ICMA Centre

Henley Business School
UNIVERSITY OF READING

Discussion Paper

Forecasting and Forecast Narratives: The Bank of England Inflation Reports

November 2016

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

J James Reade
Department of Economics, School of Politics, Economics and International Relations, University of Reading

Discussion Paper Number:
ICM-2016-10

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

Forecasting and Forecast Narratives: The Bank of England Inflation Reports*

Michael P. Clements
ICMA Centre
Henley Business School
University of Reading
M.P.Clements@reading.ac.uk

J. James Reade
Department of Economics
School of Politics, Economics and International Relations
University of Reading
J.J.Reade@reading.ac.uk

November 21, 2016

Abstract

We analyze the narratives that accompany the numerical forecasts in the Bank of England's *Inflation Reports*. We focus on whether the narratives contain useful information about the future course of key macro variables over and above the point predictions, in terms of whether the narratives can be used to enhance the accuracy of the numerical forecasts. We also consider whether the narratives are able to predict future changes in the numerical forecasts. We find that sentiment related to specific aspects of the economic outlook (say, demand conditions, or supply conditions) can enhance point forecast performance, and changes in sentiment (expressed in the narrratives) can predict subsequent changes in the point forecasts.

Keywords. Sentiment indices, inflation forecasting, output growth, forecast encompassing. C55, E37, E66.

*The authors would like to thank seminar participants at the University of Birmingham, De Nederlandsche Bank, BI Norwegian Business School, and conference participants at the 16th OxMetrics User Workshop in Washington DC, and the 36th International Symposium on Forecasting for their helpful comments.

1 Introduction

Forecasts are sometimes accompanied by a narrative account. This narrative is often used to provide additional information regarding the numerical forecasts and the outlook more generally, and might include an assessment of the extent of uncertainty associated with the forecast, as well as a discussion of some of the key features and drivers of the forecast numbers. The narrative provides an accompanying 'story' to the forecast, that sets the scene, and in so doing may enhance the credibility of the forecast, by providing explanations for the future trends and developments implicit in the forecast numbers.

Although the narrative is qualitative in nature, it is nonetheless possible to convert this qualitative information into quantitative data. This then facilitates a number of areas of research. These include the following inter-related issues: whether the narrative contains information not reflected in published forecasts; whether the narratives can be used to improve the numerical forecasts; whether the narratives provide an 'early-warning' system for adverse developments which are not fully reflected in the numerical forecasts; whether various behavioural biases often found in agents' forecasts affect the numerical and narrative forecasts in the same way; whether the implicit 'loss function' for the narratives and numerical forecasts is the same; whether the narratives are informative about forecast uncertainty (whether the numerical forecasts take the form of central projections and/or 'fan charts' or probability distributions describing future outcomes).

The use of narrative analysis in economic forecasting is in its infancy, and in this paper we focus on whether the narratives contain useful information that can be used to enhance the accuracy of the numerical forecasts, in a traditional expected squared-error sense. We also look at whether the narratives 'lead' the numerical forecasts. This second aspect is potentially important. For example, when the structure of the economy changes, forecasts often fail, in the sense of being worse than expected based on the model's past performance, see e.g., Clements and Hendry (2006). Hence, advance warning of a break may partially offset the deterioration in forecast performance: see, e.g., Castle et al. (2011). The narratives may hint at the possibility of impending breaks, even though it is felt that such developments are too speculative to be reflected in the numerical forecasts.

Relatedly, the recent literature has suggested that macroeconomic forecasters may be heterogeneous in terms of their loss functions, and particularly in terms of the degree of asymmetry of such loss functions: see, e.g., Capistrán and Timmermann (2009). If this is the case, then it may be reasonable to suppose that over-optimistic and too-pessimistic narrative statements are weighed differently by forecasters compared to the errors they make in their numerical forecasts. In short, forecasters may be more candid in their narratives than their numerical forecasts. Alternatively, the production/reporting of a point forecast - as opposed to a density forecast - may count against providing an 'early warning'.

Finally, the literature on the psychology of judgement under uncertainty (e.g., Kahneman and Tversky's 'heuristics and biases' research, as reviewed in O'Hagan et al. (2006, ch. 3)) suggests that the way in which judgments are elicited may matter. An example of this is that there can be systematic differences between forecasts of central tendencies reported as point predictions as opposed to those implicit in histogram forecasts (see Engelberg et al. (2009) and Clements (2009, 2010)). Hence there is no reason to necessarily expect the narratives and numerical forecasts to mirror each other.

In this paper we study the Bank of England's quarterly forecasts of the two key macrovariables: inflation and output growth, and the associated narratives taken from the Bank of England Quarterly *Inflation Reports*. The narratives accompanying the text are translated into quantitative variables. We consider the *Inflation Reports* from to 1993 to the present, and in so doing contribute to a small but growing literature that converts institutional narratives into quantitative indices.

The plan of the rest of the paper is as follows. In Section 2 we review relevant literature. Section 3 outlines the construction of qualitative indices from the narratives. Section 4 explains the various facets of our forecast-based evaluation of the narratives. Section 5 describes the Inflation Report data, section 6 presents our empirical results, and section 7 offers some concluding remarks.

2 Literature Review

There is a long tradition in economics of converting some types of qualitative data into numerical indices, such as, say, the proportion of survey respondents expecting their sales to increase over the next 6 months, as opposed to staying the same, or falling. See, for example, the review of survey expectations of Pesaran and Weale (2006). However, there is much less use of the wide-ranging and nuanced narratives of the sort that appear in the *Inflation Reports*. A key paper is the investigation by Stekler and Symington (2016) of the 'narratives' that constitute the minutes of the Federal Open Market Committee (FOMC) between 2006 and 2010. Their study develops sentiment 'scores' for current and future trends in the economy, from the minutes, which are then calibrated to current and next-quarter output growth. Ericsson (2016) considers the Stekler and Symington (2016) quantitative indices as forecasts, and evaluates their properties, and focuses on the method of calibrating the scores to the historical data to generate the indices. Romer and Romer (2000) had already made informal use of FOMC narratives to consider the link between the FOMC forecasts, the Staff ('Greenbook') forecasts, and the conduct of monetary policy by the FOMC. They looked at three episodes when the Staff and FOMC forecasts were at odds with each other. Stekler and Symington's is a formal statistical analysis of the FOMC narratives.

Castle et al. (2015) consider forecasts and their accompanying narratives from the standpoint of policy-makers enacting changes in policy, and term the forecasts and narratives *forediction*. Their interest is in what forecast failure implies about the status of the forecasting model, the forediction, and the validity of the policy. Clements and Hendry (2005) showed that forecast failure need not necessarily invalidate the economic theory underlying the forecasting model. However, it is argued that forecast failure generally does entail forediction failure and the invalidity of the policy. Their policy-oriented focus is different from ours, which is squarely on forecasting, but their contribution serves to illustrate the range of questions that can potentially be addressed when narratives are available, as discussed in the introduction.

Fawcett et al. (2015) evaluate the Bank forecasts generated by their COMPASS DSGE model. These forecasts include measures of uncertainty surrounding the central tendencies or point predictions, and so continue the Bank tradition that began with the publication of the 'fan charts'. This allows the possibility in future work of comparing these measures of uncertainty with the narratives.

A fundamental building block of the analysis of narratives is a method for mapping textual data into real numbers. Given the richness of the English language, Di Fatta et al. (2015) argue that standard sentiment indices need to be adapted to the context in which they are being applied, noting that words often have different connotations and meanings in different contexts, and further focusing only on adjectives and adverbs as the more descriptive words in any use of the English language.

If we consider each word to be an observation in our analysis, we are entering into the world of 'big data'; datasets of a size to challenge standard methods of data analysis. Such datasets are often microeconomic in nature, but there are macroeconomic examples too. For example, Meinusch and Tillmann (2015) consider the extent of disagreement on Twitter regarding the timing of monetary policy

Number	Sentence	Score
1	Excluding the boost to growth from the rebound in activity following the heavy snow in 2010 Q4, however, GDP was broadly flat.	1.5
2	Within that, and abstracting from the effects of snow, manufacturing and services output grew moderately, but there was a sharp fall in construction output.	-1.5
3	The extent of spare capacity within businesses is uncertain: the growth rate of companies' effective supply capacity appears to have slowed during the recession, but it is likely that some margin of spare capacity remains.	0
4	Employment has recovered somewhat but unemployment remains elevated.	1.5

Table 1: Example sentences and sentiment scores.

changes in the US. They classify tweets into whether they advocate an early or later tightening of policy, and construct a disagreement index.

3 Constructing the Sentiment Index

We consider two types of forecasts: the standard quantitative forecasts produced by the Bank of England, and qualitative forecasts inferred from the *Inflation Report* narratives. A central part of our work is the construction of suitable quantitative indices from qualitative information. We require a series, QI_t, which captures the information in the narrative in each *Report*, for all the reports in our sample. Such a measure transforms thousands of words into a single number. This task is particularly troublesome because of the inability to verify its accuracy, since there is no "true" value for the level of sentiment expressed in a given set of text.

