ICMA Centre

Henley Business School
UNIVERSITY OF READING

Discussion Paper

Are Macro-Forecasters Essentially the Same? An Analysis of Disagreement, Accuracy and Efficiency

October 2016

Michael P Clements
ICMA Centre, Henley Business School, University of Reading

Discussion Paper Number:
ICM-2016-08

The aim of this discussion paper series is to disseminate new research of academic distinction. Papers are preliminary drafts, circulated to stimulate discussion and critical comment. Henley Business School is triple accredited and home to over 100 academic faculty, who undertake research in a wide range of fields from ethics and finance to international business and marketing.

admin@icmacentre.ac.uk

www.icmacentre.ac.uk

© Clements, October 2016

Are Macro-Forecasters Essentially The Same? An Analysis of Disagreement, Accuracy and Efficiency.

Michael P. Clements
ICMA Centre,
Henley Business School,
University of Reading,
Reading RG6 6BA
m.p.clements@reading.ac.uk.

October 13, 2016

Abstract

We investigate whether there are systematic differences between forecasters in terms of their levels of disagreement and the accuracy of their forecasts, and whether these differences are related to whether or not a forecaster efficiently uses their available information. We find that forecasters are not interchangeable. At any point in time, the level of disagreement between forecasters is more likely to be due to a given set of forecasters, as opposed to any randomly-selected set of forecasters. In terms of forecast accuracy, we also find persistence, in that forecasters who are more (less) accurate in one period tend to be more (less) accurate in a subsequent period. Finally, we reject efficiency for around half of all forecasters at short horizons (depending on the variable in question), and find that efficient forecasters tend to be more accurate and less contrarian. Our results do not support the notion that contrarian forecasts stand apart by virtue of having superior information - knowing something that others do not.

Keywords: Expectations formation, Disagreement, Accuracy, Forecast Efficiency. C53, E37.

1 Introduction

Recent years have seen much innovative work on expectations formation, and in particular on explaining why forecasters disagree. The full-information rational expectations (FIRE) model in which all agents know the true structure of the economy and have access to the same information set leaves no room for differences in expectations across agents. Informational rigidities (IR) have been used to explain disagreement, and in a way in which the basic notion that forecasters form their expectations rationally, given the information constraints they face, remains intact. The two key models of informational rigidities are sticky information,[1] and noisy information.[2] Sticky information assumes that in each period, each agent updates their information (relative to the previous period) with a given probability. When they do update, they acquire full information and act as FIRE agents. The noisy information model assumes agents base their forecasts on the latest information, but only ever observe noisy signals about economic fundamentals. However, they filter the signal optimally, and conform to the rational expectations hypothesis conditional on their information set.

The sticky information model as usually used assumes all agents have the same probability of updating their forecast at each point in time. The noisy information model as implemented assumes the noise-variance contaminating agents' signals is equal across agents. Coupled with the assumption that agents share a common model of the economy,[3] IR forecasters are effectively *identical* or *interchangeable*: this period (t) Forecaster A may happen to receive a more accurate signal than Forecaster B, or may update her forecast in period t whereas Forecaster B does not, resulting in A's forecast being more accurate than B's, but next period ($t+1$) B is just as likely to produce the superior forecast as A. We will consider whether individual forecasters are essentially the same in terms of their ability to produce accurate forecasts. We will also consider whether there are persistent differences between forecasters in terms of their degree of non-conformity with the consensus (or disagreement). Lastly, we consider whether forecasters who satisfy forecast efficiency are more or less likely to report accurate forecasts, or to adopt a contrarian stance. For example, if Forecaster A tends to produce systematically more accurate forecasts than B, is it the case that a) Forecaster A is systematically closer to (further from) the consensus than B? and/or b) that Forecaster A tends to produce efficient forecasts whereas Forecaster B does not? The latter bears on whether a more accurate forecaster has access to superior information, or uses the (public) information more efficiently.

Our main question is whether there is empirical support for the interchangeability of forecasters, in terms of the characteristics of disagreement and accuracy, and the relationship between these two characteristics and efficiency. This has implications for the standard IR literature, as well as the wider question of how expectations are formed. If forecasters are essentially interchangeable, then the IR literature appears well micro-founded, but otherwise there may be a case for generalizing these models. For example, Giacomini, Skreta and Turen (2015) consider relaxing the common model assumption, as a way of improving the ability of their model to explain the data. We take a step back and consider

[1] See *inter alia* Mankiw and Reis (2002) and Mankiw, Reis and Wolfers (2003), and Coibion and Gorodnichenko (2012, 2015).

[2] See Woodford (2002), Sims (2003) and Coibion and Gorodnichenko (2012, 2015), *inter alia*.

[3] The literature aims to match certain key moments of the data by making as few departures from FIRE as possible (as made explicit in Andrade and Le Bihan (2013) and Andrade, Crump, Eusepi and Moench (2014), *inter alia*). The moments of the data of interest are usually taken to be (cross-section) aggregate measures, for example, the responses of mean errors and forecaster dispersion to shocks.

whether the different aspects of interchangeability hold by directly examining the forecast records of actual forecasters. We consider the forecasts made by US Professional Forecasters (the US SPF) over (approximately) a quarter of a century from 1990 to 2013.

At first sight, it might appear that the properties we consider are not distinct. However this is incorrect. For example, forecast efficiency is distinct from forecast accuracy. A forecast can be efficient, but inaccurate, or accurate but not efficient. Efficiency (in the sense of Mincer and Zarnowitz (1969), and as used in this paper) means that the forecaster efficiently uses the information available at the forecast origin. Depending on the individual's information set, an efficient forecast may be more or less accurate than a comparator. Conversely, a forecast may be accurate without exploiting all the information available to the forecaster. Secondly, consider the relationship between non-conformist behaviour and accuracy (or efficiency). A forecaster who 'stands out from the crowd' may have superior information and produce more accurate forecasts than average (and these forecasts may or may not be efficient), or may not have superior information, and simply produces relatively inaccurate forecasts.

Of course forecasters may not be attempting to produce the most accurate forecasts possible: forecasters may be driven by motives other than minimizing expected squared forecast error. For example, according to Lamont (2002, p. 265), forecasters may set their forecasts to 'optimize profits or wages, credibility, shock value, marketability, political power...' (and see in addition Laster, Bennett and Geoum (1999) and Ottaviani and Sorensen (2006), *inter alia*). These and related explanations of forecaster behaviour implicitly assume that forecasters are heterogeneous. In this paper our focus is on the evidence of forecaster heterogeneity from examining the forecasts of the individual survey respondents.[4]

Finally, there are a number of ways of measuring and testing for these properties. For instance, the recent literature stresses the multivariate nature of forecasting, and argues for multivariate loss functions for assessing accuracy, and multivariate measures of disagreement to determine 'how far from the crowd' in terms of a vector of forecasts . The measures we adopt reflect a number of these developments, and we are careful to check the robustness of our results to different approaches.

We consider US professional forecasters expectations of consumption, investment and output, because the growth rates of these variables tend to move together, and allow us to investigate the importance of adopting multivariate approaches to assessing forecast disagreement and accuracy (relative to univariate approaches).[5] There are few studies looking at multivariate disagreement, and the studies there are arguably consider sets of variables about which there is less likely to be a consensus about the relationships between the variables (discussed further below).

Our paper is related to a large literature on disagreement, [6] but the number of papers using multi-

[4] Other motivations, such as seeking 'shock value' may tend to hide innate differences between forecasters in terms of their ability to produce accurate or efficient forecasts, but we do not directly consider their effects here. Nor do we directly consider the reasons why forecasters might be different from one another. For example, in addition to those already mentioned related to information rigidities, forecasters may disagree because they: receive a public signal but interpret the signal differently (see, e.g., Kandel and Zilberfarb (1999) and Manzan (2011)); have heterogeneous beliefs about long-run outcomes (see Patton and Timmermann (2010)); or have heterogeneous degrees of loss asymmetry (Capistrán and Timmermann (2009)), *inter alia*.

[5] In addition, a number of studies have considered whether there are constant long-run or equilibrium relationships between the log levels of these variables. King, Plosser, Stock and Watson (1991) found support for the 'great ratios' of Kosobud and Klein (1961) on data up to 1990, consistent with balanced growth paths (of the Solow-Ramsey model), whereas more recently two-sector models (such as, e.g., Whelan (2003)) predict that the key NIPA aggregates grow at constant but different rates. Clements (2016) considers whether imposing theory-based, steady state restrictions on long-horizon forecasts improves forecast accuracy, with mixed results.

[6] See, for example, Zarnowitz and Lambros (1987), Bomberger (1996), Rich and Butler (1998), Capistrán and Timmer-

variate approaches is much smaller. Banternghansa and McCracken (2009) was one of the first papers to consider multivariate approaches to forecast disagreement, and Dovern (2014) uses multivariate disagreement to assess whether forecasters disagree because they have different views about the outlook for the economy e.g., whether an expansion is likely to continue or give way to a period of slower growth or even a contraction, or whether disagreement occurs because forecasters have different views about how the economy operates (even though they might be of a similar opinion regarding the prospects for growth, for example). In the context of forecasting inflation, GDP growth and the unemployment rate, Dovern (2014) finds that the disagreement-based evidence for agents sharing a common model of the economy is weak. However, this may simply be because some of the actual data correlations between these variables are small. If the actual data series were generated independently of each other forecasters would not be expected to adopt a common model of their joint determination. On the other hand, there are good reasons to expect the variables we consider to be relatively highly correlated, so it is of interest to assess whether these correlations are present in the forecast data. Moreover, the balanced growth paths (of the Solow-Ramsey model) or two-sector models (such as, e.g., Whelan (2003)) all predict that the key National Income and Product Accounts aggregates grow at constant (possibly different) rates in the long run, which imply relationships between the short-run forecasts of these variables.

The plan of the remainder of the paper is as follows. In section 2 the forecast data used throughout the paper are discussed. Section 3 presents the multivariate measures of disagreement, and section 3.1 applies these measures to analyze disagreement for the most prolific forecasters. Section 4 describes the assessment of individual-level forecast accuracy, and section 4.1 presents the results of this assessment. Section 5 describes the tests of forecast optimality, and section 5.1 the results. Section 6 presents a simple model which brings out the relationship between forecast efficiency and forecast accuracy. Section 7 directly addresses the question of whether more accurate forecasters are more contrarian, and section 8 whether efficient forecasters are more or less accurate, or more or less contrarian. Section 9 offers some concluding remarks.

2 Forecast Data: SPF Respondents' Forecasts

We use the US Survey of Professional Forecasters (SPF). The SPF is a quarterly survey of macroeconomic forecasters of the US economy that began in 1968, administered by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER). Since June 1990 it has been run by the Philadelphia Fed, renamed as the Survey of Professional Forecasters (SPF): see Croushore (1993). The SPF is made freely available by the Philadelphia Fed, allowing results to be readily reproduced and checked by other researchers. Its constant scrutiny is likely to minimize the impact of respondent reporting errors. An academic bibliography of the large number of published papers that use SPF data is available at: http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/academic-bibliography.cfm.

We use the SPF multi-horizon forecasts of GDP, consumption and investment from 1990:4 onwards, i.e., from when it was administered by the Philadelphia Fed. It is tempting to use the earlier survey data, but the SPF documentation warns of its suspicion that the forecast identifiers may not have been

mann (2009), Lahiri and Sheng (2008), Rich and Tracy (2010) and Patton and Timmermann (2010).

uniquely assigned over the earlier period - newcomers may have been given the identifiers once associated with participants who have left the survey. Given our focus on indvidual behaviour, it seems preferable to forego the additional survey data.

Forecasts are made of the current quarter (i.e., the quarter in which the survey takes place), and of the quarterly values of the variables in each of the next four quarters, so that the longest-horizon quarterly forecast is of the same quarter of the year in the following year. Forecasts are also provided of the current (survey-quarter) calendar-year levels of the variables, and of the levels in the following year. Hence for surveys made in the first quarter of a year, the next year forecast approximately corresponds to a 2-year or 8-quarter forecast, whereas for fourth quarter surveys these forecasts have a horizon of 5-quarters. Hence from the first-quarter surveys we obtain an annual series of 2-year ahead calendar-year forecasts (as well as the annual series of 1-year ahead calendar-year forecasts from the forecasts of the current year).

We use the 92 surveys from 1990:4 to 2013:3 inclusive. Table 1 provides details concerning the actual and forecast data.

