FEDERAL RESERVE BANK
PHILADELPHIA

Research Brief
RESEARCH DEPARTMENT

Forecast Accuracy and Forecaster Disagreement in the Survey of Professional Forecasters

Patrick Doelp
Federal Reserve Bank of Philadelphia

PUBLISHED
December 2023

DOI: https://doi.org/10.21799/frbp.rb.2023.dec.31

Forecast Accuracy and Forecaster Disagreement in the Survey of Professional Forecasters

Patrick Doelp

Patrick Doelp is a research associate in the Research Department's Real-Time Data Research Center of the Federal Reserve Bank of Philadelphia.

The views expressed by the author are not necessarily those of the Federal Reserve.

Research Briefs are timely reports on economic topics of interest to researchers and the public written by economists and analysts in our Research Department.

THE SURVEY OF PROFESSIONAL FORECASTERS (SPF) HAS BEEN an important and well-known forecasting tool for economists and policymakers for over 50 years. The survey provides projections for important economic measures such as real GDP, inflation, unemployment, and interest rates. Economic forecasts can, however, be inaccurate because the models on which the forecasts rely do not include all the important features of the economy. Forecasters are always looking for ways to improve the quality of their models and forecasts, hoping to identify the reasons for the inaccuracies in their projections. The Federal Reserve Bank of Philadelphia, which has produced the SPF since the early 1990s, has a long history of tracking the accuracy of the SPF's forecasts to better understand the reasons for forecast inaccuracy and how the accuracy changes over time. See Stark (2010), Mboup and Wurtzel (2021), and Doelp and Mboup (2021) for recent examples.

In this research brief, I will add to the Philadelphia Fed's literature on the SPF's forecast accuracy by studying if rising forecast disagreement between SPF panelists leads to relatively poorer SPF forecast accuracy when compared to the forecast of a benchmark model. I conclude that where I find a difference in forecast accuracy between the SPF and a benchmark model, conditional on forecaster disagreement, the SPF becomes the more relatively accurate forecast as forecaster disagreement rises.

Data

The SPF forecasts numerous variables measuring alternative types of economic activity.[1] I have selected nine variables from the SPF that measure the activities of consumers, businesses, and governments. These variables are industrial production, nominal GDP, real GDP, GDP price index, real federal government consumption and gross investment, real state and local government consumption and gross investment, real personal consumption expenditures, real residential fixed investment, and real

1 All Survey of Professional Forecasters data come from the Real-Time Data Research Center, Federal Reserve Bank of Philadelphia, https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters.

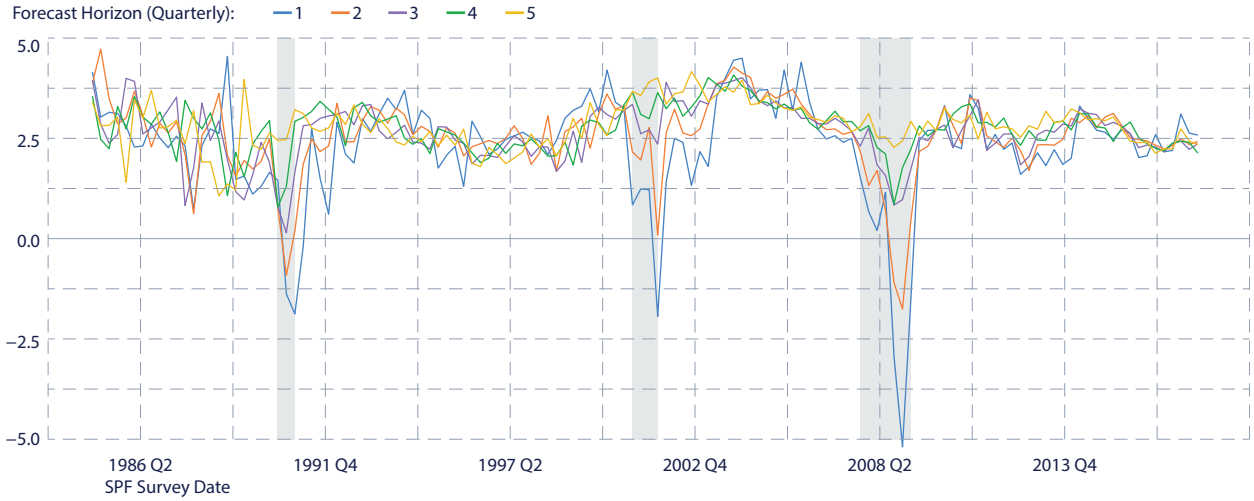
nonresidential fixed investment.[2] This paper focuses on the SPF surveys conducted over the period 1985 Q1 to 2017 Q4 (inclusive).

The SPF is conducted quarterly, and all variables are forecast over the five quarters running from the quarter when the survey was conducted to the quarter that follows four quarters into the future. The forecast for the survey's current quarter is the first horizon forecast, and the forecasts for the following four quarters are the second through fifth horizons. Figure 1 shows the SPF's real GDP growth forecasts over the various horizons. The shorter horizon forecasts tend to have the most variation over time, while longer horizon forecasts tend to be more stable. The figure shows three recessions over the sample period (shaded areas), and the majority of the time is in expansion. For a more detailed discussion about the SPF and its history, see Croushore and Stark (2019).

FIGURE 1

SPF Forecasts for Real GDP Growth

Growth rate (annualized); 1985 Q1–2017 Q4



Note: Shading reflects National Bureau of Economic Research (NBER) recession dating, retrieved via Haver Analytics.

