Nowcasting Norwegian GDP in Real-Time: A Density Combination Approach *

Knut Are Aastveit[†] Karsten R. Gerdrup[‡] Anne Sofie Jore[§] Leif Anders Thorsrud[¶]

October 6, 2010

Prelimenary and incomplete - Please do not quote

Abstract

In this paper we use an expert combination framework to produce density combination nowcasts for Norwegian Mainland-GDP from a system of VARs, leading indicator models, factor models and a DSGE model. We update the density nowcasts from our forecast combination framework several times during the quarter and highlight the importance of new data releases. We first show that the logarithmic score of the predictive densities for Norwegian Mainland-GDP increase monotonically as new information arrives during the quarter. Second, we show that the predictive densities for our combination approach is well-calibrated throughout the quarter, while this is not the case for all of the individual models and model classes. Especially, the predictive densities for the DSGE model do not match the performance of the other model classes.

JEL-codes: C32, C52, C53, E37, E52.

Keywords: Density combination; Forecast densities; Forecast evaluation; Monetary policy; Nowcasting; Real-time data

*We thank John Geweke, Francesco Ravazzolo and Shaun Vahey for helpful comments. The views expressed in this paper are those of the authors and should not be attributed to Norges Bank.

[†]*Corresponding author*: Knut Are Aastveit, Norges Bank, Economics Department, Bankplassen 2, 0107 Oslo, Norway. Telephone: +47 22 31 61 21. Fax: +47 22 42 40 62. Knut-Are.Aastveit@norges-bank.no

[‡]Norges Bank, Economics Department, Karsten.Gerdrup@norges-bank.no

[§]Norges Bank, Economics Department, Anne-Sofie.Jore@norges-bank.no

[¶]Norges Bank, Economics Department, Leif-Anders.Thorsrud@norges-bank.no

1 Introduction

Policy decisions in real-time are based on assessments of the recent past and current economic condition under a high degree of uncertainty. Many key statistics are released with a long delay, are subsequently revised and are available at different frequencies. In addition, the data generating process is unknown and is likely to change over time. As a consequence, there has been a substantial interest in developing a framework for forecasting the present and recent past, i.e. nowcasting. In a seminal paper, Giannone, Reichlin, and Small (2008) provide important amendments to the approximate dynamic factor model, as they adapt the model to account for an unbalanced dataset. Their framework allows study of the impact of different data releases and their importance for reducing the root mean square forecasting error (RMSFE) of U.S. GDP nowcasts.[1]

Until now, the academic literature on nowcasting has been focusing on developing single models that increase forecast accuracy in terms of point nowcast. This differs in two important ways from policy making in practice. First, policy makers are often provided with several different models which may provide rather different forecasts. This leads naturally to the question of model choice or combination.[2] Second, if the policy maker's loss function is not quadratic or if the world is nonlinear then it no longer suffices to focus solely on first moments of possible outcomes (point forecasts). To ensure appropriate monetary policy decisions, central banks therefore must provide suitable characterizations of forecast uncertainty. Density forecasts provide an estimate of the probability distribution of the forecasts. Mitchell and Hall (2005) and Hall and Mitchell (2007) provide some justification for density combination.

In this paper we use an expert combination framework to produce density combination nowcasts for Norwegian Mainland-GDP from a system of different model classes. To ensure

[1] Evans (2005) proposes an alternative framework that also allows for the use of non-synchronous data releases (jagged edge problem). This model is however not a factor model and only suitable for a limited number of variables.

[2] The idea of combining forecasts from different models was first introduced by Bates and Granger (1969). Their main conclusion is that a combination of two forecasts can yield lower mean square forecasts error than either of the original forecasts when optimal weights are used. Timmermann (2006) surveys combination methods and provides theoretical rationales in favor of combination - including unknown instabilities, portfolio diversification of models and idiosyncratic biases.

relevance for policy makers, we include vector autoregressive models (VARs), leading indicator models, factor models and a dynamic stochastic general equilibrium (DSGE) model. These four model classes are the most widely used at central banks. Our recursive nowcasting exercise is applied to Norwegian real-time vintage data. We update the density nowcasts from our forecast combination framework for every new data release during a quarter and highlight the importance of new data releases for the evaluation period 2001q2-2009q1. The density nowcasts are combined in a two-step procedure. In the first step, we group models into different model classes. The nowcasts for each model within a model class are combined using the logarithmic score, see among others Jore, Mitchell, and Vahey (2010). This yields a combined predictive density nowcast for each of the four different model classes. In a second step, these four predictive densities are combined into a new density nowcast, again using the logarithmic score. We then evaluate whether the predictive densities are well-calibrated.