The switch from a 'fixed-base-year' to chain-weighted estimates of real GDP and its components in the 1990's may potentially have an impact on our analysis, as with chain-weighting it is no longer true that the GDP components sum to GDP, or to intermediate sub-aggregates. The investment series is the sum of private non-residential investment, and residential investment, although strictly-speaking these components are not summable. However, it seems likely that any resulting distortions will be of secondary importance.

In the course of this research, a small number of aberrant observations were identified, and these observations were replaced by missing values. Appendix 1 details the small number of changes which were made to the published SPF data.

3 Multivariate Measures of Disagreement

Banternghansa and McCracken (2009) argue for a multivariate approach to the analysis of forecaster disagreement. Survey respondents are typically asked to report forecasts for a number of variables. Banternghansa and McCracken (2009) argue that one might then consider multivariate measures of disagreement, which consider the distances between the vectors of forecasts, rather than analyzing disagreement about individual variables in isolation of each other. Since the vector of forecasts reflects the forecaster's beliefs about the inter-dependencies that exist between the variables, it is reasonable to take the correlations across variables into account in determining the extent to which forecasters disagree. They present an illustration (see their Figure 1): if two variables are positively correlated, then this would be expected to be reflected in forecasts of these two variables, and an individual who records forecasts of the two variables of different sign might could reasonably be said to disagree to a greater extent than a forecaster who produces forecast of the same sign, even if the (Euclidean) distance of the two pairs of forecasts from the consensus is the same.

To capture this idea, they first define the cross-sectional forecast covariance matrix as:

$$S_{t|t-h} = N_{t,h}^{-1} \sum_{i=1}^{N_{t,h}} \left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right) \left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right)' \tag{1}$$

where $y_{i,t|t-h}$ is the vector of forecasts made by i at time $t-h$ for a target y_t, $N_{t,h}$ is the number of forecasters of y_t at time $t-h$, and $\overline{y}_{t|t-h} = N_{t,h}^{-1} \sum_{i=1}^{N_{t,h}} y_{i,t|t-h}$. Then they define their multivariate disagreement measure for individual i forecasting the vector y_t at forecast origin $t-h$ as the Mahalanobis distance:

$$D_{i,t|t-h} = \sqrt{\left(y_{i,t|t-h} - \overline{y}_{t|t-h}\right)' S_{t|t-h}^{-1} \left(y_{i,t|t-h} - \overline{y}_{t|t-h}\right)}. \quad (2)$$

To illustrate, suppose y consists of just two variables, and $y_{i,t|t-h} - \overline{y}_{t|t-h} = (1,1)'$, so that a respondent's forecasts of both variables differ from the consensus forecasts by a positive amount (of 1 unit). Then the Euclidean measure of disagreement (by setting $S = I_2$) is $\sqrt{2}$. Suppose the diagonal elements of S are unity and the off-diagonal element is ρ. If $\rho = 0.9$, so the cross-sectional covariance between the other respondents' forecasts of the two variables (equivalently, forecast errors) is positive, then $D = \sqrt{2/1.9}$, which is less than $\sqrt{2}$, which is in turn less than $D = \sqrt{20}$ when $\rho = -0.9$.

Dovern (2014) interprets the elements on the main diagonal of $S_{t|t-h}$ as measuring the disagreement about the outlook for the specific variables, whereas the covariances indicate (dis)agreement about how the economy operates. In the context of forecasting inflation, GDP growth and the unemployment rate, Dovern (2014) finds the unconditional cross-section correlations (full-sample averages) between the three variables are relatively low - the year ahead inflation and output growth forecasts exhibit a correlation of only 0.16, and the growth and unemployment rate forecasts a correlation of -0.25. [7] Off-diagonal terms close to zero indicate that forecasters who predict higher than average inflation, say, are just as likely to predict lower than average growth as higher than average growth: there is little agreement in terms of the forecasters' beliefs about how inflation and output growth are generated in this example.

From our perspective, the use of $S_{t|t-h}$ in (2) serves to adjust $y_{i,t|t-h} - \overline{y}_{t|t-h}$ for some periods being inherently more difficult to forecast than others. These differences are more likely to be large for a respondent who just happened to be active in highly uncertain times compared to a forecaster responding in tranquil times. When $S_{t|t-h}$ is diagonal, it is clear from (2) that the multivariate measure sums (across variables) the squared deviations of the forecasts from the consensus, after first dividing the squared deviations by the cross-sectional variances of the variables at that t. Hence included in (2) is an automatic adjustment for individuals forecasting during different economic conditions, as well as including an offset for agreement about how the economy operates. The former adjustment is potentially important because few if any forecasters are ever present over the whole period, and many forecasters respond to around a half of the surveys between 1990:4 and 2013:3.

3.1 Multivariate Disagreement and the Most Prolific Individuals

Forecasters are obviously heterogeneous in the sense that they do not report identical forecasts at each point in time. The more interesting question is whether differences in forecasters are systematic, in the sense that some respondents' forecasts tend to systematically differ more or less from the consensus than those of others. The alternative would be that overall disagreement at any point in time is as likely to be due to any one forecaster disagreeing with the consensus as any other forecaster.

Table 2 reports the evidence for forecaster heterogeneity based on the multivariate disagreement

[7] The correlation between output and unemployment is known as Okun's Law (see, e.g. Ball, Jalles and Loungani (2015) for a cross-country analysis from a forecasting perspective), but there is perhaps less reason to expect a strong correlation between output growth and inflation.

measure (equation (2)). The table records results for the 15 most prolific forecasters in response to the 92 surveys between 1990:4 and 2013:3. The forecasters are ordered numerically by forecaster id., and the average disagreement for each individual recorded in the second and fourth columns is the average of eqn. (2) across all surveys to which the individual responded, for $h = 0$ and $h = 4$, respectively.

The top half of the table is based on the multivariate disagreement measure which takes into account the correlations between variables, whereas the bottom half assumes $S_{t|t-h}$ is diagonal, and so simply sums the scaled disagreement for each variable.[8]

For the $h = 0$ forecasts we report a formal test of whether the population means of the $D_{i,h}$ differ across individuals, i.e., of the null that $H_0 : \mu_{i,h} = \mu_{m,h}$ versus $H_1 : \mu_{i,h} \neq \mu_{m,h}$ for individuals i, where m is the individual with the average level of disagreement at $h = 0$, and where $\mu_{i,h}$ denotes a population mean. The $\left\{D_{i,t|t-h}\right\}$ are regarded as realizations, and we report a standard t-test for the equality of two population means allowing the variances to be unequal. p-values are reported, calculated such that a large p-value indicates the individual has a significantly larger than average value of disagreement, and a value close to zero indicates a smaller than average value of disagreement. Formally we find that for around two thirds of forecasters the p-values exceed 0.90 or are less than 0.10 for both weighting schemes. Hence there is evidence of significant differences between forecasters in terms of their disagreement.

A method of assessing the persistence in individual forecasting behaviour, which doesn't require pairwise comparisons of each individual to the average forecaster, is to compare the ranks of forecasters based on their average levels of multivariate disagreement in the first and second halves of the sample. We split the sample 1990:4 to 2013:3 in half, and refer to the first (or earlier) and second (or later) samples.[9] For $h = 0$, the full sample rank and two sub-sample ranks are given in columns 5 to 7, and in columns 8 to 10 for the $h = 4$ forecasts. Note that the two respondents with the lowest and highest levels of multivariate disagreement in the first sample (for $h = 0$), i.e., those ranked 1 and 15, are also 1 and 15 in the later period. However, there are large switches too.

As a formal test of whether the rankings are the same over the two sub-samples, we report tests based on Spearman's rank correlation coefficient. In addition to comparing forecaster behaviour over time, in terms of disagreement, we also address the constancy of forecaster behaviour across horizon, and the effects on these comparisons of adopting a true multivariate measure as opposed to summing disagreement for the individual variables. Table 3 reports the p-values of rank correlation tests of the null hypotheses that there is no relationship between forecaster disagreement: *i*) across time - between the earlier and later periods - for a given h ($h = 0, 4$): Panel A; *ii*) between short ($h = 0$) and long-horizon forecasts ($h = 4$), across all surveys and in each of the two sub-periods: Panel B. We carry out *i*) and *ii*) for the multivariate disagreement measure, and for the sum of the individual variable measures.

The first panel of table 3 records the results for the 15 forecasters of table 2. The bottom panel records results for the 44 most prolific forecasters, as a robustness check.

Consider the top panel for the 15 most prolific forecasters. For each test, the table records the small-sample test statistic and the p-value for the large-sample approximation to the test statistic,

[8] $S_{t|t-h}$ is calculated using the forecasts of all the respondents (who responded to 12 or more surveys).

[9] For any individual, there may be different numbers of forecasts in the two sample - an individual may have been ever-present in the later-period surveys, say, but not in the earlier period. When we consider the top 44 forecasters this becomes problematic, since some individuals make too few forecasts in one of the two samples to reliably estimate disagreement. Such individuals are not included in the tests we report comparing the behaviour of individual forecasters across the two samples.

which is normally distributed under the null.[10] We reject the null of no relationship in the rankings of disagreement between the earlier and later sample periods for both forecast horizons, and for both disagreement measures (referred to as 'S' and 'Diag. S' in the table). In terms of short and long-horizon disagreement, we reject the null of no relationship in the rankings across forecasters for the whole forecast sample and both sub-samples, whether we use the multivariate or 'univariate' measure (i.e., the sum of the disagreement measures for each variable).

When we consider the top 44 forecasters, the results are for the most part as for the top 15: extending the number of forecasts does not change the results. The rejections of the null are more emphatic - at the 1% level - except for the test across sub-samples for the $h = 4$ forecasts using the diag. S matrix.

We conclude that there is considerable evidence that forecasters are not interchangeable in terms of their degrees of conformity with the consensus. Moreover, the results are generally not sensitive to whether the disagreement measure is adjusted for the degree of agreement about how the economy operates.

4 Forecast Accuracy

We ask whether the more (less) accurate forecasters over a given period remain the more (less) accurate over a subsequent period. The forecast accuracy measures we use are the trace and the determinant of the Mean-Squared Forecast-Error Matrices (MSFEMs) for $h = 0$ and $h = 4$ forecasts. The determinant is a multivariate measure, whereas the trace simply sums the individual-variable MSFEs. Clements and Hendry (1993) propose the determinant as an invariant measure of forecast accuracy for 1-step forecasts: it is invariant to forecasting linear transformations of the vector of variables. For $h = 4$ an invariant measure would be the Generalized Forecast Error Second Moment Matrix (GFESM), as discussed by Clements and Hendry (1993), although we have relatively small samples of forecasts at our disposal to calculate such a measure (but see Hendry and Martinez (2016)). Komunjer and Owyang (2012) propose a multivariate loss function which does not pre-suppose independence between the different variables' forecast errors in the loss function. In their approach, the multivariate loss function reduces to the sum of the individual variable losses (cf. the trace) when loss is symmetric. We leave for future research the question of whether individuals' preferences over the three variables are non-symmetric, and whether this qualitatively affects our findings.

We adjust for individuals forecasting during different economic conditions by controlling for differences over time in the average accuracy of all forecasters, following D'Agostino, McQuinn and Whelan (2012) and Clements (2014).[11] Not controlling for the degree of difficulty in forecasting at time t might distort the inter-personal comparisons of forecast accuracy.

Letting $e_{i,t+h|t}$ denote the forecast error made by individual i in response to forecast survey t, at

[10] The statistics and small-sample critical values are recorded in the notes to the table.

[11] An extreme example illustrating why such adjustments may be necessary is provided by investment around the time of the recent Crisis. Investment fell by about 12% in 2009:1 relative to 2008:4 (not annualized). The magnitude of the fall was unforseen, and those who happened to respond to the 2008:1 survey registered much larger 4-step ahead forecast errors than those made in response to any other survey.

each t we calculate (separately for each of the three variables) normalized forecast errors as:

$$\widetilde{e}_{i,t+h|t} = \frac{e_{i,t+h|t}}{\sqrt{\frac{1}{N_t}\sum_{j=1}^{N_t} e^2_{j,t+h|t}}}$$

where N_t is the number of respondents to survey t,[12] so that the denominator is the cross-section RMSE. The trace MSFE for respondent i is then simply:

$$\sum_{k=1}^{3}\left(\frac{1}{n_i}\sum_{t\in N_i}\widetilde{e}^2_{k,i,t+h|t}\right)$$

where the k subscript denotes the variable ($k = (1,2,3)$), N_i denotes the set of surveys to which i responded, and n_i is the number of elements in that set.