Perhaps the simplest method of creating a sentiment index maps words into positive (+1), negative (-1) or neutral (0) scores, and evaluates the resulting sequences. We calculated a Finn index in this way using a dictionary provided by Nielsen (2011), and took the average sentiment per word in a report.

Alba (2012) develop the sentiment index that we use. This attempts to better capture the nuances and subtleties of the language that is used in the narratives. It makes use of the *Natural Language Toolkit* (Bird et al., 2009) in the *Python* programming language. First we manually rank the 3000 most common words in historical *Inflation Reports*, restricting our attention to adverbs (1000), adjectives (1000) and nouns (1000). We score words that express positive sentiment with a 1, and words that express a negative sentiment with a -1, leaving other words with a score of 0. We categorize words into three additional classes: decrementers, incrementers, and inverters. Decrementers decrease the score attached to a subsequent word (e.g. "slightly bad"), while incrementers increase the score (e.g. "really bad"), and inverters invert the sign of the score attached to a word (e.g. "not bad"). Decrementers halve the score attached to a word, while incrementers double it.

We subsequently sum the scores attached to words in a report to provide a sentiment score for that report. In order that sentiment scores are not biased by the size of a report (reports have nearly doubled in size over the 23 years of their existence), we divide the sum by the number of sentences in a report.

In Table 1 we provide some examples of sentences from *Inflation Reports* and the sentiment score attached to them.

Sentence 1 scores 1.5 because "boost" and "growth" both score positively, whereas flat is classed to be negative, but "broadly" decrements this to a half. Sentence 2 scores positively because of the term "grew", which is decremented by "moderately", but negatively because of the "fall", which was "sharp".

Sentence 3 scores zero and shows the difficulty in evaluating sentences, because "growth" attracts a positive score, as does "effective", even though the latter is only part of the term "effective supply". The words "slowed" and "recession" score negatively and hence balance out the first two words. Finally, sentence 4, like sentence two, has "recovered" scoring positively, although "somewhat" is a decrementer, while "elevated" scores positively. Again this shows the difficulty of scoring text, since elevated when attached to the term unemployment should ideally be inverted as a score.

In addition to quantifying the sentiment of entire *Reports*, we also focus only on sentences containing particular keywords. For example, we might consider only sentences that mention the "committee" in order to focus more on the opinions recorded as having been expressed by the Monetary Policy Committee; equally, we might consider only sentences with the word "demand" in, in order to consider the demand side of the economy.

We note that a number of sentences in *Inflation Reports* are included for information purposes only, such as the brief of the Monetary Policy Committee at the start of the report, and the text beneath many of the forecast plots that describes the construction of the plot. Such sentences are factual and hence do not contain adjectives and adverbs, and as such any sentiment score they are attributed will be zero or very close to zero. Furthermore, these text blurbs are common to all *Inflation Reports*, and hence can be thought of only as having a nominal effect, rather than a real effect, on sentiment between reports.

Finally, if we suppose that there were a 'true' QI, then our imperfectly-framed index could be regarded as measuring that series with error. Under standard assumptions, the use of the measure subject to error would attenuate the estimated effect relative to the true effect. Hence we are unlikely to find an effect using our index if such an effect does not exist: if we find the narratives contain relevant information using our QI we can be reasonably confident that the finding is not spurious.

4 Forecast Evaluation

4.1 Forecast Accuracy

We denote the variable of interest at time t as y_t, and the forecast of the variable at time $t+h$, made using information at time t, as $\widehat{y}_{t+h|t}$. Each *Report* gives a forecast of the current quarter (the quarter in which the report is issued) and for each of the next 7 quarters. We adopt the convention throughout that $h = 1$ refers to a current quarter forecast. We define the forecast error as $\widehat{e}_{t+h|t} = y_{t+h} - \widehat{y}_{t+h|t}$, and use squared error loss to determine forecast accuracy, with empirical counterpart given by the mean squared forecast error (MSFE). For T forecasts of length h, this is simply $T^{-1}\sum_{t=1}^{T}\widehat{e}^2_{t+h|t}$. We will sometimes use the Root Mean Squared Forecast Error (RMSFE), the square root of the MSFE. For the quantitative forecasts at least a natural way of assessing accuracy is in terms of the RMSFE relative to a set of benchmark forecasts. For inflation forecasting 'no-change' predictors often provide stiff competition (see, e.g., Atkeson and Ohanian (2001), Stock and Watson (1999)), whereas for output growth autoregressive models (which do not impose a unit root in the growth rate) are preferred. Formal tests of predictive ability can be made using approaches such as those popularized by Diebold and Mariano (1995) (see, e.g., Clark and McCracken (2011) and Giacomini (2011) for recent reviews) to see whether differences in RMSFE reflect statistically significant differences between the forecasts. We employ a regression-based

Diebold Mariano test (Pretis, 2015) where we construct the term e_t^{DM} as:

$$e_t^{DM} = L\left(y_t - \widehat{y}_{t|t-h}\right) - L\left(y_t - \widehat{y}_{t|t-h}^{RW}\right).$$

Conventionally, the loss function L is often taken to be squared error loss, i.e., $L(e) = e^2$. Equal forecast performance implies $E(e_t^{DM}) = 0$, and this test can be implemented using the regression model:

$$e_t^{DM} = \alpha + u_t \tag{1}$$

with the null of equal accuracy, $H_0 : \alpha = 0$. The significance of α implies difference in forecast performance. If $\alpha > 0$ this implies that the benchmark forecast is 'better' than the Bank forecast, while $\alpha < 0$ implies the opposite, namely that the Bank forecast is 'better' than the benchmark.

In principle, the qualitative index (QI) could be assessed for accuracy in a similar way, and it could be formally compared to the numerical forecasts ($\hat{y}$). Two complications arise in assessing the accuracy of QI. Firstly, although the forecast origin is well defined, as the survey in which the narrative appeared, the period to which the narrative relates is typically not clear. Secondly, QI can only be directly related to the macro-variable y when it is appropriately scaled, for example, to have the same mean and variance, typically over a relatively short window of data immediately prior to the forecast origin. (Using a short window of data allows for some non-constancy in the relationship between our sentiment index and inflation).

We scale our QI measure to match the range of the variable over the sample immediately prior to the forecast origin. Letting QI_{t-h} denote the index from the *Report* at time $t-h$, the scaled index, denoted q_{t-h}, is given by:

$$q_{t-h} = \frac{(b_{t-h} - a_{t-h})(QI_{t-h} - \min_{\tau=t-h-8\ldots t-h-1}\{QI_\tau\})}{\max_{\tau=t-h-8\ldots t-h-1}\{QI_\tau\} - \min_{\tau=t-h-8\ldots t-h-1}\{QI_\tau\}} + a_{t-h},$$

where $a_{t-h} = \min\{y_\tau\}_{\tau=t-h-8,\ldots,t-h-1}$ and $b = \max\{y_\tau\}_{\tau=t-h-8,\ldots,t-h-1}$. This gives rise to a scaled series $\{q_t\}$, where the QI_t from each *Report* has been separately scaled. We always use q_{t-h} in the following. Note it is a legitimate choice as a 'real-time' forecast in that QI_{t-h} is available at time $t-h$, and the construction of q_{t-h} from it only makes use of information known at time $t-h$. The single subscript on q denotes the forecast origin (i.e., the date of the *Report*) and the target date is deliberately left unspecified given the nature of the forecast.

There are a number of other dimensions beyond forecast accuracy in which it will prove useful to compare the forecasts, including the extent to which the two sets of forecasts contain complementary information on future movements in the macro-variable.

4.2 Forecast Efficiency

A popular test of forecast efficiency (or optimality) is that of Mincer and Zarnowitz (1969), with recent extensions due to Patton and Timmermann (2012). The Mincer and Zarnowitz (1969) (MZ) regression tests forecast optimality at a given horizon. The regression is:

$$y_t = \alpha + \beta\widehat{y}_{t|t-h} + u_t, \tag{2}$$

where the observations range over t for a given h, and the null of optimality is that $\alpha = 0$ and $\beta = 1$. By writing:

$$y_t - \widehat{y}_{t|t-h} = \alpha + (\beta - 1)\,\widehat{y}_{t|t-h} + u_t \tag{3}$$

we obtain $Cov\left(y_{T+h} - \widehat{y}_{T+h|T}, \hat{y}_{T+h|T}\right) = (\beta - 1)\,Var\left(\widehat{y}_{T+h|T}\right)$, so that the covariance is zero iff $\beta = 1$. Hence it is a test of whether the forecast efficiently uses all the information available at the forecast origin, such that the resulting forecast error is not systematically related to the forecast origin information, as filtered via the forecast. The condition $\alpha = 0$ and $\beta = 1$ is sufficient for unbiasedness (but not necessary: see Holden and Peel (1990)).