First, we show that the logarithmic score for the predictive densities for Norwegian Mainland-GDP increases monotonically as new information arrives during the quarter. In particular, releases of leading indicators are important and improve the predictive densities the most. Interestingly, the weights attached to the four different model classes change during the quarter in correspondence with new data releases. In this way, our combination procedure attaches a higher weight to models with new and relevant information.

Second, we evaluate the nowcast densities for our combination approach as well as for the four different model classes using probability integral transforms (pits), see Diebold, Gunther, and Tay (1998). We show that the predictive densities from our density combination approach is well-calibrated throughout the quarter, while this is not the case for all of the four model classes.

Our paper is similar to Bache, Jore, Mitchell, and Vahey (2009), Amisano and Geweke (2009) and Gerdrup, Jore, Smith, and Thorsrud (2009) in that we combine different model classes. While, Bache, Jore, Mitchell, and Vahey (2009) focus on the performance and the weight attached to a DSGE model compared to a cluster of VARs, Gerdrup, Jore, Smith, and Thorsrud (2009) compares the approach of combining densities from different model classes rather than combining individual model densities directly. Amisano and Geweke (2009) focus on deriving optimal weights attached to three different classes of models, a VAR, a factor model and a DSGE. Our paper differ from all these papers in that we are interested in

providing nowcasts based on a density combination approach, where we focus on how new information changes the combined density throughout the quarter. To our knowledge, this is the first paper to study density combination in such a nowcasting framework.

The rest of the paper is organized as follows. In the next section we describe the modeling framework and discuss the rational for combining densities for different model classes. In the third section we describe the real-time data set and the suite of individual models, while the fourth section describes the recursive forecasting exercise. The fifth section contains the results of the out-of-sample nowcasting experiment. Finally, we conclude in the sixth section.

2 Model

The problem at hand requires aggregating N density forecasts for forecasters i $(i = 1, ..., N)$ for a variable y_t at time t $(t = 1, ..., T)$.[3] The modeler is thus left with many possibilities for choosing both combination method(s), weights and individual forecasters. Below we describe how we in this application aggregate a set of individual density forecasts, how we derive the individual model weights using scoring rules, and describe the set of N individual forecasters. For details and a more thorough description of possible scoring rules, combination strategies and derivations, see for example Timmermann (2006) and Hall and Mitchell (2007).

2.1 Combining densities

One popular approach to solve the aggregation problem is to take a linear combination of the individual density forecasts, the so called linear opinion pool:

$$p(y_{\tau,h}) = \sum_{i=1}^{N} w_{i,\tau,h}\, g(y_{\tau,h}|I_{i,\tau}), \quad \tau = \underline{\tau}, ..., \overline{\tau} \tag{1}$$

where $I_{i,\tau}$ is the information set used by model i to produce the density forecast $g(y_{\tau,h}|I_{i,\tau})$ for variable y at forecasting horizon h. $\underline{\tau}$ and $\overline{\tau}$ are the period over which the individual forecasters's densities are evaluated, and finally $w_{i,\tau,h}$ are a set of non-negative weights that sum to unity (see section 2.2).

Combining the N density forecasts according to equation 1 can potentially produce a combined density forecasts with characteristics quite different from those of the individual

[3] In our application, the individual forecasters will denote different econometric models.

forecasters. As Hall and Mitchell (2007) notes; if all the individual forecasters' densities are normal, but with different mean and variance, the combined density forecast using the linear opinion pool will be mixture normal. This distribution can accommodate both skewness and kurtosis and be multimodal, see Kascha and Ravazzolo (2010). If the true unknown density is non-normal, this is a appealing feature. If on the other hand, the true unknown density is normal, combining the individual forecast densities using equation 1, will in general get the distribution wrong. Further, since the combined density is a linear combination of all the individual forecasters' densities, the variance of the combined density forecast will in general be higher than that of individual models. However, this is not necessarily deleterious, as the combined density may perform better than the individual density forecasts when evaluated.

