0.001942	1.305	0.25493
	Beveridge-Nelson	2	9.690***	0.000103	1.393	0.23948
	Refined BN	1	8.353***	0.004338	8.353***	0.00434
	Average	2	0.006	0.993685	0.473	0.49257
	Trimmed Mean	2	0.030	0.970873	0.371	0.54335
	Bates-Granger	2	5.047***	0.007415	1.803	0.18106
Real-Time	Linear Trend	2	0.177	0.837886	0.479	0.48988
	Quadratic Trend	2	0.414	0.661556	1.666	0.19846
	HP Filter	1	3.121*	0.079018	3.121*	0.07902
	Unobserved Components	6	1.124	0.350803	2.149	0.14442
	Hamilton	2	1.613	0.202302	0.970	0.32596
	Modified Hamilton	6	3.217***	0.005207	3.109*	0.07959
	Beveridge-Nelson	2	1.384	0.253418	3.309*	0.07060
	Refined BN	1	5.440**	0.020819	5.440**	0.02082
	Average	2	0.161	0.851107	0.585	0.44541
	Trimmed Mean	2	0.082	0.921553	0.601	0.43918
	Bates-Granger	6	1.985*	0.070798	0.309	0.57896

Notes: Tests the null hypothesis that the output gap does not Granger-cause CPI inflation. The *Optimal Lag* columns report results using the lag length selected by the BIC (SC) criterion (maximum of 6 lags). The *Fixed Lag 1* columns use a fixed one-quarter lag. Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$.

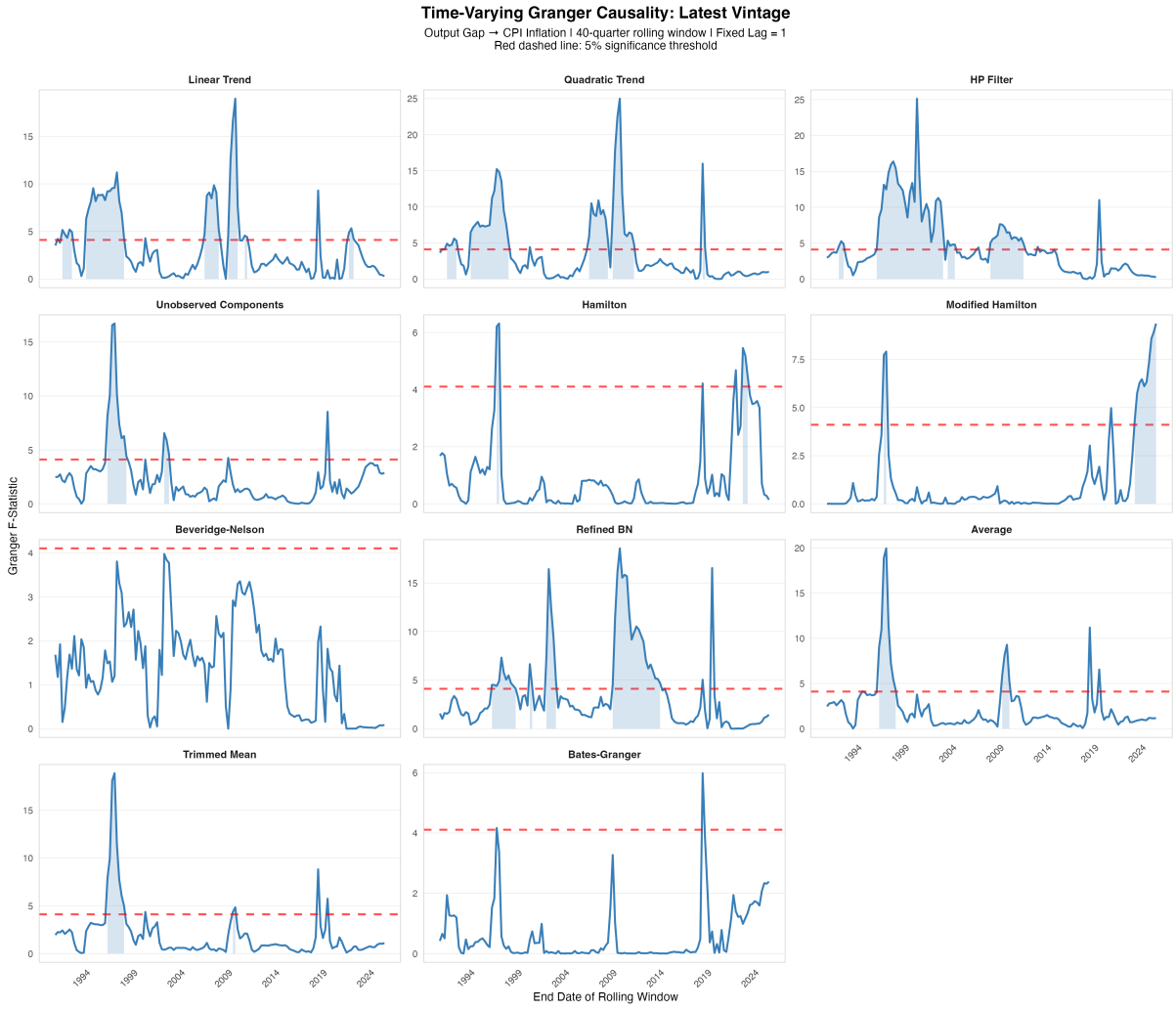
The results show three main patterns. Firstly, predictive content depends critically on both the decomposition method and data vintage. For the latest vintage, model based methods such as Beveridge Nelson, refined BN filter and Modified Hamilton along with the Bates-Granger aggregate show strong predictive power at the 1% significance level under optimal lag selection. In real time data, refined BN, modified Hamilton, HP filter, and Beveridge-Nelson show significant causality. Secondly, deterministic trend methods (linear and quadratic) show no predictive power for inflation in either vintage, regardless of lag specification. This aligns with their failure to produce stationary series in the ADF tests and suggests these simple approaches may not adequately capture the cyclical component relevant for price movement. Lastly, the aggregate measures show mixed results: the Bates-Granger aggregate is significant both in latest vintage (at 1%) and real time (10 %), while simple average and trimmed mean show no

significance at all. Combining methods does not necessarily strengthen the inflation predictive signal, possibly because different methods capture different aspects of cyclical variation.