Autocorrelation consistent (AC) standard errors are used for multi-step forecasts ($h > 1$) to account for the overlapping forecasts phenomenon.

4.3 Forecast Encompassing

The MZ regression (2) can be supplemented with additional variables known at $t - h$, e.g.,

$$y_t = \alpha + \beta \widehat{y}_{t|t-h} + \kappa' z_{t-h} + u_t, \tag{4}$$

where (as shown) z_{t-h} may comprise a vector of such variables. The null is now that $\alpha = 0$, $\beta = 1$ and $\kappa = 0$. Of course, the additional variable(s) could include the q index, so for example:

$$y_t = \alpha + \beta \widehat{y}_{t|t-h} + \gamma q_{t-h} + u_t \tag{5}$$

which takes the form of the test for forecast encompassing between forecasts $\widehat{y}_{t|t-h}$ and q_{t-h} suggested by Fair and Shiller (1990). Tests of forecast encompassing (see Chong and Hendry (1986)) assess whether *ex post* a linear combination of forecasts results in a statistically significant reduction in the mean squared forecast error (MSFE) relative to using a particular forecast. Forecast encompassing is formally equivalent to 'conditional efficiency' due to Nelson (1972) and Granger and Newbold (1973), whereby a forecast is conditionally efficient if the variance of the forecast error from a combination of that forecast and a rival forecast is not significantly less than that of the original forecast alone. The null that $\widehat{y}_{t|t-h}$ encompasses q_{t-h} is given by $\gamma = 0$ against the (usually) one-sided alternative that $\gamma > 0$. Alternative forms of forecast encompassing tests include:

$$y_t - \widehat{y}_{t|t-h} = \alpha + \beta q_{t-h} + u_t$$

where the null is as before, and the interpretation is whether q_{t-h} can help explain the forecast errors associated with $\widehat{y}_{t|t-h}$. Forecast encompassing of one model, say $M2$ by another, say $M1$, is a more stringent requirement than simply that $M1$ is more accurate than $M2$ on RMSE (see, e.g., Ericsson (1992)). $M2$ might still contain useful information not in $M1$, raising the possibility that a simple linear combination of the two might produce superior forecasts in the future.

4.4 Forecast Updating

Patton and Timmermann (2012) show that the actual value of y_t can be replaced by a short-horizon forecast, say, $\widehat{y}_{t|t-h_1}$, under certain circumstances, to give:

$$\widehat{y}_{t|t-h_1} = \alpha + \beta \widehat{y}_{t|t-h_2} + u_t \tag{6}$$

where $h_2 > h_1$. This may be useful when the vintage of the actual being targeted is unknown. Although CPI/RPI inflation data are not subject to large revisions, national accounts variables such as real output growth are. Replacing the actual by a short-horizon forecast requires that the latter is an efficient forecast of the actual values. Otherwise the nature of the test changes, and may have no power to detect misspecification, a situation described by Nordhaus (1987), p.673 as '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.'

In addition to the tests of forecast encompassing, we consider whether forecasts are efficiently updated, in the sense that forecast revisions are unpredictable from information available at the forecast origin. In terms of (6), we introduce $(q_{t-h_2} - q_{t-h_3})$ as an additional explanatory variable, and set $\beta = 1$ to give the test regression:

$$\widehat{y}_{t|t-h_1} - \widehat{y}_{t|t-h_2} = \alpha + \gamma \left(q_{t-h_2} - q_{t-h_3} \right) + u_t. \tag{7}$$

This is a test of whether $q_{t-h_2} - q_{t-h_3}$ 'leads' $\widehat{y}_{t|t-h_1}$, in the sense that past values of q predict subsequent values of $\hat{y}$. Formally, the null is that the revision between the forecasts of y_t made at time h_1 and h_2 should be unpredictable at the time the longer horizon (h_2) forecast is made (called a 'strong-efficiency' test by Nordhaus (1987)) against the alternative that the change in the value of q between $t - h_2$ and $t - h_3$ has predictive power.

Finally, when $h_2 = h_1 + 1$, so that the two forecasts are adjacent, and $h_3 = h_2 + 1$, then there is no need to use AC standard errors for inference — it is straightforward to show that the forecast revision for a target t will be uncorrelated with the forecast revision for all other targets, t_1, for $t_1 \neq t$, for optimal forecasts.

5 Bank Forecast Data

The Bank has produced forecasts of inflation for up to between seven and thirteen quarters ahead since 1993, from the current quarter at the time of publication. In August 1997 the Bank began also forecasting GDP growth in each *Report*, again for eight quarters ahead. Each *Report* contains a number of additional forecasts that the Bank produces, but the MPC signs off on these two forecasts. It also signs off on the unemployment rate forecasts, but only from 2013, so we do not consider these forecasts.

At the time of writing, since the first report in February 1993, there have been 95 such reports up to August 2016. Mean, median and mode inflation forecasts are provided, along with skew and uncertainty. Since 1998 it has provided two types of forecast for inflation: forecasts holding interest rates fixed at their current level, and forecasts assuming interests rates follow the "market expectation of interest rates" in the forecast period. This amounts to around 5,000 individual forecasts of inflation and GDP, many of which have been revised in subsequent quarterly reports. For inflation the forecasts were for RPIX up to the end of 2003, and subsequently are of CPI.

For inflation, the six groupings (mean, median and mode for market interest rates and fixed interest rate) show a very high level of correlation. The mean, median and mode forecasts have correlation coefficients of 0.99 or higher (both within market and fixed interest rate forecasts), while the correlation between market and fixed interest rate forecasts is never lower than 0.96. As such, we focus our analysis on mean forecasts in this paper.

Some of the key properties of the forecasts are evident from Figures 1 and 2. Figure 1 shows (mean)

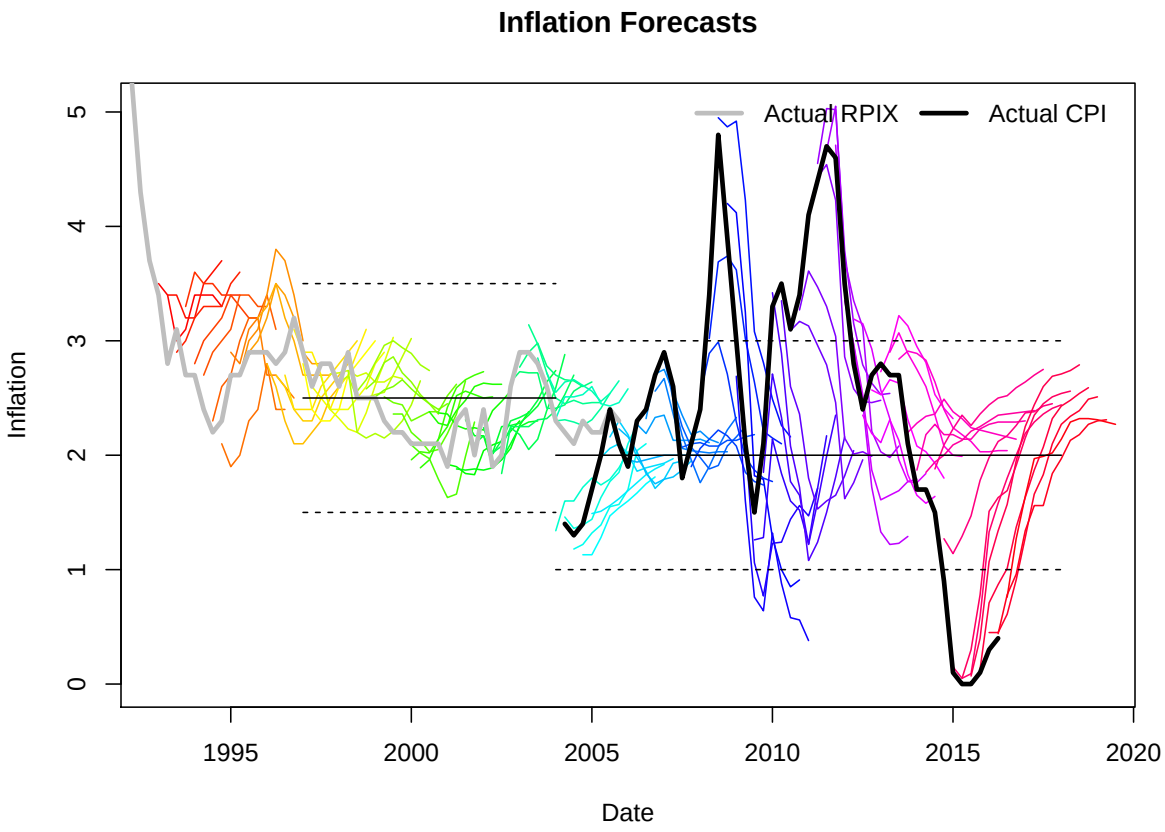

Figure 1: Forecasts and actual inflation since 1993 (forecasts to 2018).