Other alternative combination methods do exist, for example the logarithmic opinion pool. However, from a theoretical perspective, no scheme is obviously superior to the other.[4] *Comment: We will check if our results are robust to using logarithmic opinion pool instead of linear opinion pool as combination method.*

2.2 Deriving the weights

The key determinant in equation 1 is defining the weights, $w_{i,\tau,h}$. Many different weighting schemes have been proposed in the literature. For point forecast combinations the naive approach using equal weighting has proven useful, see e.g. Stock and Watson (2004). In a density combination setting, there seems to be more leverage by adopting more sophisticated strategies, see e.g. Jore, Mitchell, and Vahey (2010) and Amisano and Geweke (2009).

In this application we apply logarithmic score (log score) weights. The log score weights are recursively updated, and thus time-varying.

2.2.1 Recursive log score weights

The log score is the logarithm of the probability density function evaluated at the outturn of the forecast. As discussed in Hoeting, Madigan, Raftery, and Volinsky (1999), the log score is a combined measure of bias and calibration. The preferred densities will thus have probability

[4] Wallis (2010) finds weak statistical evidence in favor of logarithmic opinion pools in a monte carlo simulation study. Bjørnland, Gerdrup, Jore, Smith, and Thorsrud (2010) finds similar results in an empirical study.

mass centred on the correct location. Following Jore, Mitchell, and Vahey (2010) we define the log score weights as:

$$w_{i,\tau,h} = \frac{exp[\sum_{\underline{\tau}}^{\tau-h} ln\ g(y_{\tau,h}|I_{i,\tau})]}{\sum_{i=1}^{N} exp[\sum_{\underline{\tau}}^{\tau-h} ln\ g(y_{\tau,h}|I_{i,\tau})]}, \quad \tau = \underline{\tau}, ..., \overline{\tau} \tag{2}$$

where τ, h, y, N, i and $g(y_{\tau,h}|I_{i,\tau})$ are defined above. Two points are worth emphasizing about this expression. The weights are derived based on out-of-sample performance, and the weights are horizon specific.

Note that maximizing the log score is the same as minimizing the Kullback-Leibler distance between the models and the true but unknown density. Mitchell and Wallis (2010) show that the difference in log scores between an "ideal" density and a forecast density, that is the Kullback-Leibler information criterion (KLIC), can be interpreted as a mean error in a similar manner to the use of the mean error or bias in point forecast evaluation.[5]

2.3 Two stage combination approach

In this paper, we use a two stage approach to combine forecast densities. Aiolfi and Timmermann (2006) find that forecasting performance can be improved by first sorting models into clusters based on their past performance, second by pooling forecasts within each cluster, and third by estimating optimal weights on these clusters (followed by shrinkage towards equal weights). Our approach is close to Aiolfi and Timmermann (2006) in the sense that we combine models in more than one stage. However, our focus is mainly on density forecasting. Furthermore, we are particularly interested in the case of a model suite which is populated by a wide range of models which are typically considered by central banks when they form their views on the future trajectory of the economy. Instead of grouping models according to past forecast performance, we follow Gerdrup, Jore, Smith, and Thorsrud (2009) and group models that share the same information set or model structure together.

In the first step, density forecasts for Norwegian Mainland-GDP from a large number of models are grouped into different model classes and combined. In the next step, we combine

[5]The log score weighting procedure applied here can however not be interpreted as optimal in the sense that it will optimize the log score of the combined density. Amisano and Geweke (2009) show how optimal weights will produce a combined density with a log score higher or equally high as the best individual model. This is not necessarily the case with the log score weighting strategy we apply in this analysis.

the density forecasts from each model class. In both steps, we use the linear opinion pool for combination and the recursive log score weights as described above.

2.4 Forecast density evaluation

Following Diebold, Gunther, and Tay (1998), we evaluate the density relative to the "true" but unobserved density using the probability integral transform (pits). The pits summarize the properties of the densities, and may help us to judge whether the densities are biased in a particular direction, and whether the width of the densities has been roughly correct on average. More precisely, the pits defined as $z_{\tau,h}$, where $z_{\tau,h} = \int_{-\infty}^{y_{\tau,h}} p(u)du$ are the ex-ante inverse predictive cumulative distribution evaluated at the ex-post actual observations, see Geweke and Amisano (2010).