4.1.3 Time-Varying Granger Causality

Phillips curve instability over recent decades is well documented (Ball & Mazumder, 2011). I examine whether the predictive content of output gaps has varied over time using rolling-window Granger causality tests. Figure 6 and Figure 7 present F-statistics for latest and real-time series, respectively, using a 40-quarter window with fixed one-quarter lag.

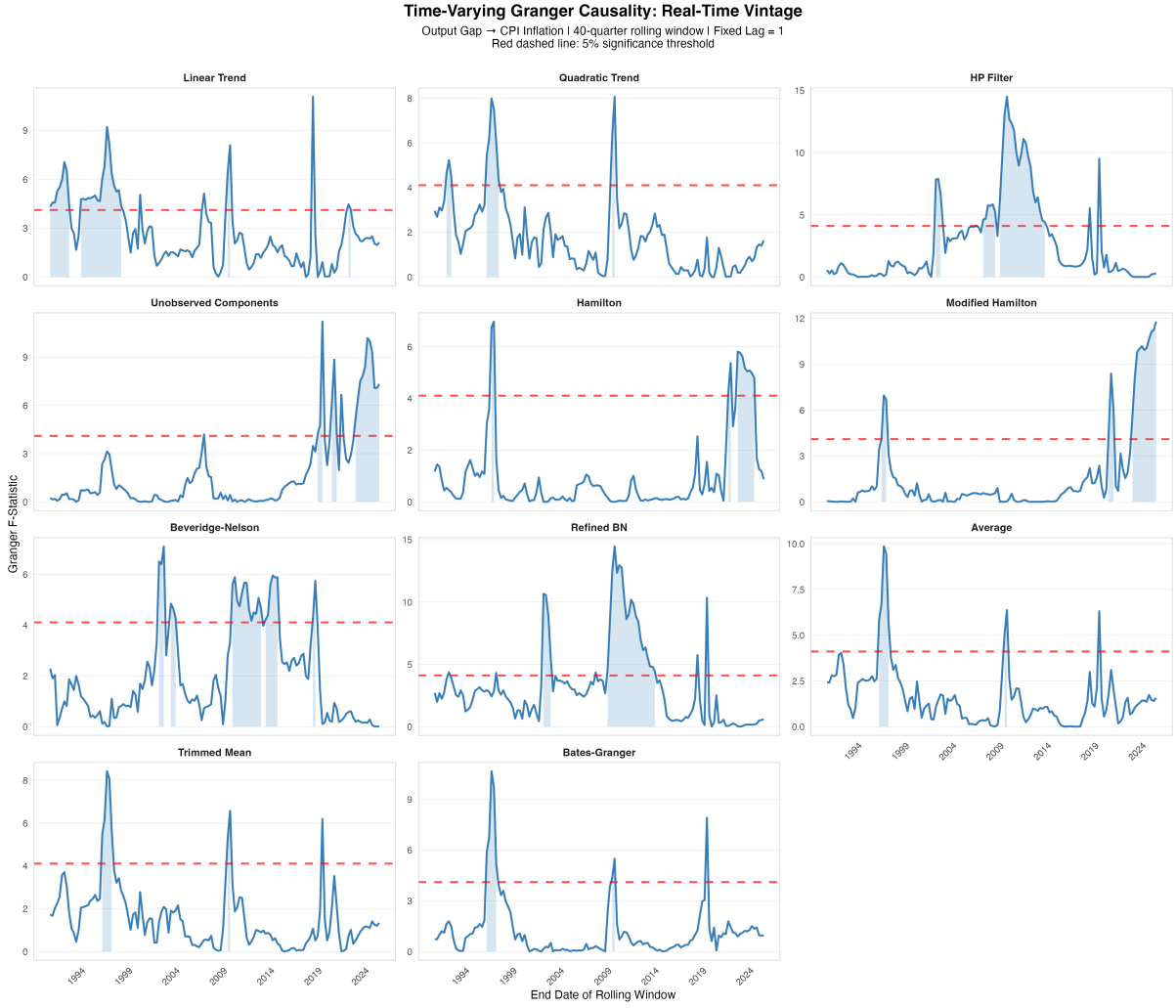
Figure 6: Time-Varying Granger Causality: Latest Vintage



Notes: F-statistics from 40-quarter rolling windows testing the null hypothesis that the output gap does not Granger-cause CPI inflation, using latest-vintage data. A fixed lag length of one quarter is used. The red dashed line indicates the 5% significance threshold.

Both figures reveal substantial time variation in the strength of the inflation-slack relationship, pointing to several distinct episodes: (1) the late 1980s through early 1990s show

Figure 7: Time-Varying Granger Causality: Real-Time Vintage



Notes: F-statistics from 40-quarter rolling windows testing the null hypothesis that the output gap does not Granger-cause CPI inflation, using real-time data. A fixed lag length of one quarter is used. The red dashed line indicates the 5% significance threshold.