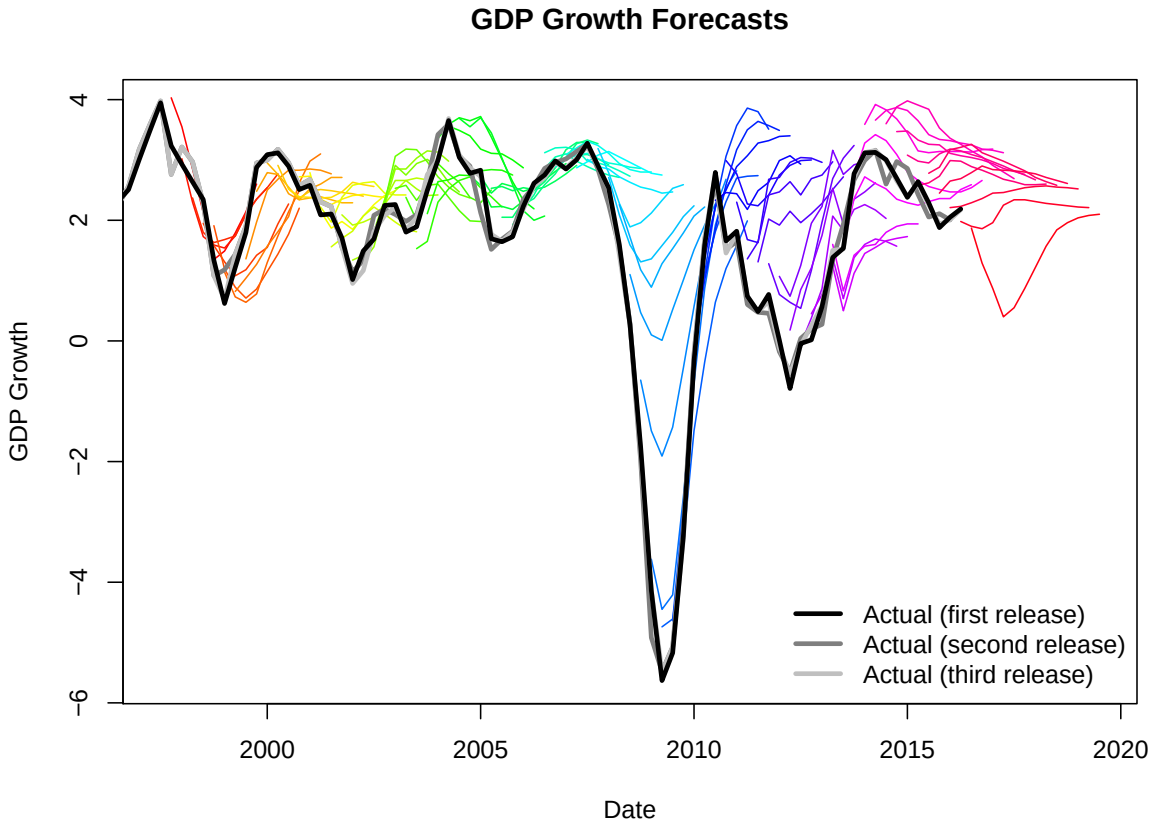

Figure 2: Forecasts and actual GDP growth since 1998 (forecasts to 2018).

forecasts since 1993 alongside outcomes for inflation. Up to 2003 the Bank was instructed to target RPIX inflation at 2.5%, but after that CPI inflation at 2%. These are marked, along with the dotted lines marking the upper and lower bounds. In both periods there is clear evidence that forecasts move towards the target values irrespective of the forecast origin values of inflation. As evident from the figure, the period up to 2003 was relatively quiescent in terms of inflation outcomes, but after this, and in particular after 2006, inflation became much more variable (and presumably less predictable). Figure 2 also suggests forecasts of output growth had an attractor, of around 2%.

The accompanying narratives have varied in size, as evident from the word counts recorded in Figure 3. Aside from the spike around the turn of the century (when MPC minutes were attached to *Inflation Reports*), the word count has generally trended upwards to its current level of around 35,000 words, from nearer to 20,000 words in 1993.

We construct two sentiment indices using the methods outlined in Section 3. We plot the resulting indices in Figure 4; the Finn index uses the dictionary provided by Nielsen (2011) and scores words, whereas the Alba index builds on words used in *Inflation Reports* classified as positive, negative, decrementers, incrementers and inverters, using the method described in section 3.

Both indices display the same broad patterns, with increasing sentiment throughout the 1990s, with record highs in the period 2000-2008 before a dramatic fall in line with the financial crisis (although

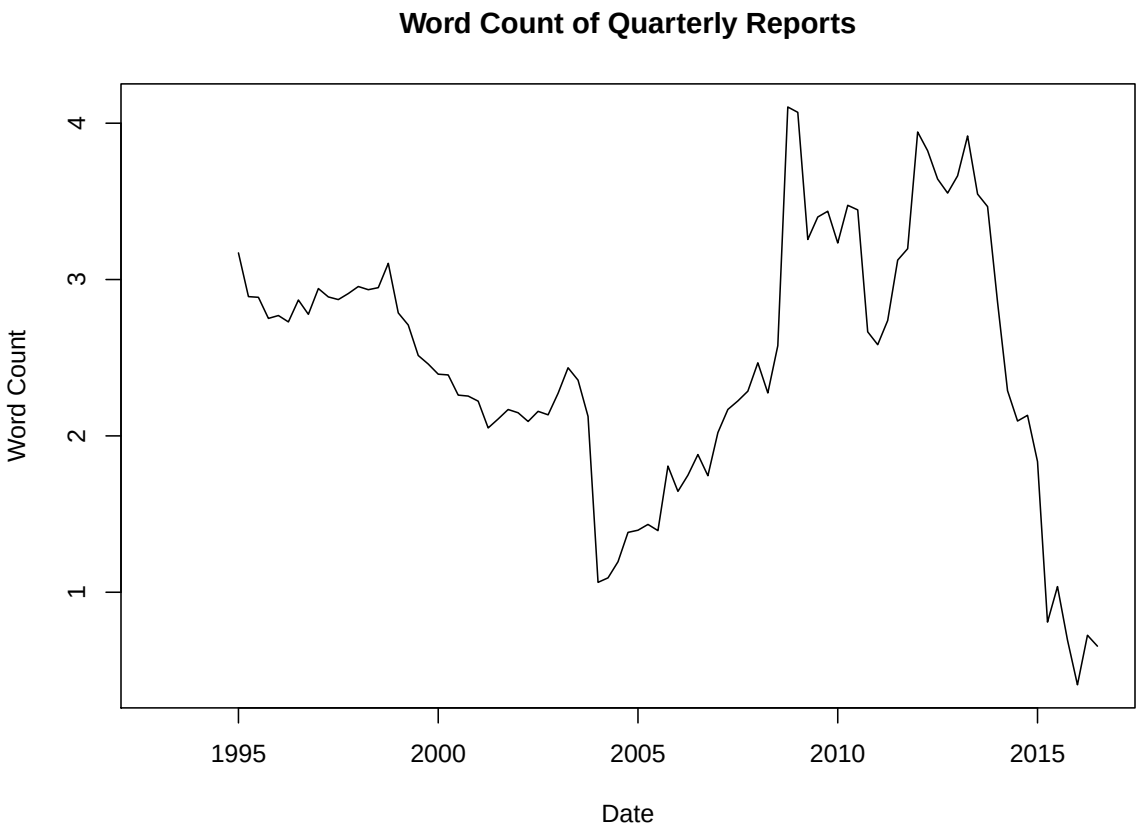

Figure 3: The word count of *Inflation Reports* since 1993.