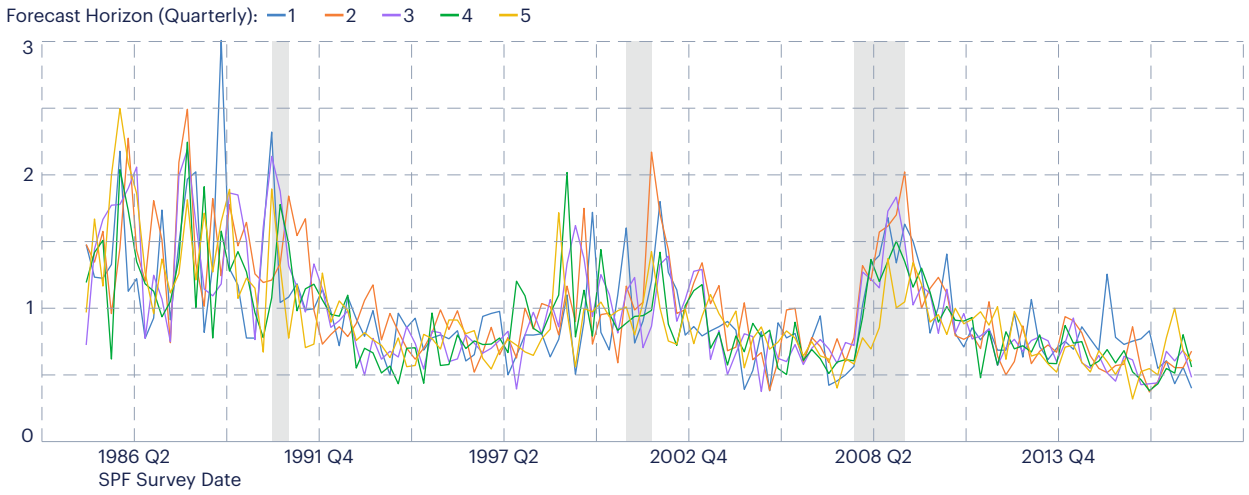
My objective is to study whether or not forecaster disagreement affects the relative accuracy of the SPF forecasts compared with the forecasts of a benchmark model. I measure forecaster disagreement as one of the survey's published measures of statistical dispersion, defined as the difference between the 75th and 25th percentiles of the SPF panelists' responses. The SPF measures disagreement in each survey, variable, and forecast horizon, so I can track how disagreement evolves over time. Figure 2 shows disagreement about future real GDP growth and how it varies over time and forecast horizon. Disagreement spikes during periods of recession. Spikes in disagreement also occur during periods of nonrecession, which is especially evident in real GDP during the 1980s. Sill (2014) discusses additional features of the survey's measure of forecaster disagreement.

2 All variables are expressed as percentage points for quarter-over-quarter annualized growth rates.

FIGURE 2

SPF Dispersion of Real GDP Growth

Growth rate (annualized); 1985 Q1–2017 Q4



Note: Shading reflects National Bureau of Economic Research (NBER) recession dating, retrieved via Haver Analytics.

Real-time historical data from the Philadelphia Fed's Real-Time Data Set for Macroeconomists (RTDSM) form the bedrock of my analysis.[3] I use the RTDSM's real-time historical vintage data to estimate and forecast a benchmark model (which I abbreviate as "BMK"). Real-time historical vintage data are snapshots of the entire history of an economic measure before the measure is fully revised by the U.S. government statistical authority. The vintage data in this analysis allows me to estimate and forecast the BMK model as if I were working in the same time period as a panelist in the SPF. This approach ensures that the BMK forecasts do not have a statistical or informational advantage.

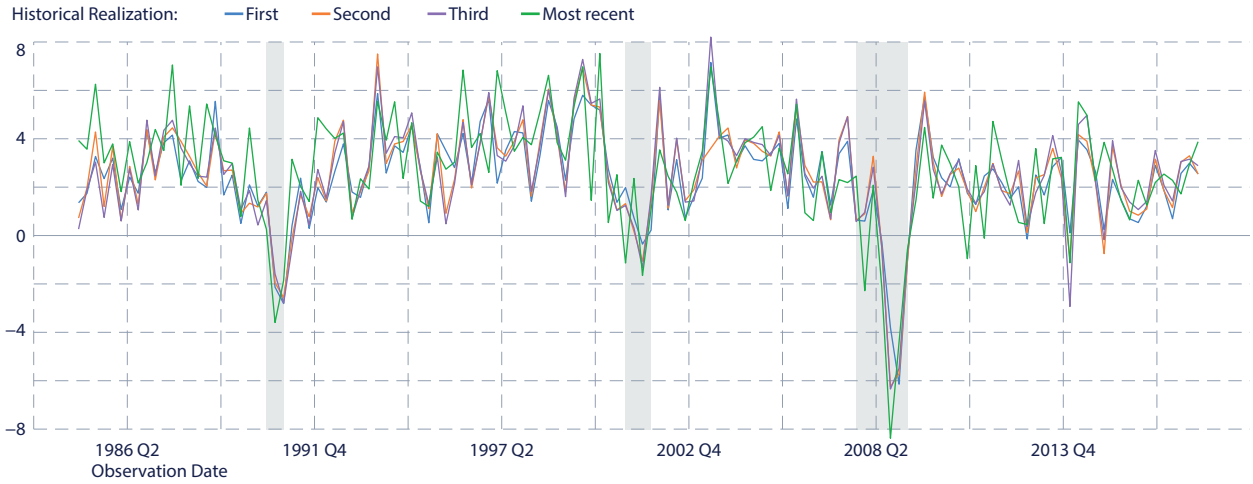
The defining characteristic of real-time historical data is that these data reflect the revisions made by the U.S. government statistical agencies over time. Preliminary or early published values reflect little or no revision and are generally thought to be less reliable than the estimates to be published much later. These alternative versions of historical values, also known as realizations, give rise to an important question about which revision is best to use for measuring the accuracy of a forecast. Figure 3 shows that revisions to the historical data can be large, perhaps large enough to affect the results of any forecast evaluation exercise, including my results. Many large revisions occur between the first release of the data and the most recent release, but noticeable revisions also happen between the first, second, and third releases. My analysis uses each alternative measure of the realization shown in Figure 3 in assessing forecast accuracy.