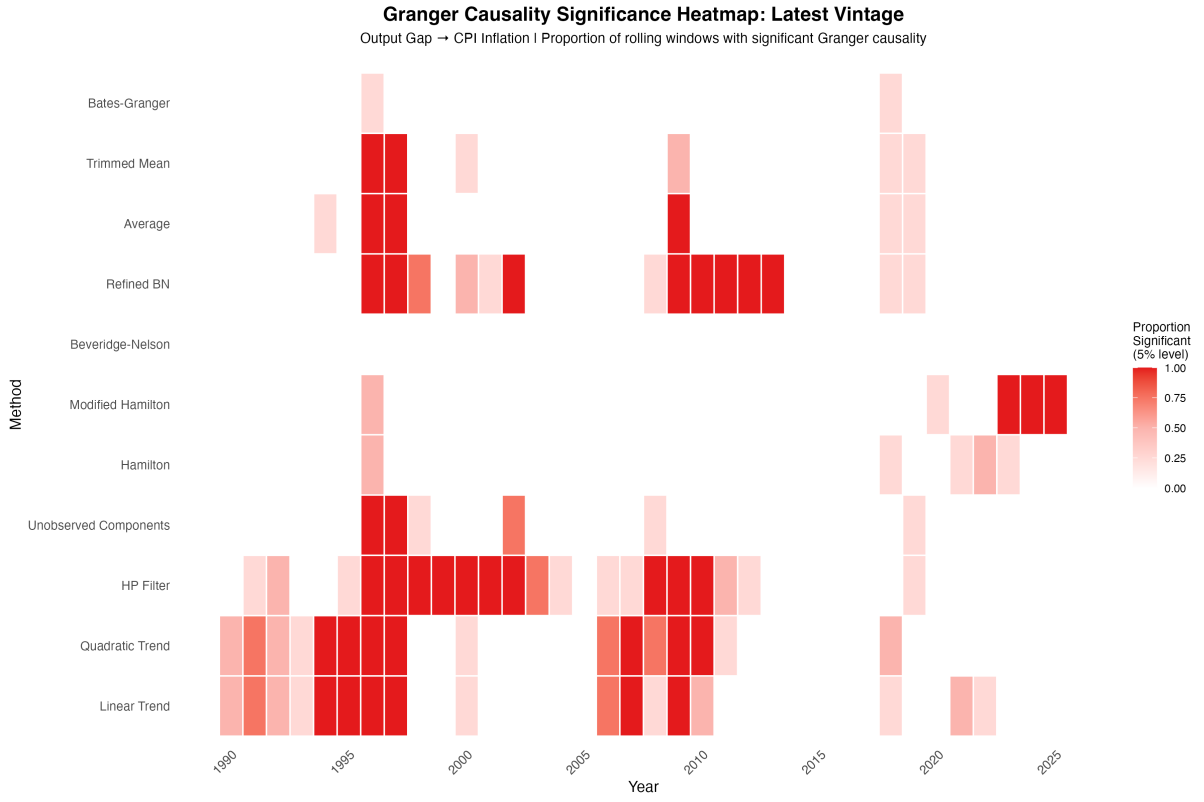
frequent significant relationships across multiple methods, particularly for trend-based approaches and the HP filter; (2) the mid-1990s through early 2000s exhibit reduced predictive power across most methods; (3) the period around the Global Financial Crisis (2007-2009) shows renewed significance for several methods, especially the HP filter and refined Beveridge-Nelson approaches; and (4) the post-2020 period demonstrates strong spikes in significance for the Beveridge-Nelson, refined BN, and Bates-Granger methods, coinciding with the extraordinary volatility of the pandemic era.

The patterns differ meaningfully across methods. The HP filter shows persistent predictive episodes throughout the sample, with particularly strong significance in the mid to late 1990s, around the 2008 crisis, and in the most recent period. The aggregated methods exhibit concen-

trated spikes of very high F-statistics around the recession periods. The Beveridge–Nelson and refined BN decompositions show especially strong predictive power in the latest vintage throughout the sample. In contrast, the linear and quadratic trend methods show their strongest performance in the early sample (1990s) and around 2008-2010, with limited significance in other periods. The Hamilton and modified Hamilton methods display episodic predictive power, with notable spikes in early 90s and post pandemic across data types.

To compare the time-varying evidence across methods, Figure 8 and Figure 9 present heatmaps showing the proportion of rolling windows with significant Granger causality at the 5% level.

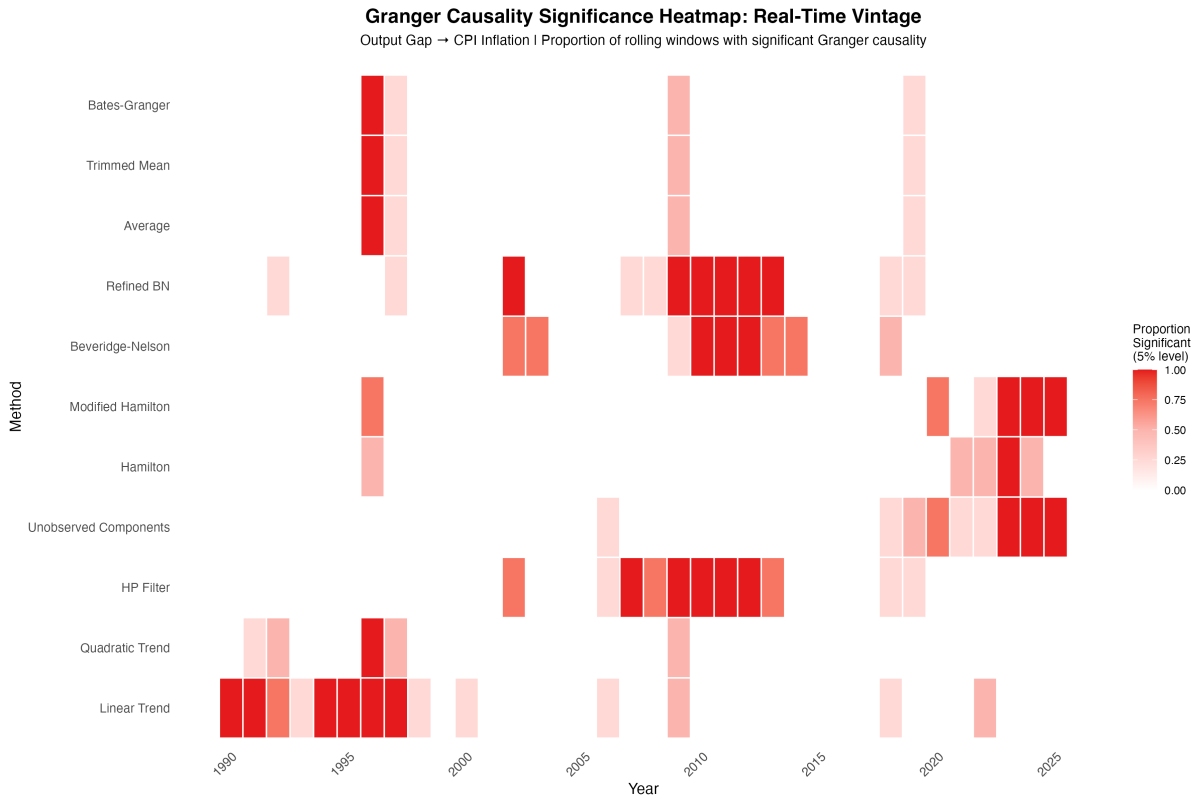
Figure 8: Significance Heatmap: Latest Vintage



Notes: The heatmap shows the proportion of 40-quarter rolling windows in which the Granger causality test rejects the null hypothesis (output gap does not cause inflation) at the 5% significance level, using latest-vintage data. Darker shading indicates a higher proportion of significant windows.