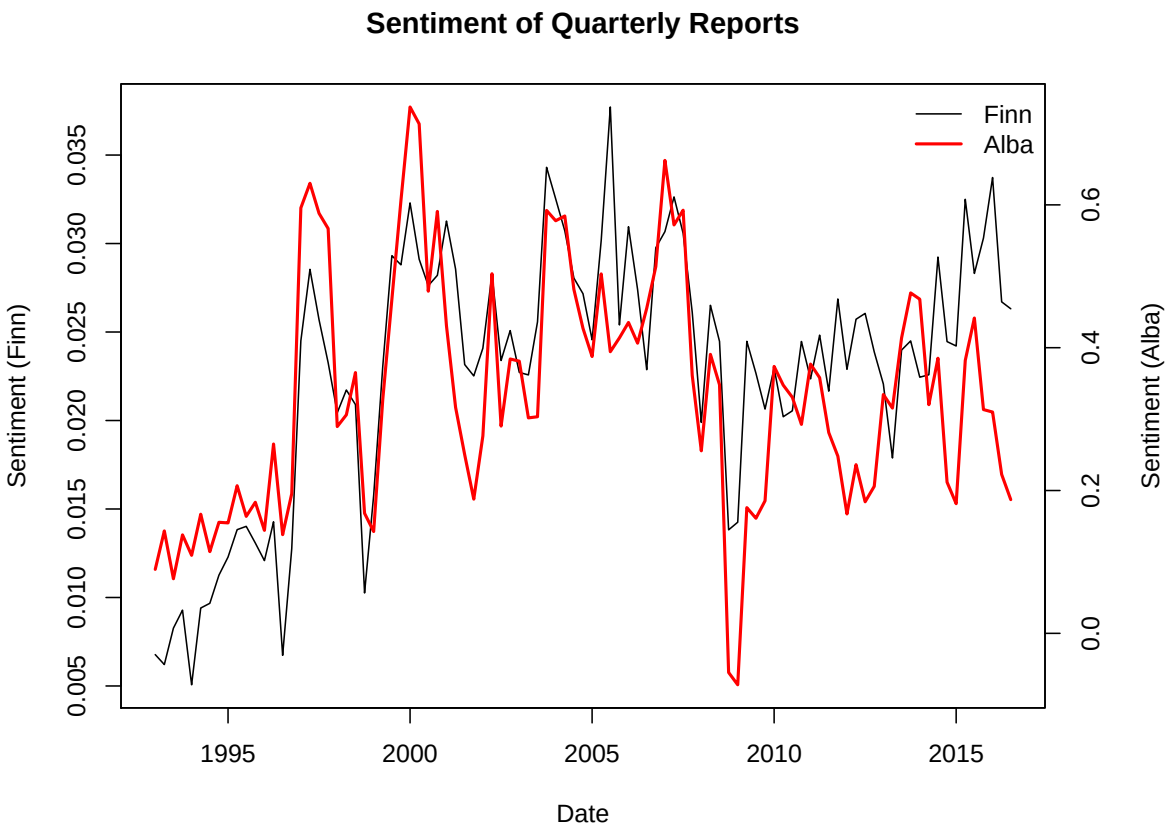

Figure 4: The two sentiment measures, plotted in raw, sentiment-per-sentence, form.

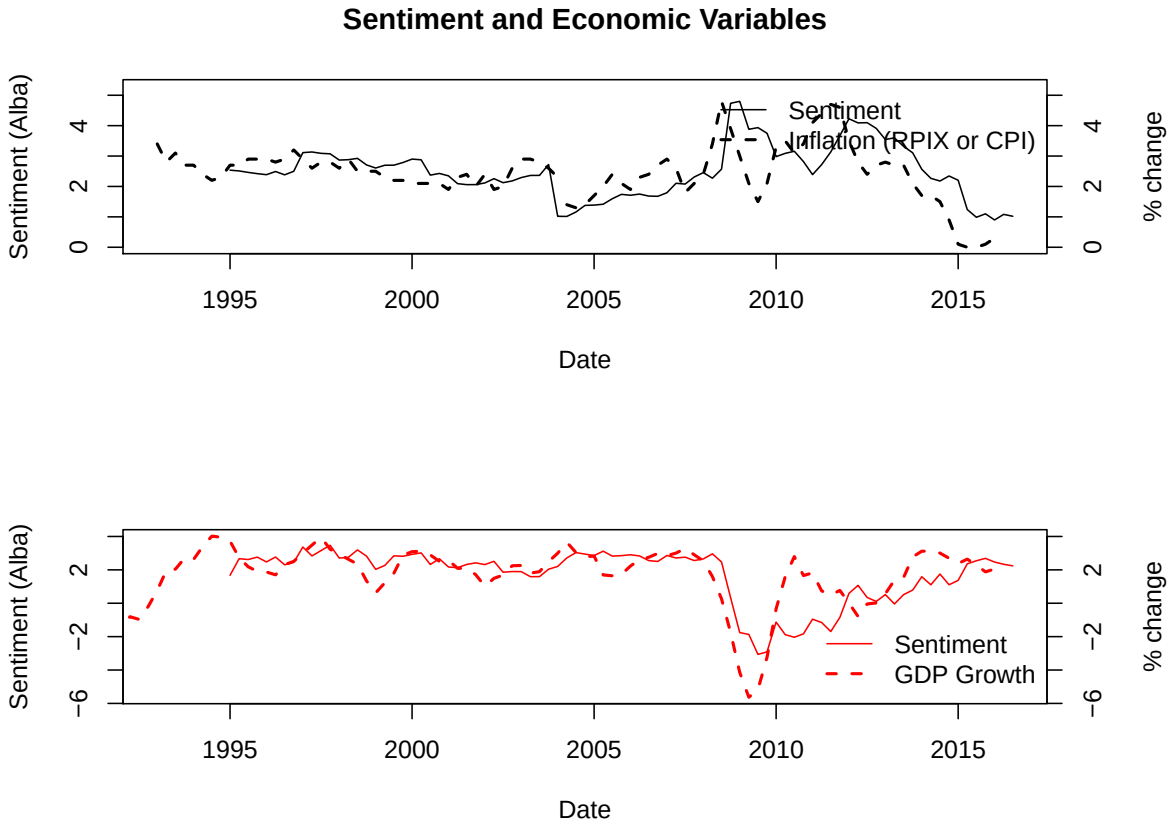

Figure 5: Sentiment, scaled to match the range of inflation (top panel), and GDP growth (bottom panel) in the most recent two years prior to each *Inflation Report*.

slightly ahead of the GDP fall), and subsequent recovery. Our subsequent analysis is based on the Alba index.

As described earlier, it is important to scale these quantitative indices in order that they can be useful as forecasts. We use the same indices for inflation and output growth, and scale them separately by recent inflation or output growth depending on the variable being forecast, as described in section 4. The resulting plots are in Figure 5, and the impact of scaling is clear, with the resulting series much more closely following the trajectories of each data series throughout the sample period.

6 Results

6.1 Forecast Accuracy

We report Diebold-Mariano regressions (specified in (1)) of the Bank's forecasts against standard time-series models, and against forecasts derived from the narratives - the sentiment indices. We might expect the Bank's forecasts to outperform the simple time-series forecasts at short horizons, but to struggle as the horizon increases, and this is broadly what we find (see Clements (2015) for a comparison of survey

expectations and time-series forecasts). For output we use an AR(1) model for the growth rate, and for inflation a 'no-change' in growth rate forecast, because the inflation rate is close to having a unit root. For output growth the model forecast will approach the unconditional mean of the growth rate for the period for which the model was estimated. So the comparison of the Bank's forecasts to the model forecasts can be interpreted as establishing the horizon at which the Bank is unable to outperform simply saying the future growth is the sample average. In Table 2 a negative (positive) coefficient indicates the Bank forecast performs better (worse) than the benchmark, and is annotated to depict statistical significance. Hence the Bank forecasts are no better than the model forecasts at around 5 and 4 quarters ahead for output growth and inflation respectively.[1]

We also compare the Bank forecasts against the sentiment index forecasts. Unsurprisingly, the Bank's numerical forecasts are more accurate than those based of the sentiment indices in this head-to-head, with large differences, although the statistical significance of these differences only holds at the shorter horizons. More pertinently, we consider below whether the sentiment indices nevertheless convey useful information about the future evolution of these variables.

6.2 Forecast Efficiency

Before considering whether the possible incremental value of the sentiment indices, we consider whether the Bank forecasts are efficient in the Mincer-Zarnowitz sense. The Mincer-Zarnowitz regressions are defined in (3), and we report outcomes for inflation in Table 3, and for GDP growth in Table 4. Because the regression equation is parameterized with the forecast error as the dependent variable, both the constant and slope ought to be zero. The null is rejected at the 10% level for inflation at $h = 1$, and at more stringent levels of significance at longer horizons, and for output growth at all horizons.

6.3 Forecast Encompassing Tests

Although the head-to head comparisons of the Bank and sentiment forecasts (q) favour the Bank's numerical forecasts, the q forecasts may nevertheless contain useful additional information not present in the numerical forecasts. We run forecast encompassing tests to see whether the numerical forecasts encompass the forecasts from the scaled quantitative index: see Table 5 for inflation, and tables 6 and 7 for output growth, where the actuals are taken to be the initial release and the second revision, respectively. The tests are based on the equivalent of (5) - we use as the dependent variable the forecast error for the numerical forecasts, and include the numerical forecast as an explanatory variable. We present each forecast horizon in a separate column, with robust standard errors in parentheses. For inflation, there is little evidence against the null hypothesis that q is forecast-encompassed: we do not reject the null hypotheses that the parameter on q is zero.

For output growth, q is significant at $h = 1$, with a coefficient of around 0.10 (using second-revision actuals), suggesting that at least in–sample, a combination of $\hat{y}$ and q provides a statistically more accurate forecast than $\hat{y}$ alone. The choice of initial release or second revision actuals for output growth makes little difference. For our measures of inflation, revisions are inconsequential. [2]

We also evaluated the sentiment of only sentences in which particular economic terms are mentioned; in doing so we created a number of sub-indices. All of these are plotted, unscaled, in Figure 6, in

[1] More accurate forecasting models could undoubtedly be found. Models which make use of related series, and higher-frequency data, as in, e.g., Ghysels et al. (2007) and Clements and Galvão (2008) might give better forecasts, but this is not our focus.