3 A small number of real-time vintages are missing data due to federal government shutdowns affecting U.S. government statistical agencies' ability to publish or due to additional issues. These vintages are replaced with the first future vintage that includes the missing values. All Real-Time Data Set for Macroeconomists data come from the Real-Time Data Research Center of the Federal Reserve Bank of Philadelphia, https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/real-time-data-set-for-macroeconomists.

FIGURE 3

Real GDP Historical Observations

Growth rate (annualized); 1985 Q4–2017 Q4



Note: Shading reflects National Bureau of Economic Research (NBER) recession dating, retrieved via Haver Analytics.

Benchmark Forecast

I have elected to use a standard AR(1) model as the benchmark (BMK) forecast against which I will judge the SPF accuracy in periods of high and low forecaster disagreement. An AR(1) model is a regression of an economic variable against the observation of itself one period prior. The mathematical form of the AR(1) model is:

$$BMK_{t,j} = C_j + \gamma_j BMK_{t-1,j} + e_{t,j}$$

The letter j indexes the variable used in this paper, C_j is a constant, γ_j is the coefficient estimated via MLE, and $e_{t,j}$ is the error. I estimate this model using a fixed window of 60 periods.

An Unconditional Forecast Accuracy Test

The Diebold and Mariano (1995) (DM) test is the gold standard for testing the unconditional accuracy between any two forecasts and is the foundation on which the conditional test I use in this paper is constructed. The DM test examines the accuracy of two competing forecasts to see if there is a statistical difference in accuracy on average. All DM-style forecast accuracy tests, including the conditional test I use for this paper's main results, require the researcher to adopt a "loss function" describing how to measure forecast accuracy. I will use the mean squared error loss function throughout this paper, defined as:

$$L_{t+\tau}\left(forecast_{t+\tau,j}, history_{t+\tau,j,r}\right) = \left(history_{t+\tau,j,r} - forecast_{t+\tau,j}\right)^2$$

The notation $L_{t+\tau}(\cdot)$ represents a general loss function, where τ represents the forecast horizon, j represents a particular variable, and r represents a historical realization. Formally, the hypothesis for the DM test is:

$$H_0: E\left[\Delta L_{t,\tau,r}\right] = 0$$

$$H_1: E\left[\Delta L_{t,\tau,r}\right] \neq 0$$

$$\Delta L_{t,\tau,r} = L_{t+\tau}\left(SPF_{t+\tau,j}, history_{t+\tau,j,r}\right) - L_{t+\tau}\left(BMK_{t+\tau,j}, history_{t+\tau,j,r}\right)$$

The null (H_0) and alternative (H_1) hypotheses respectively represent the cases where the difference between the SPF and BMK loss function is statistically zero (i.e., the forecast accuracy is the same) or the difference between the loss function is not zero (i.e., the forecast accuracy is different). There are several different methods for computing a DM test. I elect to do so using the regression method, which uses OLS to estimate the following regression:

$$\Delta L_{t,\tau,r} = d + \xi_t$$

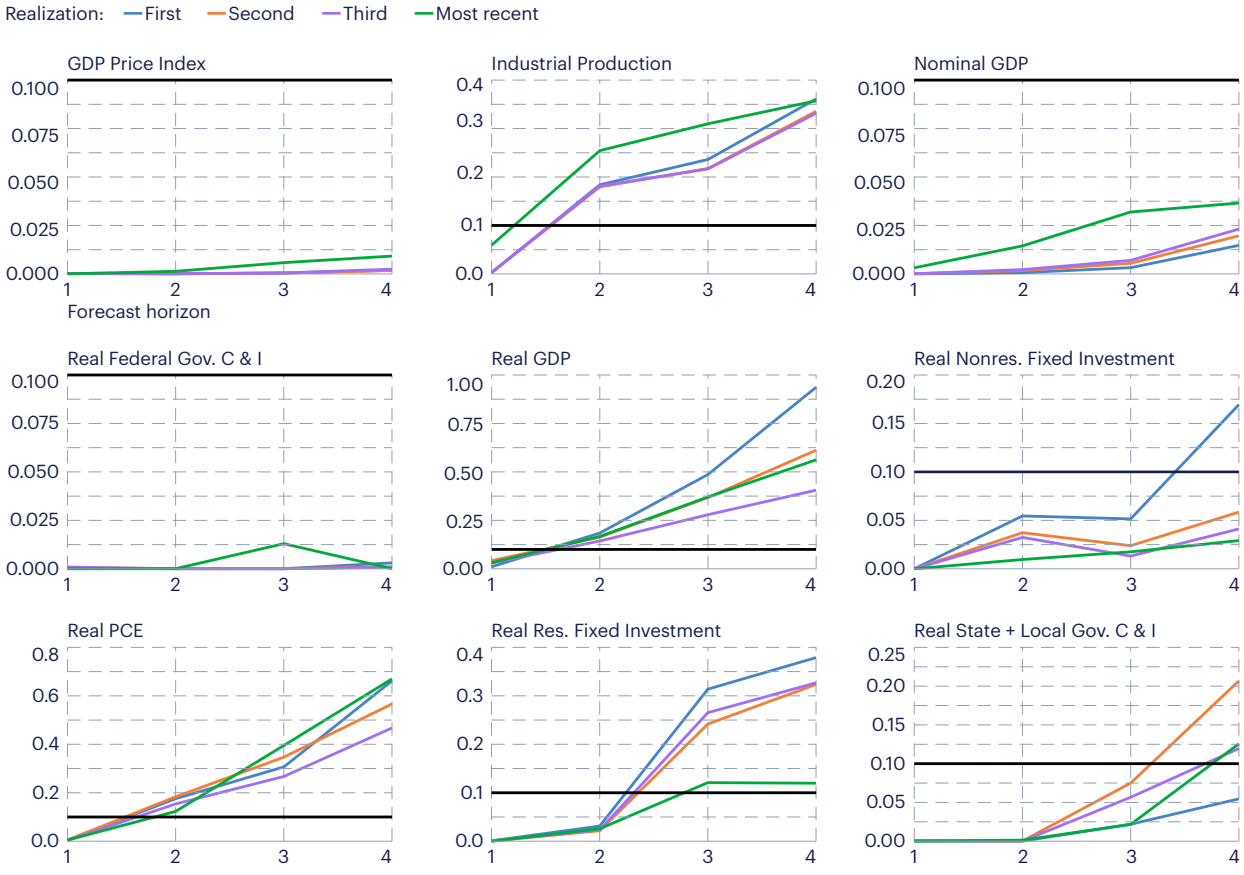
The notation d is a constant and ξ_t is the error. I have computed the DM test using the BMK model forecasts and the SPF forecasts and display the resulting p-values in Figure 4.[4] Low p-values—say, less than 0.10—indi-