The heatmaps highlight several consistent patterns across methods and vintages. For the latest vintage, the 1990s show strong predictive power across most methods, with particularly high proportions of significant windows for the HP filter, trend-based methods, and refined Beveridge Nelson method. Hamilton filter and modified version, show concentrated significance in the most recent years (2020 onward). Trend-based methods (linear and quadratic) show their strongest performance in two distinct periods - the 1990s and around 2008-2010, but

Figure 9: Significance Heatmap: Real-Time Vintage



Notes: The heatmap shows the proportion of 40-quarter rolling windows in which the Granger causality test rejects the null hypothesis (output gap does not cause inflation) at the 5% significance level, using real-time data. Darker shading indicates a higher proportion of significant windows.

exhibit weaker predictive power in the intervening years and after 2010. The aggregates do not show consistent significance but scattered predictability over time. The HP filter demonstrates the most consistent temporal coverage, with meaningful proportions of significant windows spanning from the early 1990s through the most recent period, though with variation in intensity.

The real-time heatmap shows somewhat different patterns compared to the latest vintage. The Beveridge-Nelson and refined BN methods show overall strong predictive power, while other model based approaches such as Hamilton, modified Hamilton, and unobserved components methods strengthen in the recent periods. The aggregates remain inconsistent and the late 1980s and early 1990s significance appears somewhat attenuated for most methods, particularly for HP filter and trend-based approaches.

The results present a complex picture of output gaps' predictive content for inflation. The relationship between economic slack and future inflation appears both method-dependent and highly time-varying, with distinct patterns across different historical episodes rather than

showing monotonic decline. The choice of output gap estimation method matters substantially for inflation forecasting, and the optimal choice varies across periods. The strong recent performance of Beveridge-Nelson methods following the pandemic shock contrasts with the more consistent but less dramatic utility of the HP filter over time. Trend-based methods perform well primarily in the 1990s and around the Global Financial Crisis but show limited predictive power otherwise.

The time-varying patterns do not support a simple narrative of Phillips curve collapse. Instead, predictive power concentrates during specific episodes of macroeconomic stress: the early 1990s recession, the 2008-2009 crisis, and most dramatically, the post-COVID pandemic period. This indicated that output gaps may be most informative when the economy experiences substantial shocks or regime changes.

The comparison between latest-vintage and real-time results reveals important practical considerations. While some methods show similar patterns across vintages, others, particularly trend-based approaches in the early period, show notably weaker real-time performance. The recent strength of Beveridge-Nelson methods appears robust to vintage choice, suggesting these approaches may provide more reliable real-time signals during periods of high volatility. The dramatic resurgence of predictive power for several methods in the post-2020 period is particularly striking. Whether this represents a temporary phenomenon associated with extraordinary pandemic-era circumstances or a more lasting change in inflation dynamics remains an open question requiring further investigation.

These findings underscore the challenges facing policymakers who must assess inflationary pressures in real time. Output gaps can provide valuable information, but their usefulness depends critically on both the measurement approach and the macroeconomic context. The strong recent performance of several methods means renewed relevance of slack measures for understanding current inflation developments, though historical patterns indicate this predictive power may not persist indefinitely.

4.2 Forecasting Output Growth

As a final step, I evaluate whether output gap estimates improve forecasts of quarterly output growth relative to a simple autoregressive benchmark, using a pseudo-out-of-sample design. The initial estimation period ends in 1985:Q4. At each subsequent quarter through the end of the sample, an autoregressive distributed lag (ARDL) model is estimated, in which future

output growth depends on its own lags and on lags of an output gap measure. Lag lengths, up to four for each component, are selected recursively by minimizing the Bayesian Information Criterion within each estimation window. Using both real-time and latest-vintage data, the model is estimated separately for each of the eight decomposition methods. Given the substantial variation across gap estimates, I also construct forecasts for three aggregate measures. Forecasts are produced for horizons one through four quarters ahead. A univariate ARMA model, with its order also chosen recursively by BIC, serves as the benchmark. This procedure is repeated at each forecast origin, replicating a genuine real-time forecasting setting.

Figure 10: Output Growth Forecasts by Output Gap Methods - Latest Vintage



Notes: Pseudo-out-of-sample forecasts of annualized quarterly output growth at horizons $h = 1, 2, 3, 4$ quarters ahead. Each panel shows the actual growth rate and forecasts from eight individual output gap methods, using latest-vintage data.

Figure 10 and Figure 11 plot the forecast trajectories from the individual methods alongside actual output growth. Two features are evident. First, forecast dispersion across methods is substantial, especially around business-cycle turning points, with Beveridge-Nelson methods