A density is correctly specified if the pits are uniform and, for one-step ahead forecasts, independently and identically distributed. Accordingly, we may test for uniformity and independence at the end of the evaluation period. Several candidate tests exists, but few offer a composite test of uniformity and independence together, as would be appropriate for one-step ahead forecasts. In general, tests for uniformity are not independent of possible dependence and vice versa. Since the appropriateness of the tests are uncertain, we conduct several different tests. See Hall and Mitchell (2007) for elaboration and description of different tests.

We use a test of uniformity of the pits proposed by Berkowitz (2001). The Berkowitz test works with the inverse normal cumulative density function transformation of the pits. Then we can test for normality instead of uniformity. For 1-step ahead forecasts, the null hypothesis is that the transformed pits are identically and independently normally distributed, iid N(0,1). The test statistics is χ^2 with three degrees of freedom. For longer horizons, we do not test for independence. In these cases, the null hypothesis is that the transformed pits are identically, normally distributed, N(0,1). The test statistics is χ^2 with two degrees of freedom. Other tests of uniformity are the Anderson-Darling (AD) test (see Noceti, Smith, and Hodges (2003)) and a Pearson chi-squared test suggested by Wallis (2003). Note that the two latter tests are more suitable for small-samples. Independence of the pits is tested by a Ljung-Box test, based on autocorrelation coefficients up to four for one-step ahead forecasts. For forecast horizons h>1, we test for autocorrelation at lags equal to or greater than h.

3 The Norwegian data and model classes

In this section, we describe the Norwegian real-time dataset and we list and describe the models and model classes used to construct nowcast density combinations.

3.1 Norwegian real-time dataset

We use a real-time dataset in the forecasting exercise. Vintages of the main aggregates of Quarterly National Account (QNA) are collected from Statistics Norway (SN) and stored in a real-time database at Norges Bank. The oldest vintage that is available was published in June 2000. Some of the early vintages are not complete, since SN sometimes only saved the last 8 quarters of data on their website. Since March 2003 the QNA vintages have been complete, with time series starting in 1978Q1.[6] The missing values are replaced by using growth rates from neighboring vintages.

All series used in this study have been saved in a real-time database since May 2009. To enable real-time analysis over a longer period, we have created artificial real-time series where possible. For example, unadjusted survey data are not revised. For the Business Tendency Survey (BTS) we have created vintages of real-time data by seasonally adjusting (and smoothing) the unadjusted series recursively, starting with vintages in the mid 1990s. Checking against published data the last few years indicate that these artificial real-time series provide good approximations.

Quarterly National Account are heavily revised. In a Norges Bank staff memo (not yet published) we analyze revisions to GDP and some main demand components. The main reasons for the revisions are changes in base year (which happens each year) and changing seasonal patterns. In addition, major revisions occur from time to time, but this is not a big issue in our sample. Figure 1 shows different vintages of Norwegian Mainland-GDP.

According to the revisions analysis, the revisions of growth rates (both quarter to quarter and year over year) of GDP mainland Norway seem to contain news when looking at the entire period, that is initial to second release, initial to fifth release, initial to eleventh release and so on. However, revisions from the fifth release to the eleventh release seem to mostly reduce noise.

[6] The vintage published in February 2002 was also complete.