4 I use the Newey–West HAC method to compute the variance using a truncation parameter of $2 * (\tau - 1)$.

FIGURE 4

Diebold-Mariano Test Results

P-Values; 1985 Q1–2017 Q4



cate a rejection of the null hypothesis of the equality in forecast accuracy between the SPF and BMK forecasts.

The DM test finds two patterns in the difference of forecast accuracy across variables. The first pattern can be seen in variables like nominal GDP, where the test finds a difference in forecast accuracy at all horizons and realizations of the data. The second pattern can be seen in variables like real residential fixed investment, where the DM test finds a difference in forecast accuracy at shorter horizons but no difference at longer horizons.

There are several exceptions to the patterns present in the DM test results that come in the form of one realization of the data for a particular forecast horizon not behaving like the other realizations. This exception to the pattern is best exemplified in real nonresidential fixed investment, where the first realization of the data at the fourth horizon does not produce a statistically meaningful result; however, the rest of the realizations are statistically meaningful. These exceptions emphasize the use of all four realizations of the data. Only using the first realization of the data would give us a different impression of forecast accuracy in real nonresidential fixed investment than using all four realizations.

A Conditional Accuracy Test

Giacomini and White Forecast Accuracy Test

The Giacomini and White (2006) (GW) test forms the statistical foundation for my analysis. It builds on the DM test and offers a way to examine forecast accuracy during the periods characterized by such special events as high forecaster disagreement. I adopt the same loss function I used for the DM test—mean squared error—for the GW test. The null and alternative hypotheses for the GW test are:

$$H_0: E\left[\Delta L_{t,\tau,r} | h_t\right] = 0$$

$$H_1: E\left[\Delta L_{t,\tau,r} | h_t\right] \neq 0$$

The only difference between the GW and DM hypothesis tests is the conditioning information, represented by h_t in the above notation. The GW test uses the conditioning data to give a higher or lower relative weight to observations of the loss function depending on where h_t itself has higher or lower values. Notably, the GW test is equivalent to the DM test when we give h_t an equal weighting by setting each observation of h_t equal to one.

The variable h_t in the test represents the measures of forecaster disagreement. I am interested in only using data for this study that SPF forecasters would have had available when they computed their forecast.[5] This includes the conditioning data for forecaster disagreement. For this reason, I set h_t equal to the dispersion data from the SPF survey conducted immediately prior to any given survey.

5 Each SPF forecasts five quarters into the future, so the fifth horizon forecast does not have a match from the previous survey's dispersion. I must thus restrict my study to the first through fourth forecast horizons for the SPF and BMK forecasts.

My statistical results use the standard GW test statistic given by:[6]

$$GW = n\left(n^{-1}\sum_{t=m}^{T-\tau} h_t\, \Delta L_{t,\tau,r}\right)' \Omega^{-1} \left(n^{-1}\sum_{t=m}^{T-\tau} h_t\, \Delta L_{t,\tau,r}\right)$$

The variable n is the number of observations, T is the last time period in the history, m is the maximum sample size of the two forecasts, and Ω is a weighted variance matrix.[7]

GW Regression

A rejection of the GW test's null hypothesis suggests a difference in the conditional accuracy of the two forecasts. However, unlike the DM test, the GW test does not include a constant, which means the GW test does not account for an average difference in forecast accuracy. The GW test also does not indicate which of the two forecasts is more relatively accurate. Giacomini and White suggest a method for determining which forecast is more accurate. Their method, which I will refer to as the GW regression to distinguish it from the GW test, begins by estimating the following regression via OLS:

$$\Delta L_{t,\tau,r} = c + \beta h_t + \epsilon_t$$

The method suggests the following decision rules for determining how dispersion affects relative forecast accuracy:

- If $\beta < 0$, the SPF's relative forecast accuracy improves as disagreement increases compared to the BMK model's forecast accuracy.

- If $\beta > 0$, the BMK model's relative forecast accuracy improves as disagreement increases compared to the SPF's forecast accuracy.

I compute p-values for the GW regression to test if β is statistically indistinguishable from a value of zero. When β is no different from zero, I can conclude that the relative forecast accuracy of the SPF is unrelated to forecaster dispersion. However, when the value of β is statistically different from zero, I can conclude that either the relative accuracy of the SPF forecast increases (if $\beta < 0$) or the relative accuracy of the BMK forecast increases (if $\beta > 0$) as forecaster disagreement rises. Moreover, I use the GW regression to compute an empirical cut-off value, which I refer to as h_{cut}, showing how high the forecaster disagreement must be for the accuracy of one of the two forecasts to exceed the accuracy of the other. I calculate the cut-off value by setting ΔL_t equal to zero, which yields the following equation:

$$h_{cut} = -\frac{c}{\beta}$$

6 The p-values are calculated from a chi-squared distribution with degrees of freedom set to the number of variables in the conditioning data (h_t), which, in this case, is equal to one.