Figure 11: Output Growth Forecasts by Output Gap Methods - Real Time



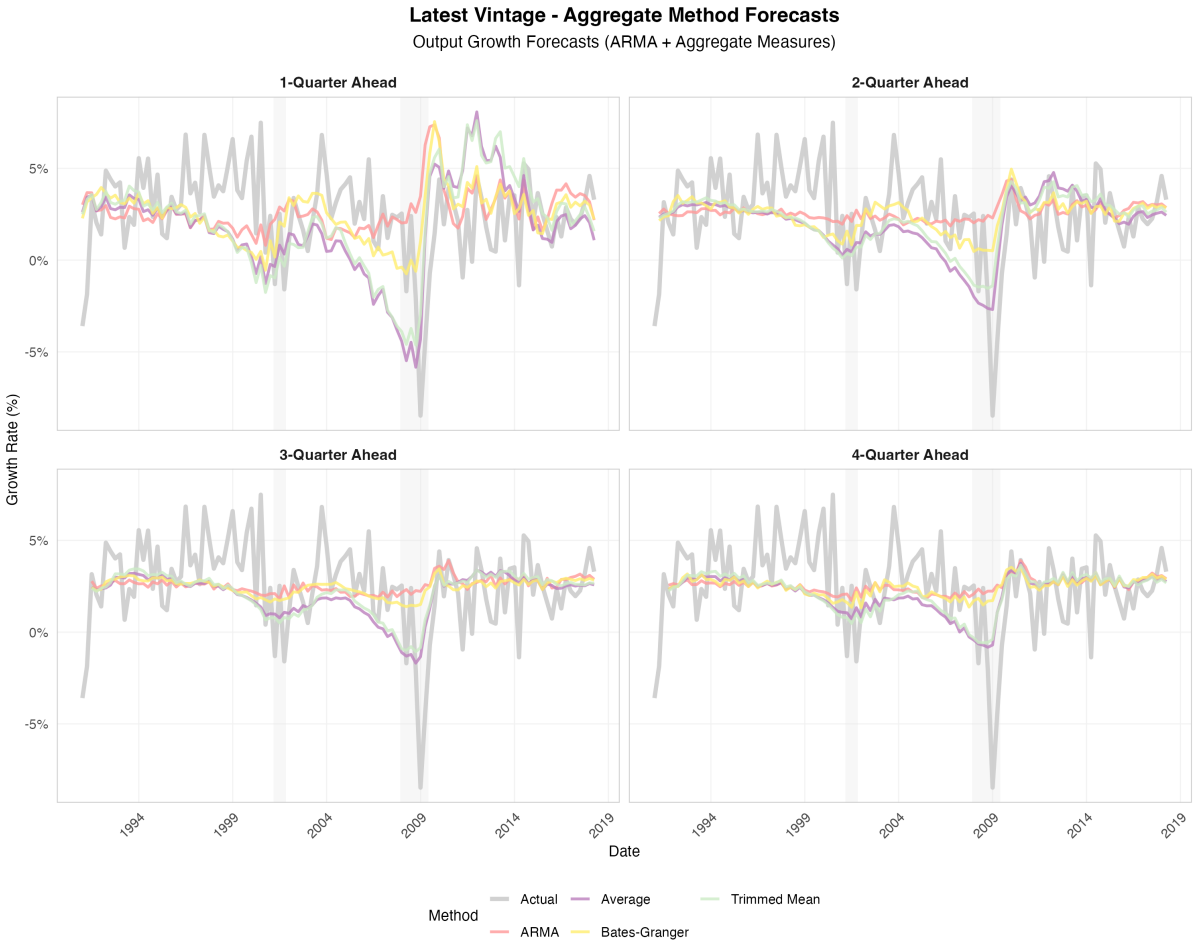
Notes: Pseudo-out-of-sample forecasts of annualized quarterly output growth at horizons $h = 1, 2, 3, 4$ quarters ahead. Each panel shows the actual growth rate and forecasts from eight individual output gap methods, using real-time data.

showing particularly volatile forecast paths. Second, dispersion is markedly larger for the real-time vintage, highlighting the additional uncertainty when relying on real-time estimates of economic slack.

The aggregate forecasts, shown in Figure 12 and Figure 13, are visibly smoother and track actual growth more reliably than most individual methods. The Bates–Granger aggregate, in particular, avoids extreme swings and maintains a relatively steady forecasting profile, even during recessions. This shows that combining multiple measures of slack can produce more robust predictions than relying on any single, potentially misleading, estimate.

To see forecasts performance, I evaluate these using root mean squared forecast errors (RMSE). Table 8 and Table 9 report relative RMSEs, the RMSE of each gap-augmented model divided by the RMSE of the ARMA benchmark, for the latest and real-time vintages, respec-

Figure 12: Output Growth Forecasts by Aggregate Output Gap Measures - Latest Vintage



Notes: Pseudo-out-of-sample forecasts of annualized quarterly output growth at horizons $h = 1, 2, 3, 4$ quarters ahead. Each panel shows the actual growth rate, the ARMA benchmark forecast, and forecasts from three aggregate output gap measures (simple average, trimmed mean, Bates-Granger), using latest-vintage data.

tively.