[2] Croushore (2011) provides a recent survey of forecasting and evaluating forecasts when data are subject to revision.

	Inflation. Bank forecasts versus 'no-change' forecasts.							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	-0.142^{***}	-0.273^{**}	-0.268^{*}	-0.087	0.188	0.341	0.250	-0.028
	(0.053)	(0.111)	(0.142)	(0.177)	(0.224)	(0.306)	(0.347)	(0.347)
Observations	95	94	93	92	91	90	89	86
	GDP growth. Bank forecasts versus an autoregressive model.							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	-1.430^{**}	-2.102^{**}	-2.205^{**}	-1.650^{*}	-0.689	0.111	0.674^{**}	0.993^{***}
	(0.575)	(0.894)	(1.002)	(0.901)	(0.606)	(0.417)	(0.331)	(0.301)
Observations	75	74	73	72	71	70	69	68
	Inflation. Bank forecasts versus sentiment forecasts.							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	-0.732^{***}	-0.799^{***}	-0.668^{***}	-0.395^{*}	-0.105	0.023	-0.093	-0.307
	(0.215)	(0.249)	(0.223)	(0.211)	(0.321)	(0.516)	(0.472)	(0.411)
Observations	87	86	85	84	83	82	81	80
	GDP growth. Bank forecasts versus sentiment forecasts.							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	-1.696^{**}	-2.135^{*}	-2.502	-2.545	-2.341	-1.953	-1.428	-0.973
	(0.845)	(1.151)	(1.512)	(1.749)	(1.602)	(1.372)	(1.148)	(0.630)
Observations	75	74	73	72	71	70	69	68
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 2: Diebold-Mariano Test Regressions of Equal Forecast Accuracy Between the Bank Forecasts, and Benchmark Forecasts or Sentiment Indices.

red, with the unscaled Alba index in black on the same plots. In total we considered twenty economic terms to create sub-indices, and each has its own plot in Figure 6. The term "disinflation" is mentioned essentially never throughout the history of Inflation Reports, and hence there is no sub-index, and for "deflation" and "committee" there are missing sections (the latter corresponding to the pre-Monetary Policy Committee years of 1993–1997). The remaining sub-indices show varying degrees of correlation with the general index; perhaps of note is the apparent stronger correlation between the "demand" sub-index and the general index (0.76) relative to the "inflation" sub-index (0.63). Similarly, there is a higher degree of correlation between overall sentiment and sentiment in "business" sentences (0.71) than there is in "household" sentences (0.49). Sentences mentioning "GDP" and anything related to the exchange rate are similarly highly correlated with overall sentiment, while the lowest correlation is between the "unemployment" sub-index and overall sentiment, which is essentially zero, and for "productivity" (0.34) and "labour" (0.31).

In the regressions with the sentiment sub-indices, the dependent variable is again the forecast error (calculated using the usual numerical forecasts), and this is regressed on the sentiment sub-indices for

	Forecast Error (quarters ahead)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.054	0.229*	0.495**	1.171***	2.228***	2.886***	2.988***	2.916***
	(0.054)	(0.132)	(0.218)	(0.322)	(0.399)	(0.405)	(0.408)	(0.470)
Forecast	−0.031	−0.109**	−0.210**	−0.475***	−0.913***	−1.198***	−1.241***	−1.211***
	(0.021)	(0.051)	(0.085)	(0.130)	(0.166)	(0.172)	(0.174)	(0.200)
Observations	95	94	93	92	91	90	89	86
R^2	0.023	0.048	0.063	0.130	0.254	0.356	0.369	0.304
Test of Efficiency	2.666*	3.069**	6.791***	15.677***	25.456***	26.857***	19.241***	
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 3: Mincer-Zarnowitz regressions for inflation.

	Forecast Error (quarters ahead)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	−0.324***	−0.475***	−0.627***	−0.757**	−0.624	0.010	1.349	2.755**
	(0.073)	(0.124)	(0.205)	(0.337)	(0.535)	(0.806)	(1.078)	(1.271)
Forecast	0.046	0.076	0.084	0.061	−0.054	−0.361	−0.911**	−1.454***
	(0.029)	(0.050)	(0.083)	(0.135)	(0.216)	(0.319)	(0.417)	(0.483)
Observations	75	74	73	72	71	70	69	68
R^2	0.033	0.031	0.014	0.003	0.001	0.018	0.066	0.121
Test of Efficiency	13.125***	9.982***	8.106***	7.723***	7.338***	8.531***	11.163***	14.151***
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 4: Mincer-Zarnowitz regressions for GDP growth (first release).

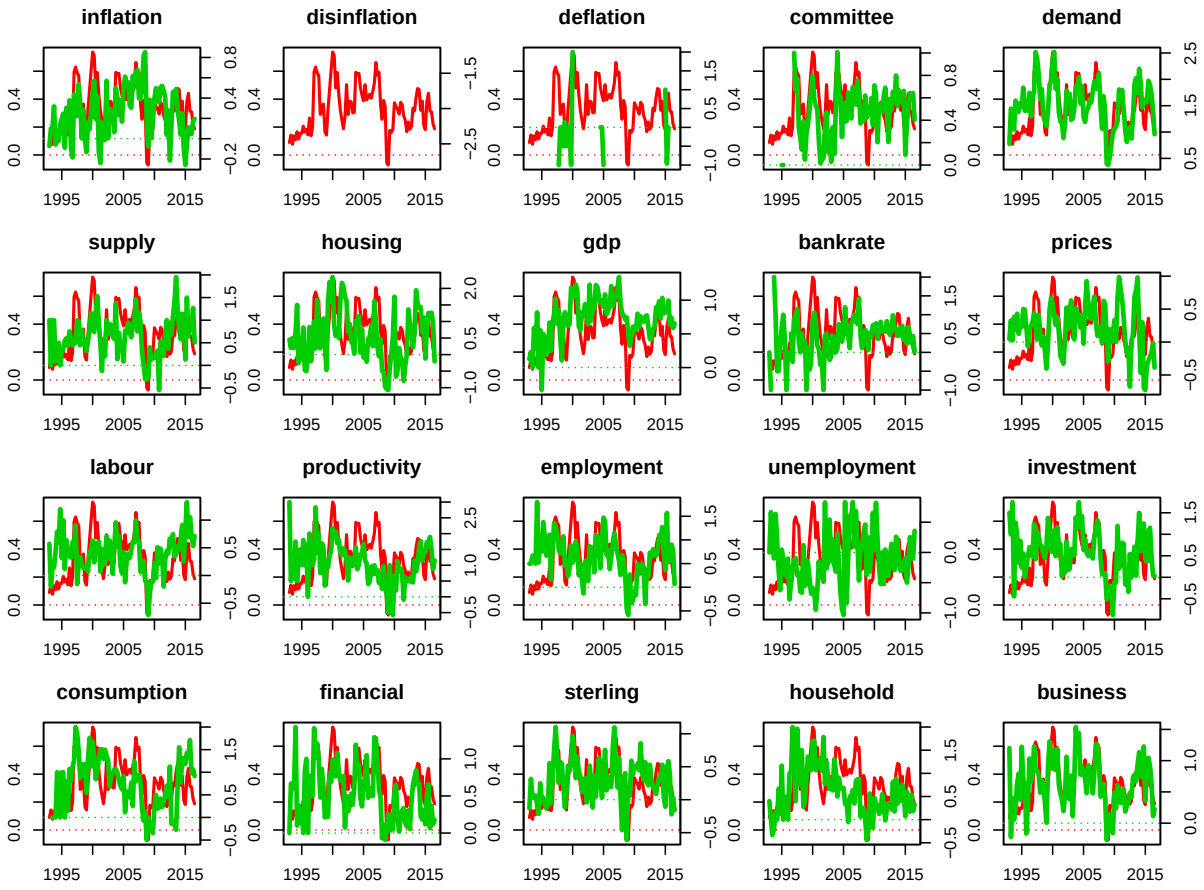

Figure 6: The sentiment sub-indices (red), plotted against the raw Alba index (black).

Table 5: Encompassing regression for each forecast horizon for inflation forecasts, regressing forecast errors on scaled sentiment forecasts. Robust standard errors in parentheses.

	Forecast Error (quarters ahead)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.013	0.087	0.158	0.562	1.665***	2.724***	3.460***	3.594***
	(0.060)	(0.163)	(0.280)	(0.425)	(0.554)	(0.635)	(0.706)	(0.777)
Forecast	−0.023	−0.090	−0.204**	−0.447***	−0.944***	−1.333***	−1.483***	−1.434***
	(0.023)	(0.056)	(0.092)	(0.142)	(0.200)	(0.219)	(0.235)	(0.253)
Sentiment	0.014	0.048	0.136	0.220*	0.242*	0.162	0.002	−0.084
	(0.026)	(0.060)	(0.090)	(0.115)	(0.128)	(0.131)	(0.135)	(0.140)
								Observations
Observations	87	86	85	84	83	82	81	80
R^2	0.012	0.030	0.066	0.132	0.239	0.346	0.358	0.307
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 6: Encompassing regression for each forecast horizon for GDP growth (initial release) forecasts, regressing forecast errors on scaled sentiment forecasts. Robust standard errors in parentheses.