7 The variance matrix Ω is computed via HAC using Newey–West's method. The truncation parameter is set to $2 * (\tau - 1)$.

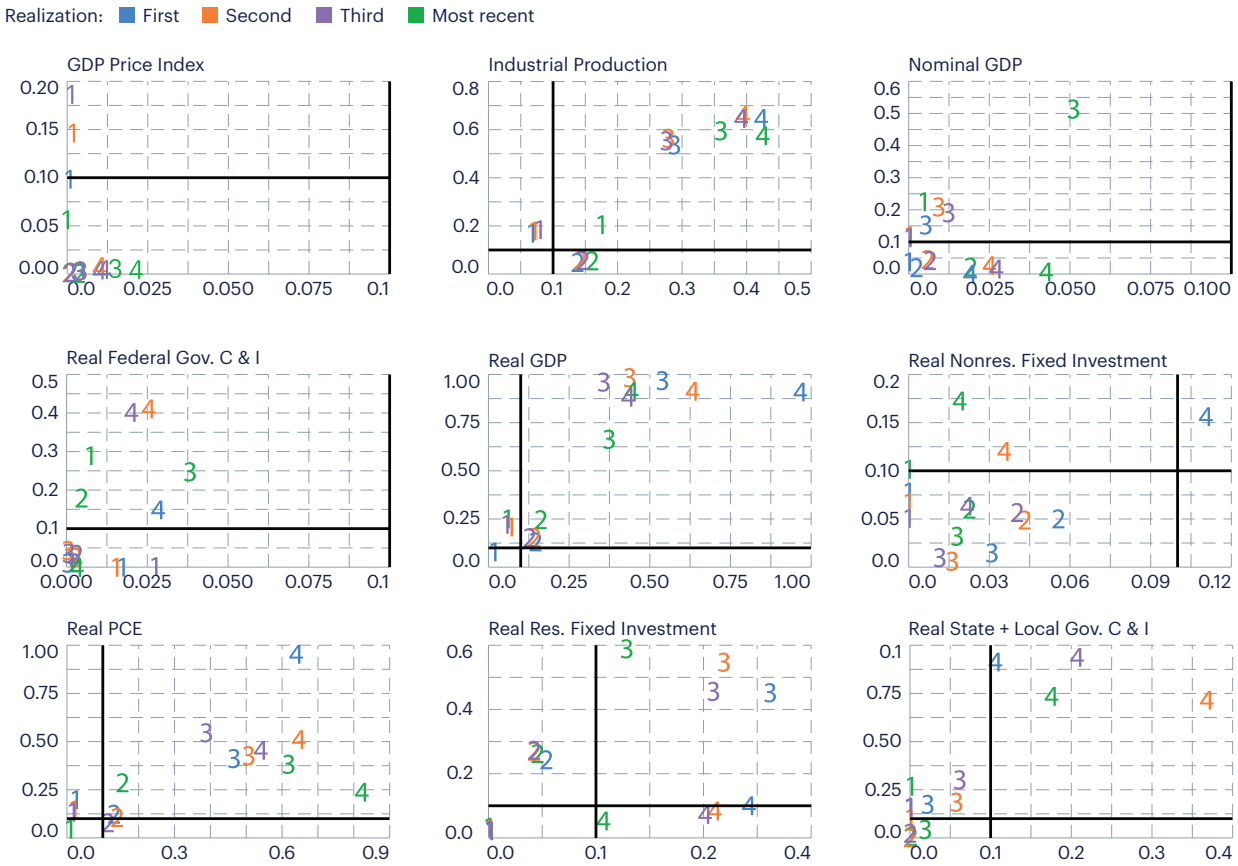
Results

I have plotted the p-values from the GW test and the GW regression in Figure 5. The vertical black line represents the 0.10 significance threshold for the GW test and the horizontal black line represents the same threshold for the GW regression. Any point shown in the bottom-left quadrant formed by these two threshold lines indicates a conditionally statistically meaningful result for both the GW test and the GW regression. Unless otherwise specified, when I discuss a conditionally statistically meaningful result, I am referring to results that are conditionally statistically meaningful in forecast accuracy for both the GW test and the GW regression. Additionally, I have computed the GW test and GW regression over the first half and second half of the sample as a robustness check.

FIGURE 5

Decision Rule Regression P-Values vs. GW P-Values

Full Sample: 1985 Q1–2017 Q4



Note: Plotted numbers represent forecast horizon; verticle and horizontal lines represent p-values equal to 0.10.

Each sample has a total of 144 tests across nine variables, four horizons, and four realizations of the historical data. The GW test by itself finds 94 (65.3 percent) statistically meaningful tests for the full sample, 78 (54.2 percent) for the first-half sample, and 67 (46.5 percent) for the second-half sample. Using results from both the GW test and the GW re-