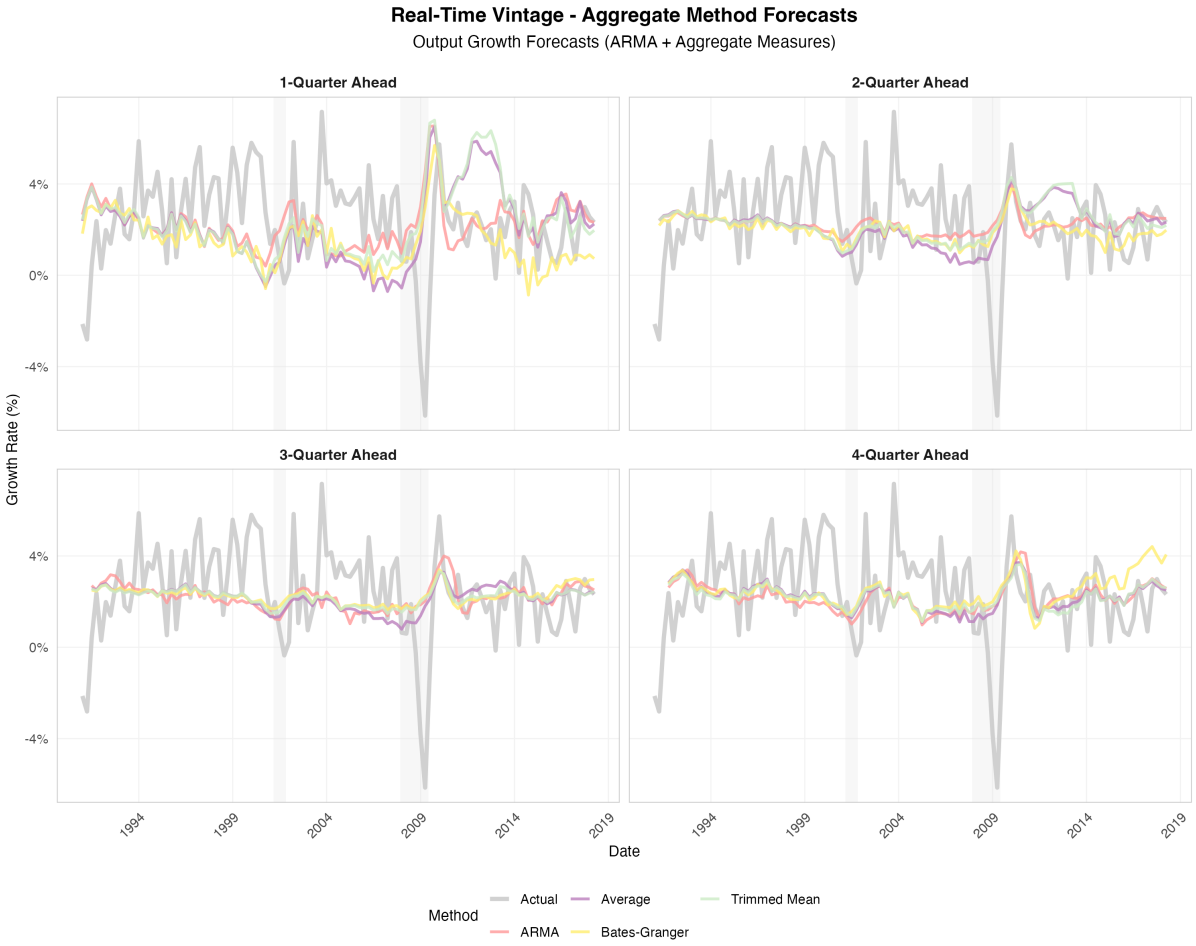
Table 8: Relative RMSE - Latest Vintage

Horizon	ARMA	Linear Trend	Quadratic Trend	HP Filter	Unobserved Components	Hamilton	Modified Hamilton	Beveridge-Nelson	Refined BN	Average	Trimmed Mean	Bates-Granger
1	1.000	1.201	1.103	0.951	0.991	0.976	0.984	1.012	1.102	1.115	1.107	**0.945**
2	1.000	1.027	0.984	**0.851**	0.985	0.999	1.010	1.034	1.028	1.008	0.982	0.955
3	1.000	0.977	0.940	**0.844**	0.944	0.979	0.986	1.021	1.011	0.954	0.941	0.957
4	1.000	0.972	0.942	**0.893**	0.938	0.993	0.992	0.997	1.001	0.958	0.949	0.978
Avg(1-4)	1.000	1.044	0.992	**0.885**	0.965	0.987	0.993	1.016	1.036	1.009	0.995	0.959

Notes: Relative root mean squared forecast error (RMSE) for h-quarter ahead output growth forecasts. Values are the ratio of the RMSE from a model including the specified output gap measure to the RMSE of a univariate ARMA benchmark. Values below 1 indicate superior performance relative to the benchmark. The *Avg(1-4)* row is the simple average across horizons 1–4.

For the latest vintage, the HP filter yields the lowest relative RMSE across horizons, with an average reduction of about 11.5% relative to the ARMA benchmark. The unobserved components model and the Bates–Granger aggregate also show stable gains, with average relative

Figure 13: Output Growth Forecasts by Aggregate Output Gap Measures - Real Time



Notes: Pseudo-out-of-sample forecasts of annualized quarterly output growth at horizons $h = 1, 2, 3, 4$ quarters ahead. Each panel shows the actual growth rate, the ARMA benchmark forecast, and forecasts from three aggregate output gap measures (simple average, trimmed mean, Bates-Granger), using real-time data.

Table 9: Relative RMSE - Real-Time Vintage

Horizon	ARMA	Linear Trend	Quadratic Trend	HP Filter	Unobserved Components	Hamilton	Modified Hamilton	Beveridge-Nelson	Refined BN	Average	Trimmed Mean	Bates-Granger
1	1.000	1.176	1.146	1.115	**0.965**	0.976	0.986	1.305	1.077	1.095	1.101	1.005
2	1.000	1.028	1.030	1.025	1.018	0.995	1.010	**0.993**	1.005	1.044	1.049	1.002
3	1.000	0.971	0.982	0.995	0.969	0.978	0.970	**0.966**	0.985	0.972	0.970	0.979
4	1.000	**0.956**	0.963	0.988	0.987	0.993	0.984	1.079	0.987	0.969	0.977	1.037
Avg(1-4)	1.000	1.033	1.030	1.031	**0.985**	0.986	0.988	1.086	1.013	1.020	1.024	1.006

Notes: Relative root mean squared forecast error (RMSE) for h-quarter ahead output growth forecasts using real-time data. Values are the ratio of the RMSE from a model including the specified output gap measure to the RMSE of a univariate ARMA benchmark. Values below 1 indicate superior performance relative to the benchmark. The *Avg(1-4)* row is the simple average across horizons 1–4.

RMSEs of 0.965 and 0.959 respectively. In contrast, Beveridge-Nelson and refined BN methods underperform the ARMA benchmark, with average relative RMSEs of 1.016 and 1.036. For the real-time vintage, the unobserved components model achieves the lowest average relative RMSE (0.985), while the Hamilton and modified Hamilton closely follow. Both Beveridge-

Nelson and refined BN methods perform poorly in real time as well, with an average relative RMSE well above 1. The HP filter, which performed strongly with latest-vintage data, shows no advantage in real time. This suggests that the endpoint uncertainty and data revisions inherent in real-time estimation weaken its predictive value.