After normalising the forecast errors, D'Agostino *et al.* (2012) consider whether some forecasters are innately more accurate than others, by testing a null hypothesis that all forecasters are equally accurate. Their test statistic compares various percentiles of the empirical distribution of forecasters' accuracy against a bootstrapped distribution derived under the assumption of equal forecaster ability. They consider a single variable, whereas our interest is in a vector of variables. We leave for future research the possibility of extending their approach to a multivariate setting. Sinclair, Stekler and Carnow (2015) do present a multivariate analysis. They suggest evaluating a vector of forecasts of a number of variables by considering the Mahalanobis distance between the vector of forecasts and outcomes. They compare the mean vectors - the mean vector of forecasts and the mean vector of actual values - and consider whether the forecasts and outcomes have the same population means. Their approach does not directly evaluate the accuracy of the forecast errors and hence is not directly relevant to our concerns.

4.1 Forecast Accuracy Results

Table 4 ranks the 15 most prolific forecasters by the two forecast accuracy measures for the whole sample, and for each of the two sub-samples, using the normalized forecast errors. The sub-table reports Spearman rank tests of the null that the rankings across the two sub-samples are unrelated. At the 10% level we do not reject the null for the short-horizon forecasts, but we do reject at the 5% level for $h = 4$ using the determinant. If instead we consider the top 44 forecasters, the null is rejected for $h = 0$ at the 1% level, but there is no evidence against the null for $h = 4$, unless the determinant measure is used.

The importance of the normalization is evident from table 5, where the 'raw' forecast errors are used to calculate the forecast accuracy measures. Without the normalization, the null of no relationship between the two sub-periods is clearly rejected for the top 15 forecasters for $h = 0$. On the raw errors, for example, id 431 is ranked either 10th or 6th (on the trace or det. measures). On the normalized errors, id. 431 jumps to 3rd or 1st.

In summary, if we widen the net to include the top 44 forecasters, there is clear evidence that the rankings of forecasters across the two periods are systematic (for $h = 0$). The results are shown to depend on whether or not the forecast errors are normalized, indicating the importance of controlling

[12] N_t includes all the forecasters (who made more than 12 returns), not just the most prolific.

for differences over time in the difficulty of forecasting, given that each individual responds to a subset of the surveys.

5 Tests of Forecast Optimality

Tests of forecast optimality typically consider whether the forecasts efficiently make use of all the available information available to the forecaster at the time the forecast was made. There are various aspects to this, and ways of making operational the notion of 'all the available information'.

The Mincer and Zarnowitz (1969) (MZ) regression tests forecast optimality at a given horizon. The regression is:

$$y_t = \delta_0 + \delta y_{t|t-h} + u_t \tag{3}$$

where the observations range over t for a given h, and the null of optimality is that $\delta_0 = 0$ and $\delta = 1$. Consider the covariance between the forecast error and the forecast:

$$Cov\left(y_t - y_{t|t-h}, y_{t|t-h}\right) = Cov\left(\left(\delta - 1\right) y_{t|t-h} + u_t, y_{t|t-h}\right).$$

Unless $\delta = 1$, the forecast and forecast error will be systematically related, and this correlation could be used to generate a superior forecast. For $\delta = 1$, the forecast error will be biased unless $\delta_0 = 0$. For multi-step forecasts, HAC standard errors are used to account for the overlapping forecasts phenomenon.

The MZ test could be viewed as a minimal requirement, in the sense that more stringent tests might test the orthogonality of the forecast error and specific variables in the agent's information set at $t - h$. Rather than going down that route we follow Patton and Timmermann (2012), who propose the univariate optimal revision regression (henceforth ORR), which is applicable when fixed-event forecasts (see, e.g., Nordhaus (1987) and Clements (1995)) are available, as here. This test can be motivated by writing a short horizon forecast (e.g., $h_1 = 1$) as:

$$y_{t|t-h_1} \equiv y_{t|t-h_H} + d_{t|h_1,h_2} + \ldots d_{t|h_{H-1},h_H} \tag{4}$$

where $h_1 < h_2 < \ldots < h_H$, with h_H the longest horizon forecast of the target y_t, and $d_{t|h_j,h_{j+1}} = y_{t|t-h_j} - y_{t|t-h_{j+1}}$. Then rather than regressing y_t on $y_{t|t-h_1}$, say, as in (3), the ORR test substitutes for $y_{t|t-h_1} = y_{t|t-h_H} + \sum_{i=1}^{H-1} d_{t|h_i,h_{i+1}}$ in (3), and allows a free coefficient on each of the components of $y_{t|t-h_1}$. We then estimate:

$$y_t = \delta_0 + \delta_H y_{t|t-h_H} + \sum_{i=1}^{H-1} \delta_i d_{t|h_i,h_{i+1}} + u_t, \tag{5}$$

and the null hypothesis is that $\delta_0 = 0$ and $\delta_1 = \delta_2 = \ldots = \delta_H = 1$. Under the null, the error for the short-horizon forecast $y_{t|t-h_1}$ is uncorrelated with all forecasts of the target y_t made at earlier times (and hence on smaller information sets). Equation (5) becomes $y_t = y_{t|t-h_1} + u_t$. Hence the ORR test has power to reject the null against the alternative that the short-horizon forecast error is systematically related to revisions in earlier forecasts of the target value.

Patton and Timmermann (2012) also show that a variant can be obtained by replacing the actual value of y_t by a short-horizon forecast, say, $y_{t|t-h_1}$, to give:

$$y_{t|t-h_1} = \delta_0 + \delta y_{t|t-h_2} + u_t \tag{6}$$

where $h_2 > h_1$, and e.g., :

$$y_{t|t-h_1} = \delta_0 + \delta_H y_{t|t-h_H} + \sum_{i=2}^{h_H-1} \delta_i d_{t|h_i,h_{i+1}} + u_t \tag{7}$$

when $h_H > h_{H-1} > \ldots > h_1$.[13] This requires that the short-horizon forecast is a conditionally unbiased proxy for the actual value, and the interpretation of, say, (6) is that it tests the rationality of both $y_{t|t-h_1}$ and $y_{t|t-h_2}$. A practical advantage is to obviate the need to select the vintage(s) of data to be used as actual values. US national accounts data are subject to various rounds of revisions: see, e.g., the review articles by Croushore (2011a, 2011b) as well as Landefeld, Seskin and Fraumeni (2008) and Fixler, Greenaway-McGrevy and Grimm (2014). Researchers sometimes use a vintage released soon after the reference quarter, rather than the latest-available vintage at the time of the investigation, and this is the approach we use when calculating forecast accuracy. This is because the 'fully-revised' data will typically include benchmark revisions, rebasings, and other methodological changes to the way the data are collected and measured, which would not have been foreseen when the forecast was made. In section 4 we used the data vintage available two quarters after the reference quarter. [14] This issue can be side-stepped when testing for forecast efficiency by using short-horizon forecasts in place of actuals. A further adavantage is that using forecasts may make the calculation of autocorrelation-consistent standard errors (AC-SEs) unnecessary. The calculation of AC-SEs is complicated when there are missing forecast observations, as in surveys such as the SPF: individuals do not file a response to every survey.[15]

We show in an Appendix that using adjacent horizon forecasts in the tests of forecast optimality serves to circumvent the need for autocorrelation corrections. That is, in the MZ regression given by:

$$y_{t|t-h_1} = \delta + \delta_1 y_{t|t-h_2} + u_t \tag{8}$$

where $h_2 = h_1 + 1$ (i.e., that the forecasts are adjacent) the error term in the regression will be serially uncorrelated for optimal forecasts.

However, the substitution of the short-horizon forecasts for the actual values requires the optimality of the former, otherwise tests of revisions may have no power to detect mis-specification, a situation described by Nordhaus (1987).[16] Some departures from full-information rational expectations forecasts which are not detectable using short-horizon forecasts may be readily detectable using actual values.

[13] Tests based on (7) are closely related to the weak efficiency tests of Nordhaus (1987): forecast revisions should be unpredictable from earlier revisions.

[14] The Real Time Data Set for Macroeconomists (RTDSM) maintained by the Federal Reserve Bank of Philadelphia (see Croushore and Stark (2001)) has greatly facilitated the use of real-time data in macro analysis and forecasting research.

[15] As in much of the literature, we assume that data are missing 'at random' so that the sample is representative of the population. An exception is López-Pérez (2015) who considers whether the decision to contribute is related to perceived uncertainty about the outlook.

[16] Nordhaus (1987, p. 673) described the possibility of 'A baboon could generate a series of weakly efficient forecasts by simply wiring himself to a random-number generator, but such a series of forecasts would be completely useless.'

Suppose for example that:

$$y_{t|t-h} = y_{t|t-h-1} + u_{t|t-h} \tag{9}$$

where $u_{t|t-h}$ is orthogonal to y_t and $y_{t|t-h-1}$. Then consider an MZ regression of $y_{t|t-h_1}$ on a constant and $y_{t|t-h_2}$, where $h_1 < h_2$, as in (6). Then the population values of the regression parameters are:

$$\delta = \frac{Cov\left(y_{t|t-h_1}, y_{t|t-h_2}\right)}{Var\left(y_{t|t-h_1}\right)} = 1,\ \delta_0 = E\left(y_{t|t-h_1}\right) - \delta E\left(y_{t|t-h_2}\right) = 0$$

since $y_{t|t-h_1} = y_{t|t-h_2} + \sum_{s=t-h_2+1}^{t-h_1} u_{t|s}$ and $Cov\left(\sum_{s=t-h_2+1}^{t-h_1} u_{t|s}, y_{t|t-h_2}\right) = 0$. Forecast revisions which do not add news will nevertheless be detectable using actual values, since:

$$\delta = \frac{Cov\left(y_t, y_{t|t-h_1}\right)}{Var\left(y_{t|t-h_1}\right)} = 0,\ \delta_0 = E\left(y_t\right) - \delta E\left(y_{t|t-h_1}\right) = E\left(y_t\right)$$

since neither $\delta = 1$ nor $\delta_0 = 0$ (unless $E\left(y_t\right)$ happens to equal zero).

In summary: Tests for forecast efficiency based on (6) obviate the need to apply autocorrelation correction in the presence of missing observations, and the need to take a stance on the vintage of data the forecaster is targetting. On the other hand, the use of actual values (as in equation (3)) will not fail to reject when forecasts are generated by, say, (9). Relatedly, we can directly test the efficiency of the $h = 0$ forecasts without placing any requirements on the $h = 1$ forecasts. We calculate forecast efficiency in both of these ways and contrast the results.

5.1 Forecast Efficiency Test Results

We consider forecasts made by the most prolific forecasters in response to the surveys dates between1990:4 and 2013:3 inclusive. We assess whether each forecaster is rational. We consider two ways of implementing the MZ test: using actual values as the dependent variable, and using an $h-1$ forecast on the LHS when the RHS forecast is length h, and thus obviating the need for autocorrelation correction.

The tables of results (tables 6 to 8) for the MZ test regressions of forecast efficiency suggest that the null is generally rejected for a majority of forecasters at the longer horizons, although the patterns of rejections across forecasters vary with the variant of the test. For example, for horizon $h = 1$ for the consumption growth forecasts, the sub-table of table 6 indicates that the results using the actual value as the dependent variable are only in agreement with the results using the shortest-horizon forecast as the dependent variable for just over a half of forecasters. For investment, it is close to 9 in 10 forecasters, and for output growth around a half.

The rejection rates generally increase in h. At $h = 4$, the two tests agree around 90% of the time for consumption and output, and for around two thirds of forecasters for investment.

The results of the ORR tests are reported in tables 9 to 11 and generally suggest similar rates of inefficiency across forecasters. In the following we will consider the relationship between forecast efficiency, accuracy and disagreement using the MZ results for efficiency. It seems unlikely that the results would be very different using the ORR test, and for the shortest horizon forecasts ($h = 0$) the approaches are equivalent.

6 Forecast Accuracy and Forecast Efficiency

As noted in the introduction, the notions of accuracy, efficiency and disagreement are logically distinct. In this section we show formally that inefficient forecasters may be more or less accurate than efficient forecasters.

One way of formalizing the relationship between forecast accuracy and efficiency is the following. Let $\mathcal{I}_{i,t-h}$ denote individual i's information set at time $t-h$, and individual i's optimal forecast h-step ahead forecast of y_t is the conditional expectation $y^*_{i,t|t-h} = E\left(y_t \mid \mathcal{I}_{i,t-h}\right)$. Then compare two forecasters, A and B, assuming that both produce efficient forecasts, i.e., $y_{A,t|t-h} = y^*_{A,t|t-h}$, and $y_{B,t|t-h} = y^*_{B,t|t-h}$. Then Forecaster A's expected squared forecast error will be less than B's if $\mathcal{I}_{B,t-h}$ is a proper subset of $\mathcal{I}_{A,t-h}$, and B will be more accurate if $\mathcal{I}_{A,t-h}$ is a proper subset of $\mathcal{I}_{B,t-h}$. It follows immediately from the definition of the conditional expectation that both forecasters A and B will be efficient: their forecasts and forecast errors will be uncorrelated.