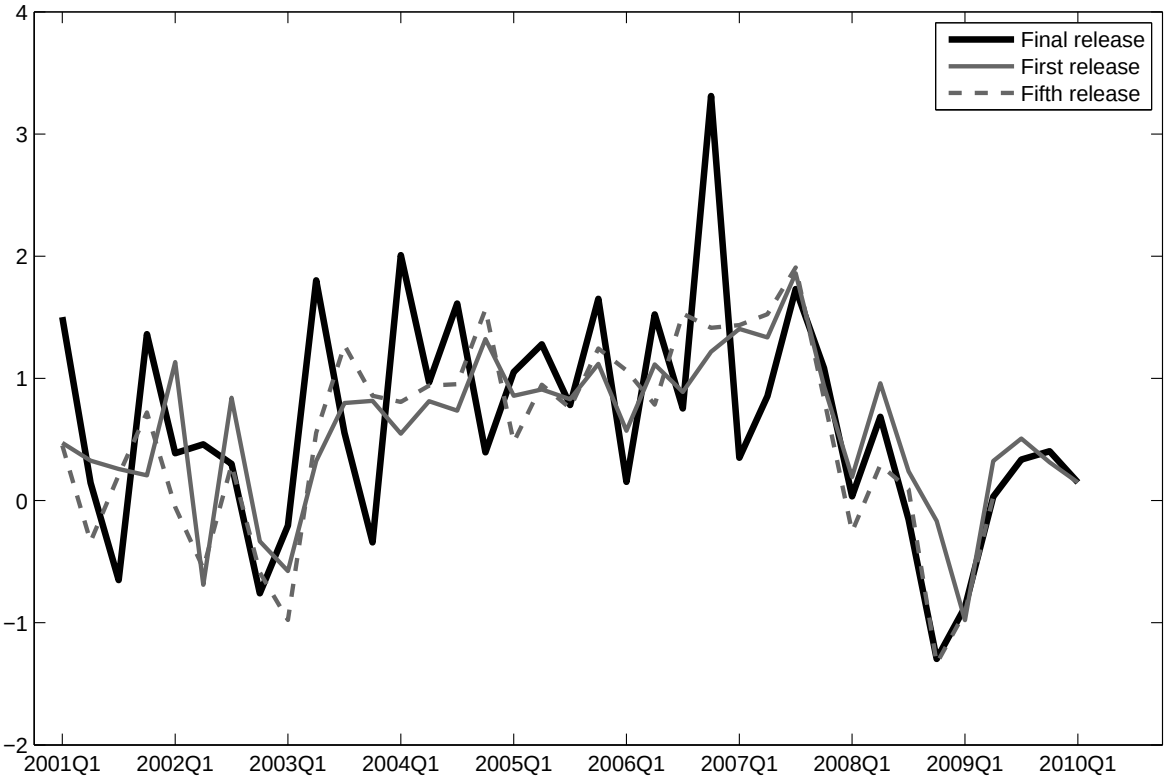


Figure 1. *First release, fifth release and final vintage of Norwegian Mainland-GDP*

3.2 Model classes

We use four different classes of models. To ensure relevance for policy makers, we include VARs, leading indicator models, factor models and a DSGE model. These are all model classes widely used at central banks. A brief description of the models are given in Table 1 and a more thorough description is given below.

3.2.1 VARs

We consider a range of AR models for GDP growth, bivariate VARs with GDP growth and inflation, bivariate VARs with GDP growth and the interest rate, and trivariate VARs with GDP growth, inflation and the interest rate. The models are estimated with maximum lag lengths of 1 to 4. We allow for the following three transformations of the variables prior to estimation; first-differences, double-differences and de-trended. Finally, recent work by Clark and McCracken (2010) suggest that VARs may be prone to instabilities. We therefore consider three different estimation periods. The whole estimation period (recursively), a short

Table 1. A summary of component model classes

Model classes	Description	Number of models
VAR	Univariate and Vector Autoregressions using GDP (and inflation and/or interest rate)	144
IND	Bivariate VARs with GDP and survey indicators	45
FM	Dynamic Factor Models	16
DSGE	Dynamic Stochastic General Equlibrium Model	1
Sum		206

rolling window of 20 quarters, and a longer rolling window of 40 quarter. In total, we have 36 AR models, 72 bivariate VARs and 36 trivariate VARs.

3.2.2 Leading indicator models

There is a large amount of studies showing that leading indicators can be useful for economic forecasting, see among others Banerjee, Marcellino, and Masten (2005), Banerjee and Marcellino (2006) and Marcellino (2006) for a survey on leading indicators in macroeconomics. Giannone, Reichlin, and Small (2008) shows that timeliness of data matters for nowcasting, that is the exploitation of early releases leads to improvement in the nowcast accuracy. Survey data is perhaps the most important type of leading indicator data. Several papers, such as Frale, Marcellino, Mazzi, and Proietti (2010) and Banbura and Rünstler (2010), have found that survey data, which provide the most timely information, contribute to substantial improvements of nowcasts. For Norway, there exist relatively few surveys of substantial time length and these are mostly quarterly series published with a lag. We include aggregate and disaggregated measures of business tendency surveys (BTS) and consumer confidence surveys (TNS) and a regional network survey (see Brekke and Halvorsen (2009)). In addition, we include different variables from the labour market, money, credit and measures of new orders in industry and construction. We include each indicator variable in a bi-variate VAR with GDP growth. In total we have 45 different bivariate leading indicator models.