The forecast combinations offer a robust middle ground. While seldom the single best performer at any horizon, the Bates–Granger aggregate consistently outperforms the worst individual methods and shows significant gains over Beveridge-Nelson approaches. The trimmed mean and simple average also avoid the poor performance of the weakest components. This reinforces the benefit of combining multiple methods to buffer against model uncertainty, particularly when individual methods like Beveridge-Nelson show high volatility. To formally test whether the improvements are statistically significant, I apply the Clark-West test for nested model comparison, evaluating whether the Bates–Granger forecasts are significantly more accurate than those from other methods. The results (Tables 10 & 11) indicate that the Bates-Granger forecasts are statistically superior to several alternatives, especially at shorter horizons. In the latest vintage, Bates-Granger significantly outperforms Beveridge-Nelson, refined BN, and trend-based methods across multiple horizons. In real time, the advantages are more modest but still significant against trend methods and Beveridge-Nelson at some horizons. These findings confirm that combining methods can yield not only more stable but also more accurate forecasts than relying on any single decomposition approach.

Table 10: Clark-West Test: Bates-Granger vs Other Methods - Latest Vintage

Horizon	ARMA	Linear Trend	Quadratic Trend	HP Filter	Unobserved Components	Hamilton	Modified Hamilton	Beveridge-Nelson	Refined BN	Average	Trimmed Mean
1	**	***	***	***	*	**	**	***	***	***	***
2		***	*		*		*	***	*	**	*
3	**	*						***	**		
4	*			**				*	*		

Notes: Clark-West test statistics evaluating whether forecasts from the Bates-Granger weighted combination are significantly more accurate than forecasts from the model listed in each column. A blank cell indicates the null hypothesis of equal predictive accuracy cannot be rejected. Significance levels: $***p < 0.01$, $**p < 0.05$, $*p < 0.10$.

Table 11: Clark-West Test: Bates-Granger vs Other Methods - Real-Time Vintage

Horizon	ARMA	Linear Trend	Quadratic Trend	HP Filter	Unobserved Components	Hamilton	Modified Hamilton	Beveridge-Nelson	Refined BN	Average	Trimmed Mean
1	**	***	***	***		**	*	***	***	***	***
2		**	***	*	*					**	**
3	**			***				*	**		
4			**					**			*

Notes: Clark-West test statistics evaluating whether forecasts from the Bates-Granger weighted combination are significantly more accurate than forecasts from the model listed in each column. A blank cell indicates the null hypothesis of equal predictive accuracy cannot be rejected. Significance levels: $***p < 0.01$, $**p < 0.05$, $*p < 0.10$.

Overall, the results show that the value of the output gap for predicting output growth is

highly contingent on how the gap is measured and whether real-time or revised data are used. Simple deterministic trends add little predictive power, while model-based filters like the HP filter and unobserved components can meaningfully improve forecasts with latest-vintage data. Beveridge-Nelson methods, despite showing strong predictive power for inflation, perform poorly for output growth forecasting, particularly in real time. Forecast combinations, particularly the Bates-Granger weighted average, provide consistently reliable performance by mitigating the weaknesses of individual methods, offering a practical hedge against the inherent uncertainty in measuring economic slack.

5 Conclusion

Output gap measurement remains deeply uncertain. Across eight decomposition methods applied to U.S. data from 1980 onward, I find stark differences in estimated slack. The differences that matter for both inflation prediction and growth forecasting. No single method dominates, and the choice between revised and real-time data compounds the policy challenge.

This paper finds some key insights that inform both academic research and practical policy analysis. There is considerable uncertainty in the measurement of economic slack. Output gap estimates vary dramatically across methods in terms of their amplitude, volatility, and even sign. For example, linear trend methods estimate a real-time output gap averaging -3.76%, while the HP filter yields -0.001% over the same period. over the same period. These aren't minor technical differences as they reflect distinct views about trend growth and cyclical fluctuations. Revision stability varies just as much. Methods like the refined Beveridge-Nelson decomposition, Hamilton and Modified Hamilton maintain high correlation between real-time and final data (0.933, 0.935 and 0.939 respectively), while quadratic trends shift substantially (correlation 0.503).

Whether gaps predict inflation depends critically on both the method and the period. The HP filter and modified Hamilton filter show significant Granger causality in certain episodes, but no method works consistently across the full sample. The slack-inflation relationship strengthens during macroeconomic stress, such as the early 1990s recession, the 2008–2009 financial crisis, and the post-2020 pandemic period. Rather than a stable Phillips curve, output gaps seem to provide periodic signals, most valuable precisely when the economy faces significant disruption.

For forecasting output growth, gaps prove useful but with important caveats. With revised data, the HP filter cuts forecast errors by roughly 11.5% relative to an autoregressive benchmark, while unobserved components and the Bates-Granger aggregate also show gains. Beveridge-Nelson methods, despite their strong inflation predictive power, actually worsen growth forecasts relative to the ARMA benchmark. In real time, the unobserved components model performs best, while the Bates-Granger aggregate shows only marginal improvement. Forecast combinations effectively blend multiple methods, particularly by mitigating the poor forecasting performance of Beveridge-Nelson approaches.

The findings show important implications for monetary policy. The choice of decomposition method is not merely technical as it can yield fundamentally different assessments of slack and thus different policy prescriptions. Real-time decisions face even sharper challenges, since preliminary data shift and filtering methods struggle at endpoints. The time-varying nature of the Phillips curve relationship suggests that policymakers should be cautious in assuming stable relationships between slack and inflation across different economic regimes, consulting multiple methods rather than betting on one. Forecast combinations, especially those that weight methods by recent performance, offer a practical hedge against specification error. And context matters as gaps appear most informative during periods of significant macroeconomic stress.

Some questions and paths remain open for future work, such as, can we develop decomposition methods more robust to real-time data limitations? Would incorporating labor market indicators or financial conditions sharpen slack measurement? How do predictive relationships evolve across different monetary policy regimes?

The challenges documented in this paper don't negate the value of output gaps. While no single method provides a definitive measure of economic slack, a careful comparison of multiple approaches, combined with an understanding of their limitations in real time, can yield meaningful insights about business cycle conditions and their implications for inflation and growth.

References

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*(1-2), 31–53.

Andriantomanga, Z., Kishor, N. K., & Kumar, L. (2025). Interest rate misalignments and monetary policy effects: Evidence from us states. *Available at SSRN 5254567*.

Atkeson, A., Ohanian, L. E., et al. (2001). Are phillips curves useful for forecasting inflation? *Federal Reserve bank of Minneapolis quarterly review*, *25*(1), 2–11.