Suppose now that forecasters are not efficient. This can be modelled as in Patton and Timmermann (2012), for example, by supposing the forecast is a linear transformation of the optimal:

$$y_{i,t|t-h} = \gamma_{i,h} + \lambda_{i,h} y^*_{t|t-h} + w_{i,t|t-h}, \quad w_{i,t|t-h} \sim D\left(0, \sigma^2_{i,w,h}\right). \tag{10}$$

Patton and Timmermann (2012) show that there are values of the vector $\left(\gamma_h, \lambda_h, \sigma^2_{w_h}\right)$ other than $(1,0,0)$ which constitute non-detectable mis-specification using standard MZ tests of forecast efficiency. From our perspective, (10) can be used to show an efficient forecaster may be more or less accurate than an inefficient forecast. Suppose $y_{i,t|t-h}$ only differs from $y^*_{i,t|t-h}$ by a random (reporting or measurement) error,[17] corresponding to $\gamma_h = 0$ and $\lambda_h = 1$, but $\sigma^2_{w_h} \neq 0$.

Then the population value of the slope coefficient MZ regression (3) is given by:

$$\delta_i = \frac{Cov\left(y_t, y_{i,t|t-h}\right)}{Var\left(y_{i,t|t-h}\right)} = \frac{Cov\left(y^*_{i,t|t-h_1} + \varepsilon_{i,t|t-h_1}, y^*_{i,t|t-h_1} + w_{i,t|t-h_1}\right)}{Var\left(y^*_{i,t|t-h_1} + w_{i,t|t-h_1}\right)} = \frac{Var\left(y^*_{i,t|t-h_1}\right)}{Var\left(y^*_{i,t|t-h_1}\right) + \sigma^2_{i,w,h}} < 1,$$

where $\varepsilon_{i,t|t-h_1} = y_t - y^*_{i,t|t-h_1}$, and so $E\left(\varepsilon_{i,t|t-h_1} \mid \mathcal{I}_{i,t-h}\right) = 0$, and where $E\left(\varepsilon_{i,t|t-h_1} w_{i,t|t-h_1}\right) = 0$ and $Cov\left(y^*_{i,t|t-h_1}, w_{i,t|t-h_1}\right) = 0$.

Hence, the reported forecasts are not efficient, unless $\sigma^2_{i,w,h} = 0$. The expected squared errors are:

$$E\left(y_t - y_{i,t|t-h}\right)^2 = E\left(\varepsilon_{i,t|t-h_1} - w_{i,t|t-h}\right)^2 = \sigma^2_{i,\varepsilon,h} + \sigma^2_{i,w,h}$$

where $\sigma^2_{i,\varepsilon,h} = E\left(\varepsilon_{i,t|t-h_1}\right)^2$. Hence, Forecaster A will be more accurate, equally accurate, or less accurate than B, depending on whether:

$$\sigma^2_{A,\varepsilon,h} + \sigma^2_{A,w,h} \lessgtr \sigma^2_{B,\varepsilon,h} + \sigma^2_{B,w,h}.$$

Depending on $\mathcal{I}_{A,t-h}$ and $\mathcal{I}_{B,t-h}$, which determine $\sigma^2_{A,\varepsilon,h}$ and $\sigma^2_{B,\varepsilon,h}$, any of the three possibilities (more, less, same accuracy) of Forecasters A and B is possible irrespective of whether either, both or neither of

[17] The assumption forecasters make idiosyncratic errors is sometimes included in studies of forecaster behaviour, e.g., Davies and Lahiri (1995).

the forecasters produces efficient forecasts.

In this simple model disagreement is also increasing in the same parameters as forecast accuracy (namely, $\sigma^2_{i,\varepsilon,h}$ and $\sigma^2_{i,w,h}$) but accuracy and disagreement need not vary positively together over forecasters. A forecaster with a more accurate private signal than others - or who more accurately interprets a public signal (as in Kandel and Zilberfarb (1999)) may be an accurate forecaster who nevertheless stands apart from the crowd.

7 Are More Contrarian Forecasters Less Accurate Forecasters?

Figures 1 and 2 present crossplots of forecast accuracy against disagreement for the top 44 and 15 forecasters, respectively. The plots are drawn for two measures of forecast accuracy, the trace and determinant of the MSFEM for the three variables, based on forecast errors scaled by the estimated difficulty of forecasting. The multivariate disagreement measures are given by equation (2), and as explained in section 3, also make an allowance for some periods being inherently more difficult to forecast than others, as well as including an offset for agreement over how the economy operates (i.e., S is non-diagonal).

Figure 1 shows a clear positive relationship between disagreement and accuracy for the short horizon ($h = 0$) forecasts, such that the more a forecaster stands out from the crowd, the less accurate that forecaster. A negative relationship would have pointed to some forecasters having superior private information which simultaneously sets them apart from the consensus and results in their forecasts being more accurate. At $h = 4$ the relationship between disagreement and accuracy is less clear: the regression line is positive for the trace measure of forecast accuracy, but not for the determinant. Nevertheless, the Spearman rank correlation test results recorded in table 12 indicate the rejection of the null of no relationship between disagreement and accuracy, with p-values of zero to two decimal places. As this test is based only on the ranks, it will not be affected by outlying observations, such as the forecaster with the extreme value of the determinant in Figure 1.

Figure 2 displays the same plots for the top 15 forecasters. The regression approach indicates a positive relationship between accuracy and disagreement for both accuracy measures and forecast horizons. The Spearman test clearly rejects the null for $h = 0$, for both accuracy measures, and at the 5% but not the 1% level for $h = 4$.

8 Are Efficient Forecasters More or Less Accurate, or More or Less Contrarian?

We use the MZ test results to categorize individual forecasters as efficient or non-efficient. Recall that efficient means that the forecasts are not correlated with the forecast errors. Our results are for each forecaster for each of the three variables. A forecaster is designated as efficient overall if we fail to reject the null for each of the 3 variables separately at the x% level. We report our findings for two values of x, where x is chosen judiciously to ensure that not all forecasters are efficient, or alternatively, not efficient. We then calculate a test of the null hypothesis that the forecasters in the two groups (Efficient, non-Efficient) have accuracy levels (or disagreement values) which are consistent with the population mean values of the two groups being equal. The sizes of the samples from the two groups are allowed

to be unequal (and will depend on x), and we do not assume that the population variances of the two groups are equal.[18]

For each pair of values $\{h, x\}$ in table 13 we present two rows: the first compares the relatively more and less efficient groups of forecasters based on (6) - using the shorter-horizon forecasts on the LHS of the MZ regression. The second characterizes forecasters using actual values on the LHS (equation (3)).

If we use the actual values to assess forecast efficiency, table 13 suggests more efficient forecasters are more accurate forecasters at $h = 4$, whether we consider the top 15 or top 44 forecasters. At $h = 0$ the evidence points in the same direction (using the MZ test with actual values), but the test outcomes are not significant at conventional significance levels. For the top 44 forecasters, there is evidence that efficient forecasts disagree less at $h = 4$. The results point in the same direction for the top 14, and for both sets of forecasters at $h = 0$, but are generally not significant.

Using (6) to characterize individuals as efficient or not generally yields to quite different findings at $h = 0$. That this might occur is perhaps not surprising given the discussion in section 5 concerning the potential differences between testing for efficiency based on (3) and (6), and the comparison of the test outcomes across individual forecasters in tables 6 to 8. For consumption we find a much greater concordance between the two tests at the longer horizon. For $h = 0$ the signs of the tests for the equality of the population means of accuracy and disagreement switch between the (3) and (6) tests. Using the latter, there is statistically significant evidence that more efficient forecasters are less accurate and more contrarian, if we restrict attention to the top 15.[19]

This may be an artefact of the way we have implemented the tests. For the comparison of forecast attributes at $h = 4$, forecast efficiency based on equation (6) regresses the shorter horizon forecast on the $h = 4$ forecast, i.e., $y_{i,t|t-3}$ on $y_{i,t|t-4}$, and (3) regresses y_t on $y_{i,t|t-4}$. The sub-tables to tables 6 to 8 show these two tests yield the same inference about efficiency for between 60 to 87% of forecasters, depending on the variable. When we consider the characteristics of the $h = 0$ forecasts, we determine their efficiency by comparing the (6) regression of $y_{i,t|t}$ on $y_{i,t|t-1}$ with the (3) regression of y_t on $y_{i,t|t-1}$. Hence to enable a comparison of the two approaches, we are effectively analyzing the relationship between the $h = 0$ and $h = 1$ forecasts for (6), and between the acuals and the $h = 1$ forecasts, and these yield quite different inference regarding the relationship between forecast efficiency, accuracy and disagreement.

For the efficiency test using actual values, we can also regress y_t on $y_{i,t|t}$: doing so suggests stronger evidence that more efficient forecasters are more accurate and less contrarian. We reject the null for both accuracy and disagreement at the 10% level in one-sided tests (not shown). Hence the discrepancies resulting from using (3) or (6) are increased by this change, and for the top 44 forecasters there is now clear evidence that the efficient forecasters are both more accurate and less contrarian.

Given the relatively small number of observations underlying the tests of equal population means, especially for the top 15 forecasters, as a robustness check we excluded the most extreme forecaster, in terms of accuracy and disagreement (it happened to be the same forecasters). Table 14 reports the results for $x = 0.005$, and using y_t on $y_{i,t|t}$ in the MZ test for characterizing the $h = 0$ forecasts. The results are essentially unchanged. Efficient forecasters are more accurate as a group, and the results point towards efficient forecasters being less contrarian, but this is only found to be statistically significant for

[18] This is the test used in section 3 to compare forecasters against the median forecaster in terms of disagreement.

[19] Note that the table entries for $h = 0$, $x = 0.005$ and $h = 0$, $x = 0.01$ are identical. This is because exactly the same individuals are identified as efficient at the two values of x: this is evident from the number of forecasters in each group (N_1 and N_2) being the same.

the larger group's 4-step ahead forecasts.

Substituting the shorter-horizon forecasts in MZ tests of forecast efficiency generates very different results concerning the relationships between efficiency, accuracy and disagreement at the short horizon.

9 Conclusions

Our investigation suggests that some forecasters tend to be more contrarian - in the sense of standing apart from the crowd - than others. That is, at any point in time the level of the overall disagreement between forecasters is more likely to be due to a given set of forecasters, as opposed to any randomly-selected set of forecasters. Our findings are based on a multivariate measure that adjusts for the degree of difficulty in forecasting at each point in time, and also downweights an individual's disagreement measure when he/she is in agreement with the consensus view about how the economy operates (albeit not the magnitudes of the future values of the variables). Much of the literature considers disagreement between forecasters as a possible proxy for uncertainty (beginning with the seminal paper by Zarnowitz and Lambros (1987)), and considers how it varies with the state of the business cycle. However, individual forecasters are not identified, and the implicit assumption seems to be that any one forecaster is as likely to make the same contribution to overall disagreement at any point in time as any other. Our findings suggest this is not the case.

In terms of forecast accuracy, we also find evidence that the SPF forecasters are not interchangeable. There is clear evidence that the rankings of forecasters across the two periods are systematic for the short-horizon forecasts, and we demonstrate the importance of controlling for differences over time in the difficulty of forecasting. At least at the short horizon, there do appear to be 'innate' differences between forecasters (in the sense of not arising from sampling variation: see D'Agostino *et al.* (2012)).

Finally, we consider whether forecasters make efficient use of their information sets when they forecast, in the sense of Mincer and Zarnowitz (1969), by testing whether forecasts and forecast errors are correlated. We reject efficiency for around half of all forecasters at short horizons (depending on the variable in question) and for even higher fractions at the longer horizon ($h = 4$). Our results broadly indicate that efficient forecasters tend to be more accurate and less contrarian. Overall this finding suggests that forecasters who persistently tend to stand out from the crowd produce less accurate forecasts, and do not use their information efficiently. Our results do not support the notion that contrarian forecasts stand apart by virtue of having superior information - knowing something that others do not.

Our results broadly hold across two groups of forecasters - the top 15 most prolific forecasters, and the top 44. For the smaller set of 15, the forecasters are on average more active as respondents (by construction), so that occurrences of specific forecasters adventitiously happening to respond mainly at difficult times (or easy times) will be lessened. (And in any case, an attempt is made to adjust the the accuracy and disagreement estimates for the difficulty in forecasting). Against that advantage, comparisons are based on only 15 data points. For the top 44, the estimates of accuracy and disagreement, and the test outcome for efficiency, are less precise being based on fewer forecast observations per individual on average.