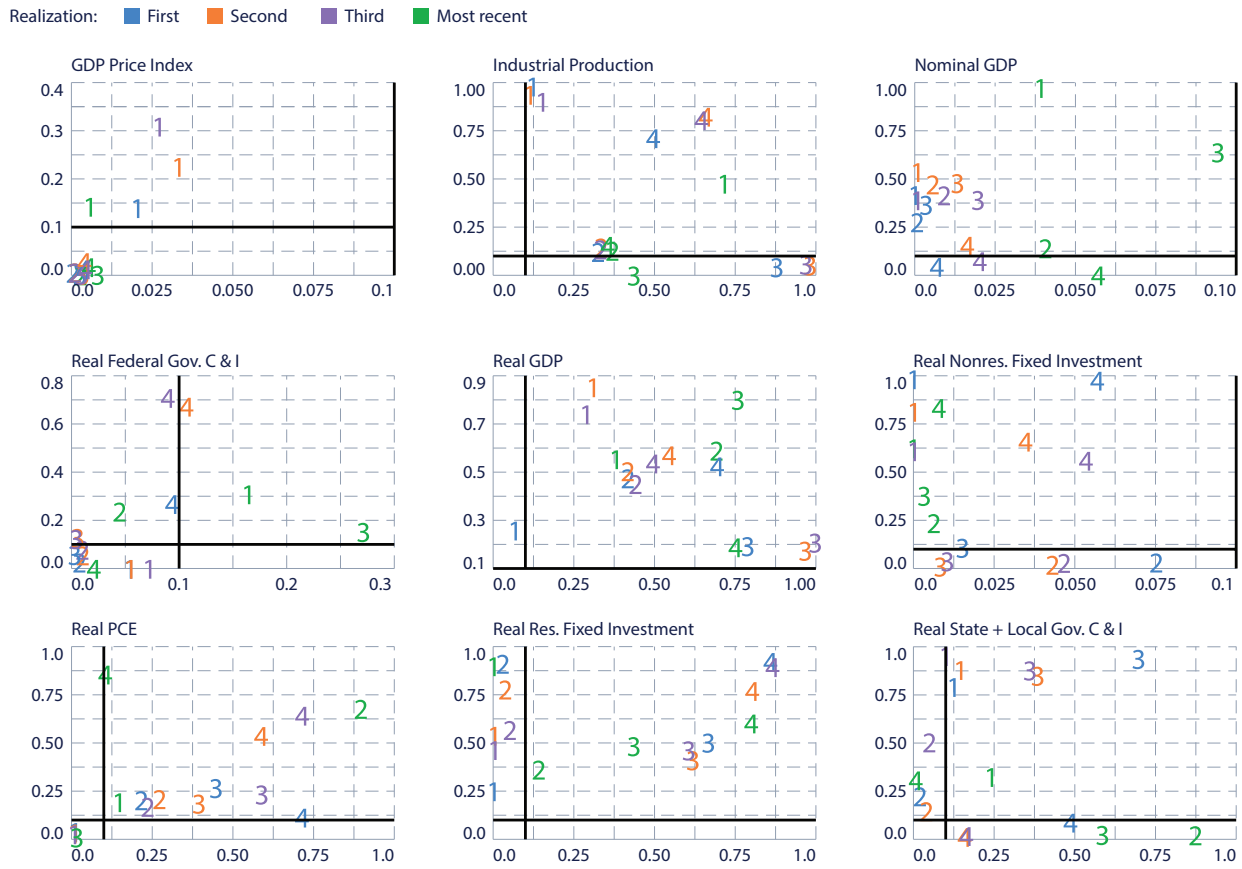
gression, I find 56 (38.9 percent) total tests with conditionally statistically meaningful differences in the full sample, 32 (22.2 percent) tests conditionally statistically meaningful for the first-half sample, and 44 (30.6 percent) for the second-half sample.

The GW regression, using the method that Giacomini and White suggested to determine which forecast is more accurate, additionally finds that the SPF is the more relatively accurate forecast conditional on forecaster disagreement for the vast majority of conditionally statistically meaningful results. Only a single result (real PCE, third forecast horizon, most recent realization, first-half sample) shows the BMK forecast as the more relatively accurate forecast.

FIGURE 6

Decision Rule Regression P-Values vs. GW P-Values

First-Half Sample: 1985 Q1–2000 Q4



Note: Plotted numbers represent forecast horizon; verticle and horizontal lines represent p-values equal to 0.10.

The variables most commonly showing conditionally statistically meaningful results in the full sample are GDP price index, real federal government C&I, real nonresidential fixed investment, and nominal GDP. Industrial production, real GDP, and real PCE have the least amount of conditionally statistically meaningful results, with most of the results for these variables located in the upper-right quadrant of Figure 5. Real GDP has few conditionally statistically meaningful results, while nom-

inal GDP and the GDP price index both have many, suggesting that the GDP price index is driving the conditional forecast accuracy of the three variables. The full sample's second horizon has the largest number of conditionally statistically meaningful results out of any sample and horizon combination, with a total of 19. The GDP price index at the first and most recent realizations of the historical data and real nonresidential fixed investment at the third realization are the only three realizations to have all four possible results be conditionally statistically meaningful in the full sample.