Ball, L. M., & Mazumder, S. (2011). *Inflation dynamics and the great recession* (Tech. Rep.). National Bureau of Economic Research.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the operational research society*, *20*(4), 451–468.

Beveridge, S., & Nelson, C. R. (1981). 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*(2), 151–174.

Blanchard, O., & Galí, J. (2007). Real wage rigidities and the new keynesian model. *Journal of money, credit and banking*, *39*, 35–65.

Canova, F. (1998). Detrending and business cycle facts. *Journal of monetary economics*, *41*(3), 475–512.

Clark, P. K. (1987). The cyclical component of us economic activity. *The Quarterly Journal of Economics*, *102*(4), 797–814.

Clark, T. E., & McCracken, M. W. (2009). Combining forecasts from nested models. *Oxford Bulletin of Economics and Statistics*, *71*(3), 303–329.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, *138*(1), 291–311.

Cogley, T., & Sargent, T. J. (2001). Evolving post-world war ii us inflation dynamics. *NBER macroeconomics annual*, *16*, 331–373.

Coibion, O., & Gorodnichenko, Y. (2015). Is the phillips curve alive and well after all? inflation expectations and the missing disinflation. *American Economic Journal: Macroeconomics*, *7*(1), 197–232.

Croushore, D. (2006). An evaluation of inflation forecasts from surveys using real-time data.

Croushore, D. (2011). Frontiers of real-time data analysis. *Journal of economic literature*, *49*(1), 72–100.

Croushore, D., & Stark, T. (2001). A real-time data set for macroeconomists. *Journal of econometrics*, *105*(1), 111–130.

Del Negro, M., Giannoni, M. P., & Schorfheide, F. (2015). Inflation in the great recession and new keynesian models. *American Economic Journal: Macroeconomics*, *7*(1), 168–196.

Edge, R. M., & Rudd, J. B. (2016). Real-time properties of the federal reserve's output gap. *Review of Economics and Statistics*, *98*(4), 785–791.

Elliott, G., & Timmermann, A. (2013). *Handbook of economic forecasting*. Newnes.

Faust, J., & Svensson, L. E. (2001). Transparency and credibility: Monetary policy with unobservable goals. *International Economic Review*, *42*(2), 369–397.

Friedman, M. (1968). The role of monetary policy. *The American economic review*, *58*(1), 1–17.

Gordon, R. J. (1981). *Inflation, flexible exchange rates, and the natural rate of unemployment* (Tech. Rep.). National Bureau of Economic Research.

Gordon, R. J., Perry, G., Modigliani, F., Okun, A., Wachter, M., Kouri, P., … Sims, C. (1977). Can the inflation of the 1970s be explained? *Brookings Papers on Economic Activity*, *1977*(1), 253–279.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

Hamilton, J. D. (2018). Why you should never use the hodrick-prescott filter. *Review of Economics and Statistics*, *100*(5), 831–843.

Harvey, A. C. (1985). Trends and cycles in macroeconomic time series. *Journal of Business & Economic Statistics*, *3*(3), 216–227.

Harvey, A. C. (1990). *Forecasting, structural time series models and the kalman filter*. Cambridge university press.

Hazell, J., Herreno, J., Nakamura, E., & Steinsson, J. (2022). The slope of the phillips curve:

evidence from us states. *The Quarterly Journal of Economics*, *137*(3), 1299–1344.
Hodrick, R. J., & Prescott, E. C. (1997). Postwar us business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, 1–16.
Kamber, G., Morley, J., & Wong, B. (2018). Intuitive and reliable estimates of the output gap from a beveridge-nelson filter. *Review of Economics and Statistics*, *100*(3), 550–566.
Kamber, G., Morley, J., & Wong, B. (2025). Trend-cycle decomposition in the presence of large shocks. *Journal of Economic Dynamics and Control*, *173*, 105066.
Kishor, N. K., & Nguyen, N. (2025). Measuring the credit gap: a forecast combination approach. *Swiss Journal of Economics and Statistics*, *161*(1), 2.
Koenig, E. F., Dolmas, S., & Piger, J. (2003). The use and abuse of real-time data in economic forecasting. *Review of Economics and Statistics*, *85*(3), 618–628.
Marcellino, M., & Musso, A. (2011). The reliability of real-time estimates of the euro area output gap. *Economic Modelling*, *28*(4), 1842–1856.
Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of econometrics*, *135*(1-2), 499–526.
Morley, J. C. (2002). A state–space approach to calculating the beveridge–nelson decomposition. *Economics Letters*, *75*(1), 123–127.
Newbold, P., & Granger, C. W. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)*, *137*(2), 131–146.
Orphanides, A. (2001). Monetary policy rules based on real-time data. *American Economic Review*, *91*(4), 964–985.
Orphanides, A. (2008). Taylor rules. In *The new palgrave dictionary of economics* (pp. 1–7). London: Palgrave Macmillan.
Orphanides, A., & Norden, S. v. (2002). The unreliability of output-gap estimates in real time. *Review of economics and statistics*, *84*(4), 569–583.
Phelps, E. S. (1968). Money-wage dynamics and labor-market equilibrium. *Journal of political economy*, *76*(4, Part 2), 678–711.
Phillips, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the united kingdom, 1861-1957. *economica*, *25*(100), 283–299.
Quast, J., & Wolters, M. H. (2022). Reliable real-time output gap estimates based on a modified hamilton filter. *Journal of Business & Economic Statistics*, *40*(1), 152–168.
Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of monetary economics*, *44*(2), 293–335.
Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of forecasting*, *23*(6), 405–430.
Stock, J. H., & Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, *39*, 3–33.
Stock, J. H., & Watson, M. W. (2008). Phillips curve inflation forecasts.
Timmermann, A. (2006). Forecast combinations. In *Handbook of economic forecasting* (Vol. 1, pp. 135–196). Elsevier. doi: 10.1016/S1574-0706(05)01004-9
Watson, M. W. (2014). Inflation persistence, the nairu, and the great recession. *American Economic Review*, *104*(5), 31–36.