Given the relatively small number of data points, particularly in the case of the top 15, we have used rank correlation tests to establish statistically significant associations, as these are more robust to extreme values.

The evidence presented in this paper suggests that macro-forecasters are not 'essentially the same' as each other, which is at odds with the implications of some popular models of expectations formation, and points to the importance of allowing agents to have different models of the economy, for example, as in Giacomini *et al.* (2015).

References

Andrade, P., Crump, R., Eusepi, S., and Moench, E. (2014). Fundamental disagreement. Working papers 524, Banque de France.

Andrade, P., and Le Bihan, H. (2013). Inattentive professional forecasters. *Journal of Monetary Economics*, *60*(8), 967–982.

Ball, L., Jalles, J. T., and Loungani, P. (2015). Do forecasters believe in Okun's Law? An assessment of unemployment and output forecasts. *International Journal of Forecasting*, *31*(1), 176–184.

Banternghansa, C., and McCracken, M. W. (2009). Forecast disagreement among FOMC members. Working papers 2009-059, Federal Reserve Bank of St. Louis.

Bomberger, W. A. (1996). Disagreement as a measure of uncertainty. *Journal of Money, Credit and Banking*, **28**, 381–392.

Capistrán, C., and Timmermann, A. (2009). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking*, **41**, 365–396.

Clements, M. P. (1995). Rationality and the role of judgement in macroeconomic forecasting. *Economic Journal*, **105**, 410–420.

Clements, M. P. (2014). Forecast Uncertainty - Ex Ante and Ex Post: US Inflation and Output Growth. *Journal of Business & Economic Statistics*, **32(2)**, 206–216. DOI: 10.1080/07350015.2013.859618.

Clements, M. P., and Hendry, D. F. (1993). On the limitations of comparing mean squared forecast errors. *Journal of Forecasting*, **12**, 617–637. With discussion. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

Clements, M. P. (2016). Long-run restrictions and survey forecasts of output, consumption and investment. *International Journal of Forecasting*, *32*(3), 614 – 628.

Coibion, O., and Gorodnichenko, Y. (2012). What can survey forecasts tell us about information rigidities?. *Journal of Political Economy*, *120*(1), 116 – 159.

Coibion, O., and Gorodnichenko, Y. (2015). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, *105*(8), 2644–78.

Croushore, D. (1993). Introducing: The Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Business Review, November*, 3–15.

Croushore, D. (2011a). Forecasting with real-time data vintages, chapter 9. In Clements, M. P., and Hendry, D. F. (eds.), *The Oxford Handbook of Economic Forecasting*, pp. 247–267: Oxford University Press.

Croushore, D. (2011b). Frontiers of real-time data analysis. *Journal of Economic Literature*, **49**, 72–100.

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

D'Agostino, A., McQuinn, K., and Whelan, K. (2012). Are some forecasters really better than others?. *Journal of Money, Credit and Banking*, **44(4)**, 715–732.

Davies, A., and Lahiri, K. (1995). A new framework for analyzing survey forecasts using three-dimensional panel data. *Journal of Econometrics*, **68**, 205–227.

Dovern, J. (2014). A Multivariate Analysis of Forecast Disagreement: Confronting Models of Disagreement with SPF Data. Working papers 0571, University of Heidelberg, Department of Economics.

Fixler, D. J., Greenaway-McGrevy, R., and Grimm, B. T. (2014). The revisions to GDP, GDI, and their major components. *Survey of Current Business*, **August**, 1–23.

Giacomini, R., Skreta, V., and Turen, J. (2015). Models, Inattention and Expectation Updates. Discussion papers 1602, Centre for Macroeconomics (CFM).

Hendry, D., and Martinez, A. B. (2016). Evaluating Multi-Step System Forecasts with Relatively Few Forecast-Error Observations. Economics series working papers 784, University of Oxford, Department of Economics.

Kandel, E., and Zilberfarb, B. Z. (1999). Differential interpretation of information in inflation forecasts. *The Review of Economics and Statistics*, **81**, 217–226.

King, R. G., Plosser, C. I., Stock, J. H., and Watson, M. W. (1991). Stochastic trends and economic fluctuations. *American Economic Review*, **81**, 819–840.

Komunjer, I., and Owyang, M. T. (2012). Multivariate Forecast Evaluation and Rationality Testing. *The Review of Economics and Statistics*, *94*(4), 1066–1080.

Kosobud, R., and Klein, L. (1961). Some econometrics of growth: Great ratios of economics. *Quarterly Journal of Economics*, **25**, 173–198.

Lahiri, K., and Sheng, X. (2008). Evolution of forecast disagreement in a Bayesian learning model. *Journal of Econometrics*, **144(2)**, 325–340.

Lamont, O. A. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior & Organization*, *48*(3), 265–280.

Landefeld, J. S., Seskin, E. P., and Fraumeni, B. M. (2008). Taking the pulse of the economy. *Journal of Economic Perspectives*, **22**, 193–216.

Laster, D., Bennett, P., and Geoum, I. S. (1999). Rational bias in macroeconomic forecasts. *The Quarterly Journal of Economics*, **114(1)**, 293–318.

López-Pérez, V. (2015). Does uncertainty affect participation in the European Central Bank's Survey of Professional Forecasters?. Working paper series no. 1807, European Central Bank.

Mankiw, N. G., and Reis, R. (2002). Sticky information versus sticky prices: a proposal to replace the New Keynesian Phillips Curve. *Quarterly Journal of Economics*, **117**, 1295–1328.

Mankiw, N. G., Reis, R., and Wolfers, J. (2003). Disagreement about inflation expectations. mimeo, National Bureau of Economic Research, Cambridge MA.

Manzan, S. (2011). Differential interpretation in the Survey of Professional Forecasters. *Journal of Money, Credit and Banking*, **43**, 993–1017.

Mincer, J., and Zarnowitz, V. (1969). The evaluation of economic forecasts. In Mincer, J. (ed.), *Economic Forecasts and Expectations*, pp. 3–46. New York: National Bureau of Economic Research.

Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *Review of Economics and Statistics*, **69**, 667–674.

Ottaviani, M., and Sorensen, P. N. (2006). The strategy of professional forecasting. *Journal of Financial Economics*, **81**, 441–466.

Patton, A. J., and Timmermann, A. (2010). Why do forecasters disagree? lessons from the term structure

of cross-sectional dispersion. *Journal of Monetary Economics*, **57(7)**, 803–820.

Patton, A. J., and Timmermann, A. (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, *30*(1), 1–17.

Rich, R., and Tracy, J. (2010). The relationships among expected inflation, disagreement, and uncertainty: Evidence from matched point and density forecasts. *Review of Economics and Statistics*, **92(1)**, 200–207.

Rich, R. W., and Butler, J. S. (1998). Disagreement as a measure of uncertainty: A comment on Bomberger. *Journal of Money, Credit and Banking*, **30**, 411–419.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, **50**, 665–690.

Sinclair, T. M., Stekler, H., and Carnow, W. (2015). Evaluating a vector of the FedâĂŹs forecasts. *International Journal of Forecasting*, *31*(1), 157–164.

Whelan, K. (2003). A Two-Sector Approach to Modeling U.S. NIPA Data. *Journal of Money, Credit and Banking*, *35*(4), 627–56.

Woodford, M. (2002). Imperfect common knowledge and the effects of monetary policy. In Aghion, P., Frydman, R., Stiglitz, J., and Woodford, M. (eds.), *Knowledge, Information, and Expectations in Modern Macroeconomics: In honor of Edmund Phelps*, pp. 25–58: Princeton University Press.

Zarnowitz, V., and Lambros, L. A. (1987). Consensus and uncertainty in economic prediction. *Journal of Political Economy*, **95(3)**, 591–621.

Table 1: Description of Forecast Data and Real-Time Data

Variable	SPF code	RTDSM code
Real GDP (GNP)	RGDP	ROUTPUT
Real personal consumption	RCONSUM	RCON
Real nonresidential fixed investment	RNRESIN	RINVBF
Real residential fixed investment	RRESINV	RINVRESID

The SPF data are from the Philadelphia Fed website (http://www.phil.frb.org/econ/spf/). For the investment series we used RNRESIN + RRESINV.
The real-time data were downloaded from http://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/.
Both sets of data were downloaded in April 2015.

Table 2: Multivariate Disagreement Statistics for Individuals with Most Responses

Weighting by cross-section covariances, S									
1	2	3	4	5	6	7	8	9	10
	$h=0$		$h=4$	Ranking for $h=0$			Ranking for $h=4$		
id.	Ave	Equal	Ave	All	1st half	2nd half	All	1st half	2nd half
20	2.91	1.00	3.13	15	15	15	15	15	15
84	1.32	0.56	1.64	9	9	7	13	12	14
411	1.00	0.00	1.12	2	2	5	4	1	6
420	1.30	0.50	1.51	8	13	4	11	14	4
421	1.50	0.96	1.68	11	8	13	14	13	11
426	1.84	1.00	1.44	14	12	14	8	4	13
428	1.11	0.03	1.35	5	3	6	7	8	8
429	1.42	0.84	1.52	10	10	9	12	10	12
431	1.09	0.03	0.96	3	4	2	1	3	2
433	0.88	0.00	1.09	1	1	1	3	5	1
446	1.11	0.05	1.07	4	7	3	2	2	3
456	1.30	0.49	1.46	7	5	10	9	11	9
463	1.59	0.99	1.47	13	11	12	10	9	10
472	1.29	0.47	1.35	6	6	8	6	6	7
484	1.53	0.96	1.32	12	14	11	5	7	5
Not weighting by cross-section covariances, diagonal S									
20	3.07	1.00	2.93	15	15	15	15	15	15
84	1.19	0.41	1.73	7	7	7	14	12	14
411	0.94	0.00	1.06	2	2	5	4	2	6
420	1.22	0.50	1.39	8	11	4	10	14	4
421	1.48	0.99	1.56	11	9	13	13	13	7
426	1.75	1.00	1.30	14	13	14	6	4	12
428	1.00	0.01	1.33	4	3	6	7	7	11
429	1.38	0.92	1.39	10	10	9	11	10	10
431	1.09	0.13	0.86	5	8	3	1	3	2
433	0.81	0.00	1.01	1	1	1	3	5	1
446	0.98	0.02	0.87	3	4	2	2	1	3
456	1.22	0.52	1.55	9	6	10	12	11	13
463	1.49	0.99	1.34	12	12	12	8	8	9
472	1.17	0.35	1.28	6	5	8	5	6	5
484	1.51	0.98	1.36	13	14	11	9	9	8

The top panel reports results for the multivariate disagreement measure, where $S_{t|t-h}$ is calculated as in eqn. (2), and the bottom panel sets $S_{t|t-h}$ to a diagonal matrix, so that $D_{i,t|t-h}$ corresponds to the Euclidean distance between the scaled forecast vector and consensus forecast vector at each point in time.
The 2nd and 4th columns denote the mean (across surveys) value of $D_{i,t|t-h}$ (eqn. 2) for $h=0$ and $h=4$. For $h=0$ we also report the p-values of testing the equality of means of each individual against the 'average' forecaster (precisely, the forecaster with the $N/2$ largest average disagreement), in column 3. The tests are constructed such that a p-value greater than 0.95 in column 3 has a larger population D_i, and a p-value less that 0.05 suggests a D_i significantly smaller than that of the average forecaster, in two-sided test at the 10% level). (The 0.50 entry in column 3 identifies the average forecaster).
Column 5 ranks the forecasters in terms of average $D_{i,t|t-h}$ (i.e., column 2) across all surveys for $h=0$, and columns 6 and 7 show the rankings if instead the averages are calculated on the first half of the sample or the second half of the sample (again for $h=0$). Columnns 8 to 10 given the ranking for the whole sample, and the two sub-samples, when $h=4$.