3.2.3 Factor Models

The objective of factor models is to summarize the information contained in large datasets, while at the same time reducing their dimension. In other words, to reduce the parameter space. This type of models have been increasingly popular at central banks as they tend to have good forecasting properties. The model that we consider is an approximate dynamic factor model similar to Giannone, Reichlin, and Small (2008). The model uses monthly information from a large unbalanced dataset for Norway including for instance financial variables, different measures of industrial production, consumer prices and commodity prices as well as labour market data. Many of these series can be revised. For some of these variables we do not have real-time data. In the analysis here, we truncate these series recursively.[7] Note that the surveys described above are quarterly variables and are not included in the factor model.[8] The model is estimated with a maximum number of factors of 1 to 4 and with a maximum lag length of 1 to 4. This yields 16 different specifications of the model. For more details about the model, its performance and the dataset, see Aastveit and Trovik (2007).

3.2.4 The DSGE Model

The DSGE model we consider is a version of NEMO.[9] It is a medium-scale New Keynesian small open economy model with a similar structure to other DSGE models developed in many central banks. The model is estimated using Bayesian maximum likelihood on seasonal adjusted data for mainland GDP growth, consumption growth, investment growth, export growth, employment, inflation (CPIATE), imported inflation, real wage growth, the real exchange rate (I44) and the nominal interest rate. The sample period starts in 1987Q1. In this version, the steady-state levels are equal to recursively updated means of the variables.

[7] Aastveit and Trovik (2008) show using U.S. real-time data that factor models are relatively robust to data revisions. That is, the factors extracted from a quasi real-time data set are very similar to factors extracted from a fully real-time data set.

[8] Schumacher and Breitung (2008) and Banbura and Modugno (2010) suggests extension to the approximate dynamic factor model allowing for mixed frequency data using the EM algorithm.

[9] NEMO is the core model used by Norges Bank for monetary policy, see Brubakk, Husebø, Maih, Olsen, and Øsntor (2006) for documentation.

4 Empirical exercise and ordering of data blocks

Our recursive forecasting exercise is intended to mimic the behavior of a policymaker nowcasting in real-time. We use real-time vintage data for the Norwegian economy for all forecasts and realizations, see section 3 for details. A key issue in this exercise is the choice of benchmark representing the "final" measure of GDP. Stark and Croushore (2002) suggest three alternative benchmark data vintages: the most recent data vintage, the last vintage before a structural revision (called benchmark vintages) and finally the vintage that is released a fixed period of time after the first release. We follow Clark and McCracken (2010) and Jore, Mitchell, and Vahey (2010) and choose the latter approach. However, we differ from the two mentioned papers as we use the fifth release of GDP as "final" measure of GDP in contrast to using the second release. Revisions in Norwegian GDP are larger than in U.S. GDP. Most of the revisions during the first year are due to "news". Hence, we choose the fifth estimate of GDP as benchmark but do check for robustness of our results with respect to other measures of '"final" GDP.

We perform a real-time out-of-sample density nowcasting exercise for Norwegian Mainland-GDP growth. The recursive forecast exercise is constructed as follows. We estimate each model on a real-time sample and compute model nowcast/backcast for GDP. For each vintage of GDP we re-estimate all models and compute predictive densities for every new data release within the quarter of interest (nowcast) and until the first estimate of GDP is released by SN. In Norway, this will be approximately 7 weeks after the end of the quarter of interest. By then the nowcast has turned into a backcast for that quarter. We have constructed a stylized calender with the full sequence of data releases during the quarter of interest until the first release of GDP. The data that are considered are either of monthly or quarterly frequency. Hence, some blocks of data will be updated every month, while others are only updated once every quarter. Series such as equity prices, dividend yields, currency rates, interest rates and commodity prices are constructed as monthly averages of daily observations. Following the standard approach, data series that have similar release dates and are similar in content are grouped together in blocks. We have defined a total of 17 different blocks (in addition to GDP itself) that are released on 10 different dates throughout the months, i.e., on some dates more than one block is released. The number of variables in each block varies from 25 in "Consumer Prices" to only 1 in for instance the "Regional Network Survey" and

"Services" blocks. In Figure 2 we depict how the 17 different blocks are released throughout any month and quarter. Note that release lags vary for different blocks, ranging from 4 months for the "Labour Market 2" block to 1 month for "Financials", "Interest Rates" and "Commodity Prices". Thus, the structure of the unbalancedness changes when a new block is released.