	Forecast Error (quarters ahead)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	−0.353***	−0.500***	−0.638***	−0.767**	−0.644	−0.053	1.179	2.509*
	(0.072)	(0.125)	(0.209)	(0.343)	(0.545)	(0.829)	(1.140)	(1.362)
Forecast	−0.010	0.037	0.071	0.052	−0.064	−0.364	−0.882**	−1.400***
	(0.037)	(0.060)	(0.093)	(0.144)	(0.221)	(0.322)	(0.424)	(0.497)
Sentiment	0.087**	0.066	0.025	0.021	0.029	0.048	0.066	0.072
	(0.038)	(0.056)	(0.079)	(0.101)	(0.119)	(0.129)	(0.135)	(0.137)
								Observations
Observations	75	74	73	72	71	70	69	68
R^2	0.098	0.049	0.016	0.004	0.002	0.020	0.070	0.124
Note:								*p<0.1; **p<0.05; ***p<0.01

each forecast horizon, as well as the numerical forecast. We carried out 'general-to-specific' model selection on the sub-indices and the numerical forecast using the Pretis et al. (2014) R package. We present the resulting final models for each quarter ahead in Table 8. As outlined in Section 4, any significant coefficients are indicative of sub-optimal forecasts: information available at the forecast origin, reflected in the *Inflation Report* narratives, is not included in the usual numerical forecasts. For inflation, as might have been anticipated from the forecast efficiency test results, the numerical forecast is retained at all forecast horizons, whereas for output growth the numerical forecast is retained at just one horizon. Hence for output growth the reported results have the simple interpretation of whether the sub-indices have explanatory power for the forecast errors.

At each horizon, different sub-indices are significant and hence retained in the model selection process. We clearly find that some sub-indices are significant: at some horizons some sub-indices do contain useful

Table 7: Encompassing regression for each forecast horizon for GDP growth (second revision) forecasts, regressing forecast errors on scaled sentiment forecasts. Robust standard errors in parentheses.

	Forecast Error (quarters ahead)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	−0.360***	−0.495***	−0.602**	−0.725**	−0.582	−0.003	1.192	2.555*
	(0.088)	(0.142)	(0.229)	(0.360)	(0.564)	(0.852)	(1.169)	(1.410)
Forecast	−0.017	0.023	0.046	0.032	−0.097	−0.397	−0.900**	−1.430***
	(0.045)	(0.068)	(0.102)	(0.152)	(0.229)	(0.332)	(0.436)	(0.517)
Sentiment	0.102**	0.088	0.044	0.031	0.044	0.074	0.085	0.085
	(0.046)	(0.064)	(0.086)	(0.106)	(0.123)	(0.133)	(0.138)	(0.140)
								Observations
Observations	74	73	72	71	70	69	68	67
R^2	0.088	0.050	0.012	0.003	0.004	0.025	0.073	0.125

Note: *p<0.1; **p<0.05; ***p<0.01

information not already embodied in the numerical forecasts. For example, for inflation, the negative estimated coefficient on the employment sentiment indicator suggests under-predictions of inflation are correlated with low employment sentiment. For output growth, financial sentiment is clearly significant and has a positive effect: underpredictions of output growth (positive forecast errors) are associated with higher financial sentiment.

Our results are 'in-sample', and the true significance of some of the explanatory variables may be inflated by their retention in the model selection step. A more demanding assessment of whether the sentiment sub-indices can be used to systematically improve upon the published numeric forecasts would require an out-of-sample evaluation, and the division of the data into a within-sample or training period, and an out-of-sample period, on which the in-sample estimated forecast combinations could be evaluated. We leave this for future research.

6.4 Forecast Updating

Turning to adjustments to forecasts, we consider updates to forecasts between adjacent *Inflation Reports*. A requirement of an efficient forecast - in the sense of making use of all the available information - is that subsequent revisions to the forecasts of the same target (here, y_t) should not be predictable using information available at the time of the original forecast. A key focus is whether sentiment, as expressed in the narratives accompanying the forecasts, 'lead' the numeric forecasts. This gives rise to the regression given in equation (7), with the (lagged) change in the sentiment index as the sole regressor. More general specifications could be allowed, and we could consider the persistence in the numerical forecast updates themselves, but we do not do so here.

Tables 10 and 11 report the results. There is clear evidence that changes in the (scaled) sentiment index lead changes in the Bank's numeric forecasts for both output growth and inflation. For output growth, the coefficients on the change in sentiments are significant at the 5% level at horizons up to one-year ahead, and sentiment has a positive effect. That is, an increase in sentiment between *Reports* predicts a subsequent upward revision to the forecast of output growth in the next *Report*.

To illustrate and clarify the timings of these effects, consider the first column regressions results

Table 8: Forecast encompassing test output employing model selection on sentiment sub-indices for inflation.

	Dependent variable:							
	Forecast Error (quarters ahead)							
Quarters Ahead	1	2	3	4	5	6	7	8
Cconst.	0.121	0.995***	1.539***	1.687***	2.557***	2.571***	3.648***	3.962***
	(0.075)	(0.198)	(0.343)	(0.318)	(0.364)	(0.396)	(0.503)	(0.504)
forc	−0.043**	−0.197***	−0.270***	−0.569***	−0.889***	−0.904***	−0.954***	−1.112***
	(0.021)	(0.047)	(0.082)	(0.131)	(0.159)	(0.170)	(0.240)	(0.248)
general.sent		2.833***	2.995***				3.445***	
		(0.472)	(0.782)				(1.138)	
committee.sent		−0.476**					−1.113**	−1.210**
		(0.202)					(0.471)	(0.482)
demand.sent			−0.541**				−1.193***	−0.957***
			(0.230)				(0.347)	(0.288)
inflation.sent						0.859**		1.419***
						(0.361)		(0.427)
supply.sent	−0.124**	−0.311***	−0.444***			−0.631***	−0.993***	−0.978***
	(0.048)	(0.097)	(0.158)			(0.217)	(0.211)	(0.220)
bankrate.sent							0.785***	
							(0.212)	
financial.sent	0.071					0.517*	0.882**	0.807**
	(0.059)					(0.264)	(0.353)	(0.314)
gdp.sent		−0.730***	−0.821***					
		(0.191)	(0.294)					
consumption.sent		−0.197**						
		(0.087)						
business.sent	0.006	−0.639***	−0.326					
	(0.057)	(0.155)	(0.238)					
prices.sent				0.707***	0.734***			
				(0.254)	(0.259)			
employment.sent				−0.708***	−0.870***	−0.756***	−0.817***	
				(0.183)	(0.198)	(0.231)	(0.257)	
sterling.sent								1.255***
								(0.339)
Observations	95	81	93	92	91	90	76	75
R^2	0.105	0.437	0.276	0.289	0.423	0.525	0.695	0.650
Note:								*p<0.1; **p<0.05; ***p<0.01

corresponding to $h = 1$. These show that the the 'nowcast' of y_t in the period t-*Report* will be changed relative to the $t-1$-*Report* by 0.188 times the change in sentiment between the $t-1$- and $t-2$-*Reports*. The second column for $h = 2$ suggests the change in the forecasts between the $t-1$- and $t-2$-*Reports*

Table 9: Forecast encompassing test output employing model selection on sentiment sub-indices for GDP growth.

	Dependent variable:							
	Forecast Error (quarters ahead)							
Quarters Ahead	1	2	3	4	5	6	7	8
Const.	−0.603***	−0.941***	−1.336***					3.186***
	(0.080)	(0.126)	(0.192)					(1.197)
inflation.sent				−1.611***	−2.604***	−3.088***	−3.125***	
				(0.565)	(0.691)	(0.779)	(0.556)	
demand.sent				−1.057***	−0.901***	−1.188***		
				(0.227)	(0.255)	(0.322)		
forc								−1.187***
								(0.444)
general.sent								−6.576***
								(1.725)
housing.sent								0.808***
								(0.269)
employment.sent						1.174**		1.518**
						(0.510)		(0.592)
financial.sent	0.348**	0.540**		1.710***	2.661***	2.204***		
	(0.153)	(0.238)		(0.518)	(0.581)	(0.654)		
supply.sent			0.705***	0.910**				
			(0.227)	(0.345)				
household.sent	0.309***	0.548***	0.677***					
	(0.109)	(0.169)	(0.227)					
Observations	75	74	73	72	71	70	69	68
R^2	0.462	0.456	0.415	0.456	0.439	0.463	0.317	0.464
Note:								*p<0.1; **p<0.05; ***p<0.01

(of y_t) will be nearly a quarter (0.232) of the change in sentiment between $t-2$- and $t-3$-*Reports*, and so on.