Table 3: Rank Correlation Tests

Panel 1. Top 15 forecasters

A. Earlier and later periods

S		Diag. S	
$h=0$	$h=4$	$h=0$	$h=4$
0.66	0.56	0.74	0.53
0.01	0.04	0.01	0.05

B. $h=0$ and $h=4$ forecasts

S			Diag. S		
Whole	Earlier	Later	Whole	Earlier	Later
0.72	0.54	0.80	0.65	0.53	0.72
0.01	0.04	0.00	0.02	0.05	0.01

Panel 2. Top 44 forecasters

A. Earlier and later periods

S		Diag. S	
$h=0$	$h=4$	$h=0$	$h=4$
0.48	0.52	0.48	0.31
0.01	0.00	0.01	0.08

B. $h=0$ and $h=4$ forecasts

S			Diag. S		
Whole	Earlier	Later	Whole	Earlier	Later
0.69	0.55	0.65	0.61	0.52	0.54
0.00	0.00	0.00	0.00	0.00	0.00

The Spearman rank correlation R lies between -1 and 1, and 0 indicates no relationship. For each test, there are two entries. The first row entry is the small-sample test statistic $\frac{6R-N\left(N^2-1\right)}{N(N+1)\sqrt{N-1}}$, with 2-sided critical values of ± 0.581 at the 5% level, and ± 0.443 at the 10% level, for $N=15$; and ± 0.298 at the 5% for $N=44$. The second row entry is the large sample test statistic, $1-\frac{6R}{N(N^2-1)}$, for which we report the chi-squared p-value.

Table 4: Forecast Accuracy Rankings, Most Prolific Forecasters. Normalized for differences over time in average forecast accuracy

id.	1990:4 - 2013:3			1990:4 - 2002:1			2002:2 - 2013:3		
	Tr.	Det.	N	Tr.	Det.	N	Tr.	Det.	N
	$h = 0$								
20	15	15	60	15	15	37	15	15	23
84	4	4	68	2	4	43	6	5	25
411	2	2	72	6	5	36	1	3	36
420	9	11	69	12	14	31	7	7	38
421	11	13	82	9	11	43	13	13	39
426	14	14	83	10	10	41	14	14	42
428	7	8	82	4	6	43	8	10	39
429	12	9	62	11	9	39	12	8	23
431	3	1	67	5	2	35	2	2	32
433	1	3	76	3	3	44	4	4	32
446	8	7	77	13	13	32	3	1	45
456	6	5	59	1	1	28	9	9	31
463	10	12	67	8	7	26	11	12	41
472	5	6	60	7	8	24	5	6	36
484	13	10	65	14	12	23	10	11	42
	$h = 4$								
20	15	15	59	15	15	36	15	15	23
84	5	5	54	11	10	31	4	2	23
411	9	7	72	5	2	36	11	8	36
420	6	8	69	13	12	31	3	4	38
421	14	13	81	14	14	43	13	13	38
426	7	9	82	10	7	41	7	11	41
428	10	10	82	7	9	43	9	10	39
429	4	3	62	6	6	39	6	5	23
431	1	1	66	3	4	35	1	1	31
433	2	2	76	1	1	44	2	3	32
446	3	4	77	4	3	32	5	6	45
456	12	12	59	8	11	28	12	12	31
463	13	14	67	12	13	26	14	14	41
472	8	11	58	9	8	23	8	9	35
484	11	6	65	2	5	23	10	7	42

The table ranks ('1' denotes the most accurate) each respondent by forecast accuracy - in terms of the trace (Tr.) or determinant (Det.) of the second moment matrix of forecast errors (MSFEM). It also shows the ranks for the Earlier and Later sub-samples.
The Spearman test of no relationship in the ranks between the two samples is reported below, where the first value is the small-sample test statistic, and the second is the p-value of the large-sample statistic. The critical values for the former are given in table 3.

$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det
Top 15 Forecasters			
0.43	0.41	0.48	0.62
0.11	0.12	0.07	0.02
Top 44 Forecasters			
0.52	0.44	0.16	0.30
0.00	0.01	0.34	0.09

Table 5: Forecast Accuracy Rankings, Most Prolific Forecasters. Not normalized for differences over time in average forecast accuracy

id.	1990:4 - 2013:3			1990:4 - 2002:1			2002:2 - 2013:3		
	Tr.	Det.	N	Tr.	Det.	N	Tr.	Det.	N
	$h = 0$								
20	14	15	60	15	15	37	14	15	23
84	2	3	68	2	2	43	3	5	25
411	7	5	72	7	8	36	6	4	36
420	11	11	69	14	13	31	10	10	38
421	15	12	82	11	12	43	15	12	39
426	12	13	83	12	11	41	12	14	42
428	9	10	82	4	4	43	13	9	39
429	4	4	62	8	7	39	4	3	23
431	10	6	67	10	5	35	9	6	32
433	3	2	76	5	6	44	2	1	32
446	8	7	77	9	10	32	7	7	45
456	1	1	59	1	3	28	1	2	31
463	5	9	67	3	1	26	8	11	41
472	6	8	60	6	9	24	5	8	36
484	13	14	65	13	14	23	11	13	42
	$h = 4$								
20	12	15	59	14	15	36	8	14	23
84	2	3	54	3	5	31	4	3	23
411	10	12	72	4	4	36	14	13	36
420	9	10	69	8	7	31	11	10	38
421	15	13	81	15	13	43	12	12	38
426	7	7	82	9	6	41	7	7	41
428	3	5	82	6	11	43	5	2	39
429	4	2	62	1	1	39	6	6	23
431	14	11	66	10	8	35	15	11	31
433	1	1	76	5	3	44	1	1	32
446	8	8	77	7	9	32	9	8	45
456	6	6	59	13	12	28	2	4	31
463	11	14	67	11	14	26	10	15	41
472	5	4	58	12	10	23	3	5	35
484	13	9	65	2	2	23	13	9	42

The table ranks ('1' denotes the most accurate) each respondent by forecast accuracy - in terms of the trace (Tr.) or determinant (Det.) of the second moment matrix of forecast errors (MSFEM). It also shows the ranks for the Earlier and Later sub-samples.
The Spearman test of no relationship in the ranks between the two samples is reported below, where the first value is the small-sample test statistic, and the second is the p-value of the large-sample statistic. The critical values for the former are given in table 3.

$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det
Top 15 Forecasters			
0.69	0.60	0.01	0.36
0.01	0.02	0.97	0.17
Top 44 Forecasters			
0.32	0.40	-0.05	-0.04
0.07	0.02	0.78	0.84

Table 6: Consumption growth: MZ tests with Short-Horizon Forecasts and Actual Values. 1990:Q4 to 2013:Q3

	MZ - LHS variable is a Forecast, Adjacent					MZ - LHS variable is an Actual Value					
id.	No. forecasts	$h_1 = 0$	$h_1 = 1$	$h_1 = 2$	$h_1 = 3$	No. forecasts	$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 4$
			$h_2 = h_1 + 1$								
20	50	0.29	0.16	0.05	0.03	62	0.00	0.00	0.00	0.00	0.00
84	49	0.06	0.05	0.03	0.09	58	0.03	0.04	0.08	0.10	0.78
411	56	0.01	0.01	0.45	0.00	70	0.10	0.60	0.16	0.69	0.03
420	48	0.62	0.19	0.00	0.00	66	0.01	0.01	0.14	0.72	0.03
421	69	0.97	0.47	0.00	0.43	78	0.15	0.30	0.24	0.00	0.00
426	74	0.72	0.00	0.00	0.02	80	0.00	0.00	0.01	0.00	0.00
428	72	0.68	0.01	0.02	0.00	79	0.00	0.01	0.01	0.01	0.01
429	53	0.40	0.18	0.00	0.00	62	0.02	0.08	0.33	0.02	0.00
431	53	0.21	0.18	0.16	0.13	66	0.01	0.35	0.27	0.46	0.09
433	66	0.04	0.00	0.00	0.00	72	0.00	0.00	0.00	0.02	0.03
446	68	0.49	0.06	0.05	0.05	73	0.05	0.44	0.03	0.00	0.05
456	47	0.00	0.05	0.00	0.00	57	0.00	0.00	0.00	0.03	0.00
463	59	0.02	0.56	0.83	0.01	64	0.21	0.01	0.14	0.02	0.00
472	46	0.06	0.06	0.00	0.00	56	0.32	0.07	0.36	0.04	0.00
484	55	0.15	0.03	0.10	0.10	62	0.00	0.00	0.08	0.00	0.00
Rejns.		0.27	0.40	0.67	0.67		0.67	0.60	0.40	0.73	0.80

Forecast Horizon	1	2	3	4
LHS variable is an Actual versus $h-1$ Adjacent Forecast				
(i) Both Reject	0.20	0.27	0.53	0.67
(ii) Neither rejects	0.33	0.47	0.13	0.20
(iii) Agree	0.53	0.74	0.66	0.87

The table displays p-values of the tests of MZ forecast optimality for each respondent, as well as the (minimum across tests) number of forecast observations for each respondent ('No. Forecasts'). The row 'Rejns' gives the proportion of respondents for which the null of optimality is rejected at the 5% level.
The bottom table gives the proportion of respondents for which the tests in the two panels (for a given h_H): (i) both reject the null, (ii) both fail to reject the null, (iii) are in agreement (the sum of (i) and (ii)); all at the 5% level.

Table 7: Investment growth: MZ tests with Short-Horizon Forecasts and Actual Values. 1990:Q4 to 2013:Q3

	MZ - LHS variable is a Forecast, Adjacent					MZ - LHS variable is an Actual Value					
id.	No. forecasts	$h_1 = 0$	$h_1 = 1$	$h_1 = 2$	$h_1 = 3$	No. forecasts	$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 4$
			$h_2 = h_1 + 1$								
20	50	0.24	0.00	0.02	0.00	62	0.00	0.10	0.00	0.00	0.00
84	49	0.14	0.27	0.15	0.20	58	0.10	0.97	0.97	0.20	0.03
411	53	0.33	0.01	0.00	0.25	68	0.42	0.50	0.51	0.00	0.00
420	48	0.15	0.02	0.13	0.01	66	0.11	0.31	0.39	0.10	0.19
421	71	0.20	0.00	0.00	0.00	79	0.40	0.23	0.04	0.00	0.00
426	71	0.14	0.00	0.03	0.00	78	0.28	0.34	0.31	0.81	0.00
428	72	0.14	0.75	0.02	0.35	79	0.44	0.38	0.57	0.01	0.11
429	53	0.19	0.08	0.15	0.19	62	0.00	0.04	0.52	0.86	0.02
431	53	0.00	0.02	0.01	0.23	67	0.01	0.03	0.81	0.18	0.14
433	66	0.85	0.12	0.00	0.00	72	0.64	0.58	0.86	0.19	0.22
446	68	0.17	0.00	0.00	0.00	73	0.05	0.75	0.65	0.09	0.00
456	47	0.73	0.09	0.94	0.35	57	0.04	0.04	0.01	0.13	0.09
463	57	0.09	0.89	0.66	0.67	63	0.57	0.10	0.21	0.08	0.25
472	46	0.28	0.01	0.18	0.00	56	0.23	0.74	0.22	0.13	0.02
484	55	0.37	0.31	0.26	0.01	62	0.28	0.53	0.88	0.25	0.00
Rejns.		0.07	0.53	0.53	0.53		0.33	0.20	0.20	0.27	0.60

Forecast Horizon	1	2	3	4
LHS variable is an Actual versus $h-1$ Adjacent Forecast				
(i) Both Reject	0.07	0.13	0.27	0.40
(ii) Neither rejects	0.80	0.40	0.47	0.27
(iii) Agree	0.87	0.53	0.74	0.67

The table displays p-values of the tests of MZ forecast optimality for each respondent, as well as the (minimum across tests) number of forecast observations for each respondent ('No. Forecasts'). The row 'Rejns' gives the proportion of respondents for which the null of optimality is rejected at the 5% level.
The bottom table gives the proportion of respondents for which the tests in the two panels (for a given h_H): (i) both reject the null, (ii) both fail to reject the null, (iii) are in agreement (the sum of (i) and (ii)); all at the 5% level.