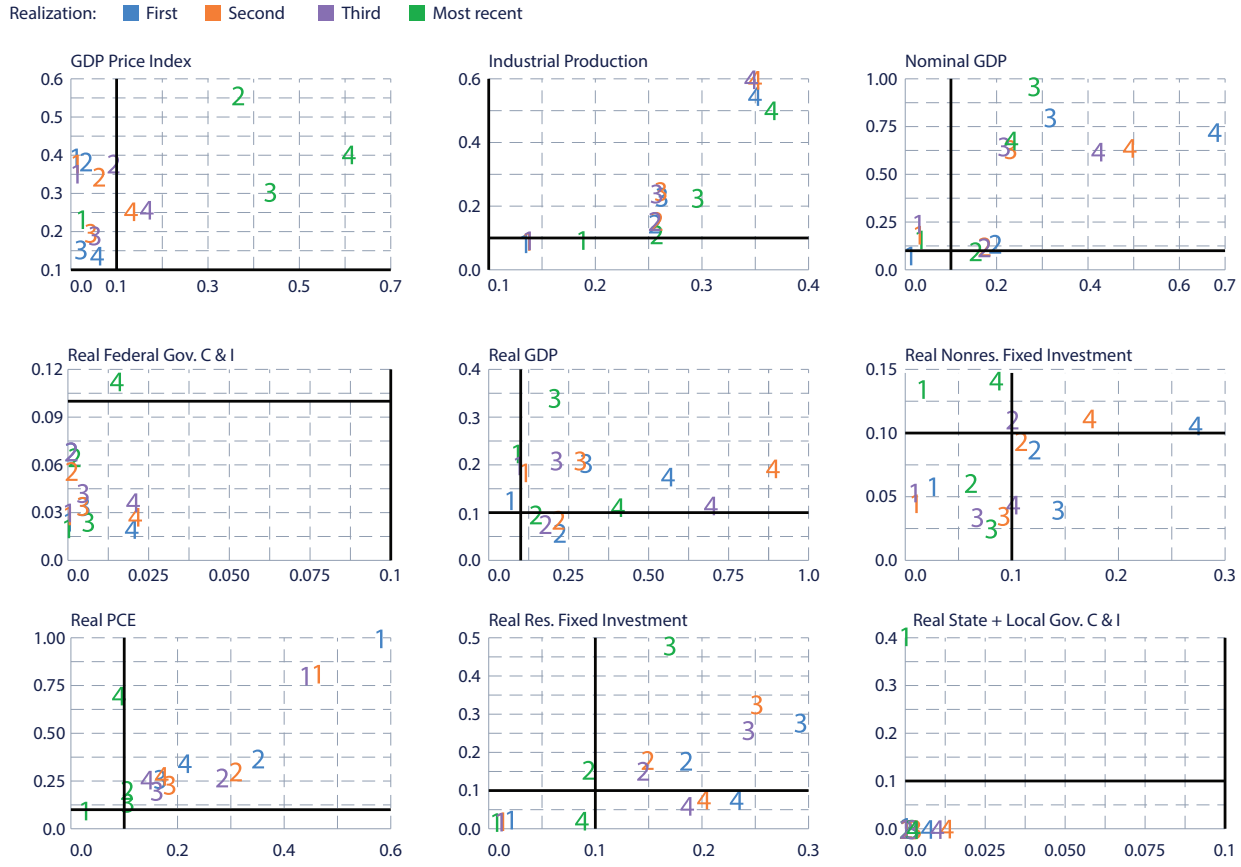
As shown in Figure 6, the first-half sample has far fewer conditionally statistically meaningful results than all other samples, at 32 total. Real PCE is the only variable that has more conditionally statistically meaningful results in the first-half sample than in the full sample.

The results for the second-half sample are displayed in Figure 7. The most notable difference between the second-half sample and the other two samples is that the GDP price index, which had a large number of conditionally statistically meaningful results in the first-half (12) and full sample (14), has no statistically meaningful results in the second-half sample.

FIGURE 7

Decision Rule Regression P-Values vs. GW P-Values

Second-Half Sample: 2001 Q1–2017 Q4



Note: Plotted numbers represent forecast horizon; verticle and horizontal lines represent p-values equal to 0.10.

Table 1: Summary Statistics of Cut-Off Values for Full Sample (1985 Q1–2017 Q4)[1]

	First Horizon		Second		Third		Fourth	
Variable	Median (min | max)	N	Median (min | max)	N	Median (min | max)	N	Median (min | max)	N
GDP Price Index	0.328 (0.318 | 0.339)	2	0.431 (0.398 | 0.481)	4	0.437 (0.421 | 0.455)	4	0.485 (0.468 | 0.514)	4
Nominal GDP	−0.002 (−0.002 | −0.002)	1	0.553 (0.529 | 0.787)	4			0.479 (0.382 | 0.739)	4
Real Federal Gov. C&I	2.094 (2.032 | 2.246)	3	1.257 (1.124 | 1.390)	3	0.940 (0.756 | 1.195)	3	0.592 (0.592 | 0.592)	1
Real GDP	0.545 (0.545 | 0.545)	1						
Real Nonres. Fixed Investment	0.536 (0.463 | 0.704)	3	2.143 (1.903 | 2.265)	4	2.063 (1.740 | 2.227)	4	1.340 (1.340 | 1.340)	1
Real PCE	0.452 (0.452 | 0.452)	1						
Real Res. Fixed Investment	4.102 (3.720 | 4.356)	4						
Real State + Local Gov. C&I			0.502 (0.385 | 0.524)	4	0.618 (0.618 | 0.618)	1		

1 Table only displays tests where the GW test and GW regression are statistically meaningful.

Cut-Off Values for Forecaster Disagreement

It is of interest to estimate the value of forecaster disagreement at which one of the forecasts, SPF or BMK, becomes more accurate than the other. More precisely, at levels of disagreement higher than my estimated "cut-off" value, one forecast dominates the other in accuracy. These cut-off values carry particular significance because they provide an insightful signal, based on the magnitude of forecaster disagreement, about the conditions under which one forecast performs particularly well. This signal speaks to the overall quality of the forecast during periods where forecast quality might come into question, such as periods with high forecaster disagreement.

My summary statistics for cut-off values (computed only when the GW test and the GW regression show statistically meaningful results) for the

full sample period are displayed in Table 1.[8] The table presents the median estimated cut-off values for each variable and forecast horizon, with the median computed across my four alternative measures of the historical realizations. The minimum and maximum values across the four realizations reveal the sensitivity of the cut-off values to the alternative measure of realizations. Because the cut-off values have the same unit of measure as the forecaster disagreement variable, a median estimate of 0.328 (GDP price index, first horizon) implies that the cross-sectional difference between the 75th percentile forecast for growth and the 25th percentile forecast must be greater than 0.328 percentage point for one forecast to dominate the other.

With one exception, the cut-off values are the point at which the SPF becomes the more conditionally accurate forecast compared with the BMK forecast as forecaster disagreement increases. Only one test, for real PCE consumption expenditures at the third horizon, shows the BMK as the more accurate projection as forecaster disagreement rises above the cut-off value.