For inflation, sentiment is significant at the two shortest horizons, and enters with a negative coefficient.

As noted in the text, although we are using multi-step forecast horizons, as we are considering the adjustment forecasts, the regression errors are not expected to be serially correlated, and inference is based on the usual formulae for coefficient estimator uncertainty.

7 Conclusions

We have considered the extent to which the narratives surrounding the Bank's published forecasts provide additional information regarding the inflation outlook. Not surprisingly, the Bank's forecasts are more accurate than simple model based forecasts at the shorter, within-year horizons, and are also

more accurate than forecasts derived from the narratives as sentiment indices. More interestingly, the sentiment indicator provides additional useful information - relative to the Bank's numeric forecast - for output growth, but not for inflation. Formally, the sentiment index is not forecast encompassed for output growth, but it is for inflation. If we consider various sentiment sub-indices there is more evidence that the narratives serve as a useful complement to the published numeric forecasts.

Finally, we also find evidence that sentiment 'leads' changes in the short-horizon numeric forecasts for output growth and for inflation, providing further evidence that the numeric output growth forecasts do not exhaust all the information in the *Reports*.

There are a number of avenues for future research. We have taken the published forecasts - the numerical forecasts and the narratives - at face value, as 'the' Bank's forecasts. Enquiry in to the forecasting process, and whether the forecasts are produced by the same people, at the same stage of the forecasting round, might shed light on our findings. We have looked at the relationship between changes in the narratives and (subsequent) changes in the numerical forecasts, although there might be effects in the reverse direction, and we could also consider persistence and more complicated dynamics more generally.

Given the interest in density forecasting, and the Bank's published history of inflation fancharts, a natural extension of our research is to a consideration of the relationship between the narratives and statistics derived from the fancharts, such as estimates of uncertainty and downside and upside risks.

Table 10: Numerical forecast updates and prior sentiment revisions.

	Forecast Adjustment for Inflation (quarters ahead)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	−0.004	0.028	0.063	0.057	0.037	−0.014	−0.074**
	(0.041)	(0.048)	(0.054)	(0.054)	(0.042)	(0.040)	(0.035)
Sentiment change	−0.273***	−0.273**	−0.192	−0.202	0.029	−0.041	−0.070
	(0.103)	(0.121)	(0.134)	(0.136)	(0.104)	(0.099)	(0.088)
							Observations
Observations	85	85	85	85	85	85	85
R^2	0.078	0.058	0.024	0.026	0.001	0.002	0.008

Note: *p<0.1; **p<0.05; ***p<0.01

Standard errors in parentheses beneath coefficient estimates.

Table 11: Numerical forecast updates and prior sentiment revisions.

	Forecast Adjustment for GDP growth (quarters ahead)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	−0.094*	−0.137**	−0.182**	−0.150**	−0.129**	−0.106**	−0.064
	(0.050)	(0.059)	(0.070)	(0.072)	(0.060)	(0.052)	(0.049)
Sentiment change	0.188**	0.232**	0.273**	0.254**	0.169*	0.098	0.038
	(0.082)	(0.097)	(0.116)	(0.119)	(0.099)	(0.086)	(0.081)
							Observations
Observations	75	75	75	75	75	75	75
R^2	0.067	0.072	0.071	0.059	0.038	0.017	0.003

Note: *p<0.1; **p<0.05; ***p<0.01

Standard errors in parentheses beneath coefficient estimates.

References

J Alba. Basic sentiment analysis. `https://github.com/fjavieralba/basic_sentiment_analysis`, 2012.

A. Atkeson and L. Ohanian. Are Phillips Curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25:2–11, 2001. (1).

S. Bird, E. Loper, and E. Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

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

J. L. Castle, N. W. P. Fawcett, and D. F. Hendry. Forecasting breaks and during breaks. In M. P. Clements and D. F. Hendry, editors, *Oxford Handbook of Economic Forecasting*, pages 315–353. Oxford University Press, Oxford, 2011.

J.L. Castle, D.F. Hendry, and A.B. Martinez. Evaluating Forediction Failure. unpublished, Department of Economics, University of Oxford, 2015.

Y. Y. Chong and D. F. Hendry. Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, 53:671–690, 1986. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press.

T. E. Clark and M. W. McCracken. Testing for unconditional predictive ability. In M. P. Clements and D. F. Hendry, editors, *The Oxford Handbook of Economic*, pages 415–440. Oxford University Press, 2011.

M. P. Clements. Internal consistency of survey respondents' forecasts: Evidence based on the Survey of Professional Forecasters. In J. L. Castle and N. Shephard, editors, *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry. Chapter 8*, pages 206–226. Oxford University Press, Oxford, 2009.

M. P. Clements. Explanations of the Inconsistencies in Survey Respondents Forecasts. *European Economic Review*, 54(4):536–549, 2010.

M. P. Clements. Are professional macroeconomic forecasters able to do better than forecasting trends? *Journal of Money, Credit and Banking*, 47,2-3:349–381, 2015. DOI: 10.1111/jmcb.12179.

M. P. Clements and A. B. Galvão. Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. *Journal of Business and Economic Statistics*, 26:546–554, 2008. No. 4.

M. P. Clements and D. F. Hendry. Evaluating a model by forecast performance. *Oxford Bulletin of Economics and Statistics*, 67:931–956, 2005.

M. P. Clements and D. F. Hendry. Forecasting with breaks. In G. Elliott, C. W. J Granger, and A Timmermann, editors, *Handbook of Economic Forecasting, Volume 1. Handbook of Economics 24*, pages 605–657. Elsevier, Horth-Holland, 2006.

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

G. Di Fatta, J.J. Reade, Jaworska S., and A. Nanda. Big Social Data and Political Sentiment: the Tweet Stream during the UK General Election 2015 Campaign. Proceedings, 8th IEEE International Conference on Social Computing and Networking (SocialCom 2015), 2015.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–263, 1995. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

J. Engelberg, C. F. Manski, and J. Williams. Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics*, 27(1):30–41, 2009.

N. R. Ericsson. Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration. *Journal of Policy Modeling*, 14:465–495, 1992.

Neil. R. Ericsson. Eliciting {GDP} forecasts from the FOMC's minutes around the financial crisis. *International Journal of Forecasting*, 32(2):571 – 583, 2016. doi: http://dx.doi.org/10.1016/j.ijforecast.2015.09.007.

R. C. Fair and R. J. Shiller. Comparing information in forecasts from econometric models. *American Economic Review*, 80:39–50, 1990.

N. Fawcett, L. Körber, R.M. Masolo, and M. Waldron. Evaluating UK point and density forecasts from an estimated DSGE model: the role of off-model information over the financial crisis. Staff Working Paper 538, Bank of England, July 2015.

E. Ghysels, A. Sinko, and R. Valkanov. MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26:53–90, 2007.

Raffaella Giacomini. Testing conditional predictive ability. In M. P. Clements and D. F. Hendry, editors, *The Oxford Handbook of Economic*, pages 441–455. Oxford University Press, 2011.

C. W. J. Granger and P. Newbold. Some comments on the evaluation of economic forecasts. *Applied Economics*, 5:35–47, 1973. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

K. Holden and D. A. Peel. On testing for unbiasedness and efficiency of forecasts. *The Manchester School*, 58:120–127, 1990.

A. Meinusch and P. Tillmann. Quantitative easing and tapering uncertainty: Evidence from twitter. Joint discussion paper series in economics, Universities of Aachen, Giessen, Göttingen, Kassel, Marburg and Siegen, 2015.

J.A. Mincer and V. Zarnowitz. The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 1–46. NBER, 1969.

C. R. Nelson. The prediction performance of the FRB-MIT-PENN model of the US economy. *American Economic Review*, 62:902–917, 1972. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

F. Å. Nielsen. Afinn, mar 2011. URL `http://www2.imm.dtu.dk/pubdb/p.php?6010`.

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

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley and Sons, Ltd., 2006.

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

M.H. Pesaran and M. Weale. Survey Expectations. In G. Elliott and A. Timmerman, editors, *Handbook of Economic Forecasting*, volume 1, pages 715–776. Elsevier, 2006.

F. Pretis, J.J. Reade, and G. Sucarrat. *gets: General-to-Specific (GETS) Modelling and Indicator Saturation Methods*, 2014. URL `http://CRAN.R-project.org/package=gets`. R package version 0.2.

C. D. Romer and D. H. Romer. The FOMC versus the Staff: Where can monetary policymakers add value? *American Economic Review*, 98:2:230–235, 2000.

Herman. Stekler and Hilary Symington. Evaluating qualitative forecasts: The FOMC minutes, 2006-2010. *International Journal of Forecasting*, 32(2):559 – 570, 2016. doi: http://dx.doi.org/10.1016/j.ijforecast.2015.02.003.

J. H. Stock and M. W. Watson. Forecasting inflation. *Journal of Monetary Economics*, 44:293–335, 1999.