Table 8: Output growth: MZ tests with Short-Horizon Forecasts and Actual Values. 1990:Q4 to 2013:Q3

	MZ - LHS variable is a Forecast, Adjacent					MZ - LHS variable is an Actual Value					
id.	No. forecasts	$h_1 = 0$	$h_1 = 1$	$h_1 = 2$	$h_1 = 3$	No. forecasts	$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 4$
			$h_2 = h_1 + 1$								
20	50	0.32	0.10	0.08	0.01	63	0.00	0.00	0.00	0.00	0.00
84	50	0.87	0.24	0.00	0.16	59	0.16	0.38	0.91	0.04	0.02
411	56	0.00	0.02	0.05	0.00	70	0.34	0.67	0.06	0.00	0.00
420	56	0.68	0.31	0.00	0.00	72	0.45	0.32	0.80	0.19	0.08
421	71	0.66	0.24	0.00	0.12	79	0.76	0.76	0.57	0.00	0.00
426	74	0.02	0.03	0.00	0.16	80	0.00	0.06	0.00	0.00	0.01
428	72	0.02	0.06	0.00	0.00	79	0.20	0.95	0.00	0.00	0.00
429	53	0.00	0.03	0.07	0.01	62	0.01	0.65	0.08	0.01	0.00
431	54	0.02	0.00	0.11	0.01	69	0.15	0.93	0.43	0.10	0.01
433	66	0.00	0.00	0.01	0.03	72	0.35	0.56	0.00	0.00	0.02
446	68	0.06	0.05	0.09	0.09	73	0.12	0.99	0.38	0.09	0.02
456	47	0.00	0.31	0.00	0.00	57	0.00	0.04	0.00	0.00	0.00
463	59	0.02	0.00	0.00	0.00	64	0.43	0.01	0.00	0.00	0.00
472	48	0.46	0.25	0.11	0.00	57	0.80	0.64	0.00	0.00	0.00
484	55	0.02	0.00	0.06	0.28	62	0.00	0.42	0.41	0.02	0.00
Rejns.		0.60	0.53	0.60	0.67		0.33	0.20	0.47	0.80	0.93

Forecast Horizon	1	2	3	4
LHS variable is an Actual versus $h-1$ Adjacent Forecast				
(i) Both Reject	0.13	0.20	0.53	0.60
(ii) Neither rejects	0.33	0.20	0.13	0.00
(iii) Agree	0.46	0.40	0.66	0.60

The table displays p-values of the tests of MZ forecast optimality for each respondent, as well as the (minimum across tests) number of forecast observations for each respondent ('No. Forecasts'). The row 'Rejns' gives the proportion of respondents for which the null of optimality is rejected at the 5% level.
The bottom table gives the proportion of respondents for which the tests in the two panels (for a given h_H): (i) both reject the null, (ii) both fail to reject the null, (iii) are in agreement (the sum of (i) and (ii)); all at the 5% level.

Table 9: Consumption growth: ORR tests with Short-Horizon Forecasts and Actual Values. 1990:Q4 to 2013:Q3

	ORR - LHS variable is a Forecast					ORR - LHS variable is an Actual Value				
id.	No. forecasts	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$	No. forecasts	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$
20	31	0.29	0.00	0.01	0.46	36	0.00	0.00	0.00	0.00
84	30	0.06	0.20	0.35	0.27	33	0.04	0.03	0.01	0.25
411	35	0.01	0.04	0.15	0.16	41	0.58	0.80	0.23	0.08
420	15	0.62	0.88	0.15	0.00	23	0.00	0.00	0.10	0.01
421	44	0.97	0.03	0.04	0.00	51	0.25	0.29	0.12	0.14
426	58	0.72	0.61	0.81	0.30	63	0.00	0.00	0.00	0.00
428	56	0.68	0.15	0.23	0.05	61	0.00	0.00	0.00	0.00
429	35	0.40	0.16	0.27	0.36	40	0.09	0.01	0.00	0.00
431	25	0.21	0.10	0.05	0.02	31	0.28	0.42	0.00	0.00
433	59	0.04	0.05	0.30	0.66	61	0.00	0.00	0.00	0.00
446	56	0.49	0.58	0.11	0.01	60	0.43	0.30	0.17	0.15
456	27	0.00	0.00	0.00	0.01	32	0.00	0.00	0.00	0.00
463	47	0.02	0.04	0.31	0.33	49	0.02	0.02	0.08	0.16
472	34	0.06	0.24	0.35	0.00	37	0.08	0.19	0.35	0.16
484	40	0.15	0.33	0.16	0.02	43	0.00	0.02	0.08	0.16
Rejns.		0.27	0.40	0.27	0.47		0.60	0.67	0.53	0.53

	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$
(i) Both reject	0.20	0.27	0.20	0.20
(ii) Neither rejects	0.33	0.20	0.40	0.20
(iii) Agree	0.53	0.47	0.60	0.40

The table displays p-values of the tests of ORR forecast optimality for each respondent, as well as the (minimum across tests) number of forecast observations for each respondent ('No. Forecasts'). The row 'Rejns' gives the proportion of respondents for which the null of optimality is rejected at the 5% level.

The last 3 rows give the proportion of respondents for which the tests in the left and right panels (for a given h_H) (i) both reject the null, (ii) both fail to reject the null, (iii) are in agreement (the sum of (i) and (ii)); all at the 5% level.

Table 10: Investment growth: ORR tests with Short-Horizon Forecasts and Actual Values. 1990:Q4 to 2013:Q3

	ORR - LHS variable is a Forecast					ORR - LHS variable is an Actual Value				
id.	No. forecasts	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$	No. forecasts	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$
20	35	0.24	0.62	0.89	0.04	40	0.05	0.26	0.00	0.01
84	31	0.14	0.59	0.61	0.80	33	0.97	0.61	0.65	0.05
411	33	0.33	0.44	0.49	0.69	38	0.50	0.29	0.37	0.61
420	15	0.15	0.36	0.55	0.45	23	0.31	0.02	0.16	0.01
421	48	0.20	0.19	0.05	0.02	55	0.24	0.14	0.19	0.00
426	53	0.14	0.24	0.26	0.14	59	0.23	0.39	0.09	0.00
428	55	0.14	0.01	0.01	0.02	61	0.39	0.02	0.05	0.04
429	35	0.19	0.33	0.21	0.00	40	0.03	0.04	0.10	0.01
431	26	0.00	0.00	0.00	0.00	32	0.04	0.07	0.00	0.00
433	59	0.85	0.97	0.26	0.16	61	0.54	0.75	0.90	0.68
446	56	0.17	0.23	0.15	0.03	60	0.72	0.86	0.93	0.94
456	27	0.73	0.89	0.97	0.99	32	0.01	0.00	0.15	0.22
463	42	0.09	0.14	0.07	0.06	45	0.07	0.14	0.08	0.11
472	34	0.28	0.00	0.00	0.08	37	0.78	0.15	0.11	0.48
484	40	0.37	0.11	0.12	0.25	43	0.50	0.32	0.70	0.76
Rejns.		0.07	0.20	0.20	0.40		0.20	0.27	0.13	0.53

	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$
(i) Both reject	0.07	0.07	0.07	0.33
(ii) Neither rejects	0.80	0.60	0.73	0.40
(iii) Agree	0.87	0.67	0.80	0.73

The table displays p-values of the tests of ORR forecast optimality for each respondent, as well as the (minimum across tests) number of forecast observations for each respondent ('No. Forecasts'). The row 'Rejns' gives the proportion of respondents for which the null of optimality is rejected at the 5% level.

The last 3 rows give the proportion of respondents for which the tests in the left and right panels (for a given h_H) (i) both reject the null, (ii) both fail to reject the null, (iii) are in agreement (the sum of (i) and (ii)); all at the 5% level.

Table 11: Output growth: ORR tests with Short-Horizon Forecasts and Actual Values. 1990:Q4 to 2013:Q3

	ORR - LHS variable is a Forecast					ORR - LHS variable is an Actual Value				
id.	No. forecasts	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$	No. forecasts	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$
20.00	32	0.32	0.62	0.98	0.96	37	0.00	0.00	0.00	0.00
84.00	31	0.87	0.80	0.44	0.42	33	0.36	0.20	0.00	0.02
411.00	35	0.00	0.02	0.08	0.11	41	0.58	0.04	0.00	0.01
420.00	25	0.68	0.61	0.36	0.33	33	0.27	0.08	0.09	0.03
421.00	48	0.66	0.39	0.24	0.38	55	0.73	0.89	0.09	0.00
426.00	58	0.02	0.03	0.06	0.11	63	0.02	0.00	0.00	0.00
428.00	56	0.02	0.02	0.09	0.14	61	0.95	0.00	0.01	0.00
429.00	35	0.00	0.01	0.02	0.00	40	0.63	0.01	0.00	0.00
431.00	26	0.02	0.01	0.02	0.00	32	0.92	0.02	0.00	0.00
433.00	59	0.00	0.00	0.00	0.01	61	0.54	0.00	0.01	0.00
446.00	56	0.06	0.00	0.00	0.00	60	0.98	0.01	0.01	0.01
456.00	27	0.00	0.00	0.00	0.00	32	0.03	0.12	0.02	0.00
463.00	47	0.02	0.03	0.09	0.08	49	0.01	0.01	0.20	0.05
472.00	38	0.46	0.71	0.48	0.04	41	0.64	0.17	0.19	0.00
484.00	38	0.02	0.06	0.21	0.08	42	0.33	0.70	0.95	0.97
Rejns.		0.60	0.60	0.33	0.40		0.27	0.60	0.67	0.93

	$h_H = 1$	$h_H = 2$	$h_H = 3$	$h_H = 4$
(i) Both reject	0.20	0.53	0.33	0.40
(ii) Neither rejects	0.33	0.33	0.33	0.07
(iii) Agree	0.53	0.86	0.66	0.47

The table displays p-values of the tests of ORR forecast optimality for each respondent, as well as the (minimum across tests) number of forecast observations for each respondent ('No. Forecasts'). The row 'Rejns' gives the proportion of respondents for which the null of optimality is rejected at the 5% level.

The last 3 rows give the proportion of respondents for which the tests in the left and right panels (for a given h_H) (i) both reject the null, (ii) both fail to reject the null, (iii) are in agreement (the sum of (i) and (ii)); all at the 5% level.

Table 12: Rank Correlation Tests: Accuracy and Disagreement

Top 15 forecasters			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det.
0.89	0.87	0.56	0.59
0.00	0.00	0.04	0.03
Top 44 forecasters			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det.
0.93	0.91	0.77	0.76
0.00	0.00	0.00	0.00

The Spearman test of no relationship in the accuracy ranks (either trace or determinant measure) and the disagreement ranks. In each row the first value is the small-sample test statistic, and the second is the p-value of the large-sample statistic. The critical values for the former are given in table 3.

Table 13: Are Efficient Forecasters More or Less Likely to be Accurate or Contrarian?

Accuracy Tr.		Accuracy Det.		Disagreement			
p-val	Stat.	p-val	Stat.	p-val	Stat.	N_1	N_2
Top 15							
$h = 0$, $x = 0.005$							
0.07	1.55	0.12	1.28	0.03	2.14	10	5
0.85	-1.20	0.82	-1.05	0.85	-1.21	10	5
$h = 0$, $x = 0.01$							
0.07	1.55	0.12	1.28	0.03	2.14	10	5
0.89	-1.31	0.85	-1.14	0.90	-1.36	7	8
$h = 4$, $x = 0.005$							
0.91	-1.45	0.95	-1.76	0.78	-0.83	3	12
0.99	-2.53	0.95	-1.80	0.83	-1.01	4	11
$h = 4$, $x = 0.01$							
0.95	-1.91	0.95	-1.83	0.68	-0.54	2	13
0.99	-2.53	0.95	-1.80	0.83	-1.01	4	11
Top 44							
$h = 0$, $x = 0.005$							
0.39	0.28	0.42	0.20	0.27	0.63	24	20
0.79	-0.82	0.73	-0.62	0.84	-1.02	29	15
$h = 0$, $x = 0.01$							
0.46	0.10	0.42	0.19	0.33	0.45	22	22
0.70	-0.54	0.64	-0.36	0.71	-0.56	23	21
$h = 4$, $x = 0.005$							
0.60	-0.27	0.92	-1.40	0.89	-1.28	12	32
1.00	-4.97	1.00	-3.32	0.99	-2.46	13	31
$h = 4$, $x = 0.01$							
0.86	-1.12	0.96	-1.79	0.95	-1.69	10	34
1.00	-4.93	1.00	-3.22	0.99	-2.40	10	34

The table consists of 3 pairs of columns - for accuracy as measured by the trace, for accuracy as measured by the determinant, and for disagreement. In each case we report the p-values and the test statistic value. The critical values for the test statistic are given in table 3. The large-sample p-values are calculated such that a value greater than 0.90 indicates the efficient group are more accurate, or have a lower level of disagreement, at the 10% level in a one-sided test. A value less than 0.10 indicates the opposite - the efficient group are less accurate or have a higher level of disagreement, at the 10% level.
For each pair of values h, x there are two rows. The first row is for forecast efficiency determined by equation (6), i.e., using a shorter-horizon forecast as the depenent variable, and the second row is for forecast efficiency determined by equation (3), i.e., using the actual value as the dependent variable.