Table 2: Summary Statistics of Cut-Off Values for First-Half Sample (1985 Q1–2000 Q4)[1]

	First Horizon		Second		Third		Fourth	
Variable	Median (min | max)	N	Median (min | max)	N	Median (min | max)	N	Median (min | max)	N
GDP Price Index			0.339 (0.284 | 0.408)	4	0.388 (0.374 | 0.402)	4	0.366 (0.341 | 0.380)	4
Nominal GDP							-0.788 (-0.948 | 0.658)	3
Real Federal Gov. C&I	3.309 (3.259 | 3.434)	3	1.821 (1.694 | 1.972)	3	1.833 (1.833 | 1.833)	1	1.598 (1.598 | 1.598)	1
Real Nonres. Fixed Investment			2.018 (1.993 | 2.034)	3	1.962 (1.813 | 2.112)	2		
Real PCE	0.597 (0.596 | 0.598)	3			0.643 (0.643 | 0.643)	1		

1 Table only displays tests where the GW test and GW regression are statistically meaningful.

An insightful finding is that my median estimates show a good deal of variation from one variable to the next and from one forecast horizon to the next, but the variation (measured by the difference between the highest and lowest values) for a particular variable and horizon is small. This result suggests that one should not apply the cut-off estimates for one variable and forecast horizon to all variables and all horizons. However,

8 Occasionally, the cut-off values can be estimated as negative values even though the dispersion measure itself is nonnegative. The interpretation of such estimates is that one forecast is always relatively more accurate than the other, regardless of the value of the dispersion measure. All data for the cut-off values, GW tests, and GW regressions for all variables, forecast horizons, historical realizations, and samples are available upon request.

the cut-off estimates for a particular variable and forecast horizon are not sensitive to the measure of the realization.

As a check on the robustness of my results over the entire sample period, I also computed the summary statistics for estimated cut-off values over the first half of my sample period, displayed in Table 2, and the second half, displayed in Table 3. The GW test and GW regression found far fewer statistically meaningful results over these shorter sample periods when compared to the full sample period. That means the subsample results are sparse, and the summary statistics rely on fewer observations. The estimated cut-off values over the first half of the sample period range from a median of -0.788 for nominal GDP at the fourth horizon to a median of 3.309 for real federal government C&I at the first horizon. Nominal GDP at the fourth horizon has the largest difference between the minimum and maximum cut-off values across all sample periods.

The results for the second half of my sample period, shown in Table 3, indicate median cut-off values that range from 0.180 percentage point for real federal government C&I at the second horizon, to 4.698 for real residential fixed investment at the fourth horizon. Real federal government C&I, which in the full and first-half samples had higher median cut-off values at earlier horizons and lower values at later horizons, shows a reverse of this pattern in the second-half sample.

Table 3: Summary Statistics of Cut-Off Values for Second-Half Sample (2001 Q1–2017 Q4)[1]

	First Horizon		Second		Third		Fourth	
Variable	Median (min | max)	N	Median (min | max)	N	Median (min | max)	N	Median (min | max)	N
Nominal GDP	0.407 (0.407 | 0.407)	1						
Real Federal Gov. C&I	0.618 (0.530 | 0.826)	4	0.180 (0.109 | 0.404)	4	0.630 (0.548 | 0.810)	4	0.986 (0.935 | 1.068)	3
Real Nonres. Fixed Investment	1.821 (1.772 | 2.134)	3	1.999 (1.999 | 1.999)	1	2.027 (1.968 | 2.108)	3		
Real PCE	0.372 (0.372 | 0.372)	1						
Real Res. Fixed Investment	4.308 (4.080 | 4.533)	4					4.698 (4.698 | 4.698)	1
Real State + Local Gov. C&I	0.549 (0.466 | 0.573)	3	0.613 (0.432 | 0.646)	4	0.567 (0.541 | 0.612)	4	0.649 (0.591 | 0.659)	4

1 Table only displays tests where the GW test and GW regression are statistically meaningful.

Conclusion

In this research brief I have continued the Philadelphia Fed's longstanding tradition of examining the forecast accuracy of the SPF by testing the survey's accuracy against a benchmark model, conditioning on forecaster disagreement. I have computed the GW test and GW regression across

three samples of the data and found that the sample choice matters, with some variables having more conditionally statistically meaningful differences in forecast accuracy in different samples. Additional evidence emphasized the importance of using all four realizations of the data rather than relying on just a single realization for forecast accuracy tests. My most general finding is that the SPF becomes more accurate than the benchmark model as forecaster disagreement rises in many of the cases that I examine.

References

Croushore, Dean, and Tom Stark. "Fifty Years of the Survey of Professional Forecasters," Federal Reserve Bank of Philadelphia *Economic Insights* (Fourth Quarter 2019), pp. 1–11.

Diebold, Francis X., and Roberto S. Mariano. "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics,* 13:3 (1995), pp. 253–263, https://doi.org/10.1080/07350015.1995.10524599.

Doelp, Patrick, and Fatima Mboup. "Battle of the Forecasts: Growth of the Median or Median of Growth as the SPF Consensus," Federal Reserve Bank of Philadelphia *Research Brief* (2021), https://doi.org/10.21799/frbp.rb.2020.jan.

Giacomini, Raffaella, and Halbert White. "Tests of Conditional Predictive Ability," *Econometrica,* 74:6 (2006), pp. 1545–1578. https://doi.org/10.1111/j.1468-0262.2006.00718.x.

Mboup, Fatima, and Ardy Wurtzel. "Battle of the Forecasts: Mean vs. Median as the Survey of Professional Forecasters' Consensus." Federal Reserve Bank of Philadelphia *Research Brief* (2021), https://doi.org/10.21799/frbp.rb.2018.dec.

Sill, Keith. "Forecast Disagreement in the Survey of Professional Forecasters," Federal Reserve Bank of Philadelphia *Business Review* (Second Quarter 2014), pp. 15–24.

Stark, Tom. "Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters," Federal Reserve Bank of Philadelphia *Research Rap* (2010).