Table 14: Efficient Forecasters and Accuracy and Disagreement: Removing the most Extreme Forecaster

Accuracy Tr.		Accuracy Det.		Disagreement			
p-val	Stat.	p-val	Stat.	p-val	Stat.	N_1	N_2
Top 15							
$h = 0$, $x = 0.005$							
0.08	1.52	0.01	2.62	0.03	2.08	9	5
0.86	-1.15	0.72	-0.59	0.67	-0.46	8	6
$h = 4$, $x = 0.005$							
0.82	-1.03	0.97	-2.18	0.61	-0.29	3	11
0.99	-2.83	0.98	-2.28	0.66	-0.45	4	10
Top 44							
$h = 0$, $x = 0.005$							
0.63	-0.34	0.77	-0.74	0.43	0.18	23	20
0.89	-1.30	0.89	-1.26	0.84	-1.03	30	13
$h = 4$, $x = 0.005$							
0.48	0.04	0.83	-0.97	0.82	-0.93	12	31
1.00	-4.95	1.00	-4.35	0.98	-2.23	13	30

As table 13.
Except that 1) Forecaster id 20 is excluded, and 2) the MZ regressions for characterizing the $h = 0$ forecasts using the actual values regress the actuals on to the $h = 0$ forecasts (not the $h = 1$ forecasts).

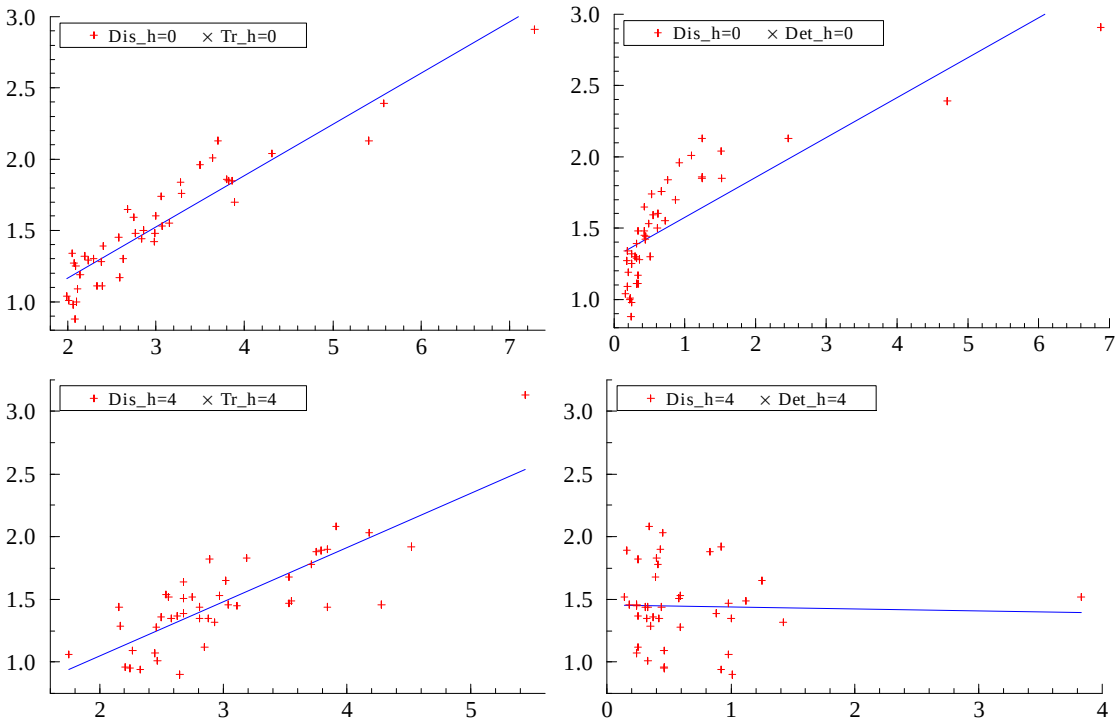

Figure 1: Crossplots of disagreement and forecast accuracy for the 44 most prolific forecasters. Tr and Det denote the trace and determinant of the MSFE matrix.

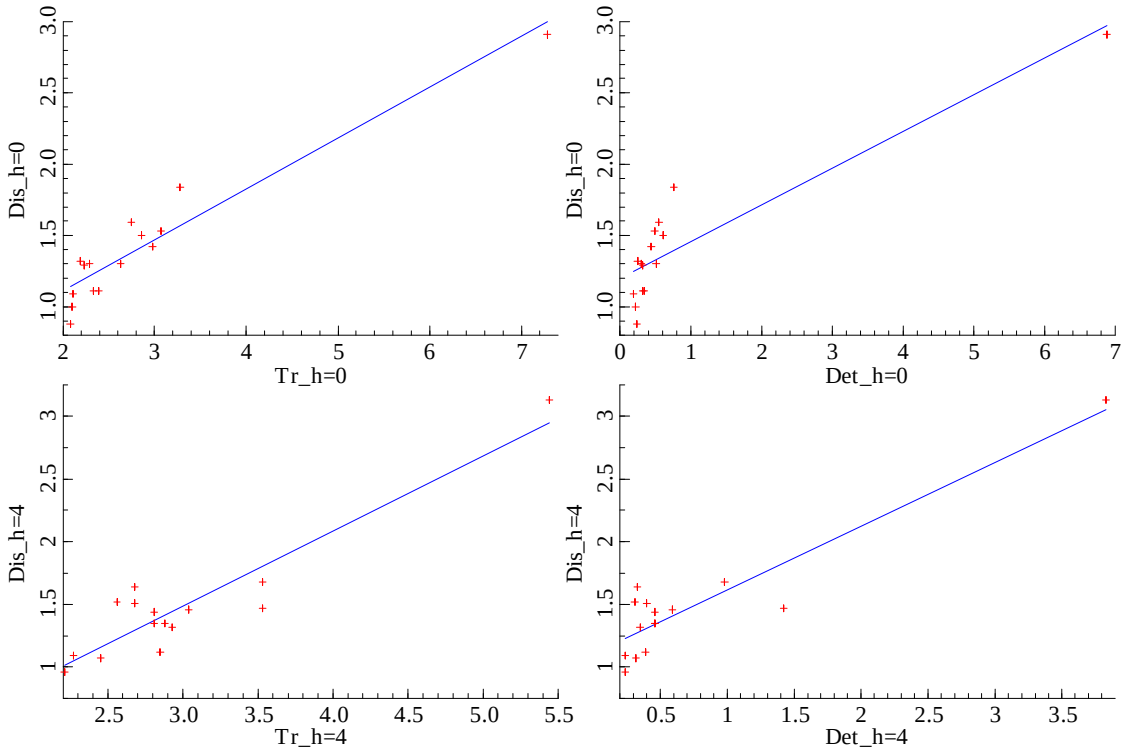

Figure 2: Crossplots of disagreement and forecast accuracy for the 15 most prolific forecasters. Tr and Det denote the trace and determinant of the MSFE matrix.

10 Appendix A

10.1 Appendix 1. Data Cleaning.

We considered a number of apparently extreme observations, but ended up discarding only 3 forecasts. In the interests of transparency and replicability, these are briefly described below.

Survey quarter 1991:4, forecaster id. 422. Real output. The 4-quarter ahead is 4135.1. This gives a marked drop from the forecasts of the previous quarters. Were it 4315 instead, the forecasts of the four quarters of 1992 would equal the reported forecast for calendar 1992. Given this additional corroboration, we replace the forecast of the 1992:4 quarter by a missing value.

Survey quarter 1993:4, forecasters id 421. Real consumption. The forecast of 1994:1 is for 3418.3, which is out of kilter with the forecasts of the other quarters and with the reported forecast for 1994. The forecast of 1994:1 is replaced by a missing.

Survey quarter 1992:4, forecaster id 414. Real non-residential investment (component of total fixed investment). Forecast of 1993:3 shows a sharp drop, inconsistent with the forecast for 1993, and is replaced by a missing value.

11 Appendix B

11.1 Actual values versus short(er) horizon forecasts

The use of adjacent horizon forecasts in the tests of forecast optimality serves to circumvent the need for autocorrelation corrections. Consider the MZ regression given by:

$$y_{t|t-h_1} = \delta + \delta_1 y_{t|t-h_2} + u_t \tag{11}$$

where $h_2 = h_1 + 1$, i.e., that the forecasts are adjacent, then under the null the error term in the regression will be serially uncorrelated. To understand why this is the case, consider for example the two rows of the regression corresponding to the targets t and $t+1$. Under the null, $y_{t|t-h_1} - y_{t|t-h_2} = u_t$ and $y_{t+1|t+1-h_1} - y_{t+1|t+1-h_2} = u_{t+1}$. Suppose the time series y_t is written as an infinite-order moving average:

$$y_t = \psi(L)\varepsilon_t = \sum_{j=1}^{h_1} \psi_{h_1-j}\varepsilon_{t-h_1+j} + \psi_{h_1}\varepsilon_{t-h_1} + \ldots + \psi_{h_2-1}\varepsilon_{t-h_2+1} + \sum_{j=0}^{\infty} \psi_{h_2+j}\varepsilon_{t-h_2-j},$$

which collapses to:

$$y_t = \psi(L)\varepsilon_t = \sum_{j=1}^{h} \psi_{h-j}\varepsilon_{t-h+j} + \psi_h\varepsilon_{t-h} + \sum_{j=0}^{\infty} \psi_{h+1+j}\varepsilon_{t-h-1-j},$$

when $h_2 - 1 = h_1 \equiv h$. The lag polynomials are written as above to clearly show the past, present and future (relative to the forecast origin) components.

In the general case,

$$\begin{aligned} y_{t|t-h_1} &= E\left(y_t \mid \mathcal{I}_{t-h_1}\right) = \psi_{h_1}\varepsilon_{t-h_1} + \ldots + \psi_{h_2-1}\varepsilon_{t-h_2+1} + \textstyle\sum_{j=0}^{\infty}\psi_{h_2+j}\varepsilon_{t-h_2-j} \\ y_{t|t-h_2} &= E\left(y_t \mid \mathcal{I}_{t-h_1}\right) = \textstyle\sum_{j=0}^{\infty}\psi_{h_2+j}\varepsilon_{t-h_2-j}, \end{aligned}$$

so that $y_{t|t-h_1} - y_{t|t-h_2} = \psi_{h_1}\varepsilon_{t-h_1} + \ldots + \psi_{h_2-1}\varepsilon_{t-h_2+1}$. (Here $\mathcal{I}_{t-h}$ denotes the information set at time $t-h$, and consists of $\varepsilon_{t-h}, \varepsilon_{t-h-1}, \ldots$.) For the target $t+1$, we have

$$y_{t+1|t+1-h_1} - y_{t+1|t+1-h_2} = \psi_{h_1}\varepsilon_{t+1-h_1} + \ldots + \psi_{h_2-1}\varepsilon_{t+1-h_2+1}$$

(by simply replacing 't' by '$t+1$'). In general, then $Cov\left(y_{t|t-h_1} - y_{t|t-h_2}, y_{t+1|t+1-h_1} - y_{t+1|t+1-h_2}\right) \neq 0$ since the two revisions have common ε's. But when $h_2 - 1 = h_1$,

$$Cov\left(y_{t|t-h_1} - y_{t|t-h_2}, y_{t+1|t+1-h_1} - y_{t+1|t+1-h_2}\right) = Cov\left(\psi_h\varepsilon_{t-h}\psi_h\varepsilon_{t+1-h}\right) = 0,$$

and no autocorrelation-correction is needed.

The advantage of using short-horizon forecasts in place of actuals is immediately apparent. It is a simple matter to check that the regression errors would be correlated for, say:

$$y_t = \delta + \delta_1 y_{t|t-h} + u_t$$

whenever $h > 1$.[20]

The upshot is that for simple MZ regressions we estimate:

$$y_{t|t-h} = \delta + \delta_1 y_{t|t-(h+1)} + u_t$$

for $h = 1, 2, \ldots$.

Note that the ORR regressions do not require autocorrelation-consistent estimation of the covariance matrix of the regression parameter estimates. Using the $h_1 = 0$ forecast as the actual value, provided the next-shortest horizon forecast $h_2 = h_1 + 1 = 1$, then under the null the regression error is $y_{t|t-h_1} - y_{t|t-h_2} = u_t$, which is serially uncorrelated by the above results.

[20] Under the null, $y_t - y_{t|t-h} = \sum_{j=1}^{h}\psi_{h-j}\varepsilon_{t-h+j}$, and $y_{t+1} - y_{t+1|t+1-h} = \sum_{j=1}^{h}\psi_{h-j}\varepsilon_{t+1-h+j}$, and so $Cov\left(y_t - y_{t|t-h}, y_{t+1} - y_{t+1|t+1-h}\right) \neq 0$ when $h > 1$.