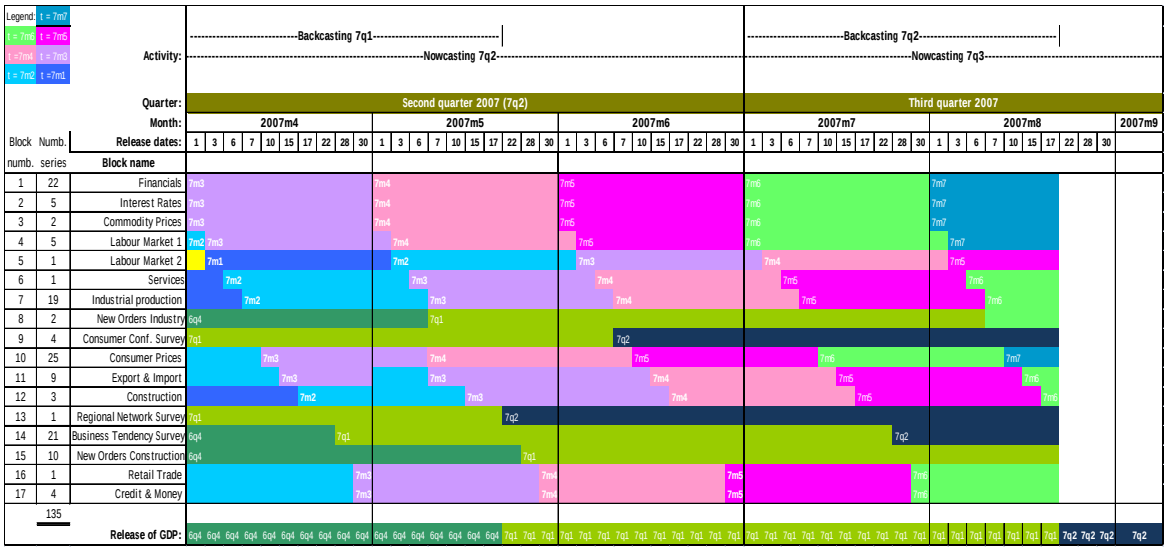


Figure 2. *Structure of data releases from the start of the quarter until the first estimate of GDP is released. April to August 2007 is used as an example. Data have either monthly or quarterly frequency. As an example, focus on the Industrial Production (IP) block. The seventh line in the figure shows that new observations become available on the 7th each month and that it is lagging two months when it becomes available, i.e., "t-2" according to the legend in upper left corner. The IP block contains 19 variables. The bottom line depicts the date of the currently available GDP through the quarter.*

Figure 2 illustrates the data release calender of a generic quarter, exemplified by the second and third quarter of 2007. The quarters are divided into three months and each month into 10 release dates. Horizontal bars indicate the release dates of the 17 different blocks of data. The color scheme of the bars indicate that when released, the data are lagging in varying degree as explained by the legend in the upper left corner (t refers to the third month of the quarter). Moreover, the date, year and month of the latest available data at any point in time during the quarter is printed on the bars. Thus the bars illustrate the jagged edges in

the data set.

The bottom line of the figure indicates the calendar for GDP releases. Midmonth in the second month of the quarter, the first estimate of GDP for the previous quarter is released. Hence, as indicated at the top of the figure, in the first month and a half of the quarter we can use the data to backcast the previous quarter GDP and to nowcast the current quarter GDP. The nowcast will then be a two-step ahead forecasts. After GDP is released for the previous quarter, the nowcast will be a one-step ahead forecast for the remaining part of the quarter. At the end of the quarter, it is still 7 weeks before the first estimate of GDP is released. The one-step ahead forecast made after the end of the quarter of interest will therefore be a backcast. In this exercise we are interested in investigating the information in all data releases from the beginning of a quarter until the first estimate of GDP is released. The procedure is done recursively so that once GDP is released we extend the sample by one quarter, re-estimate each model and compute nowcasts/backcast for all the different blocks for the new quarter. The exercise is repeated over the evaluation period, starting in $2001q2$ and ending in $2009q1$.

5 Results

In this section, we analyze the performance of our two-stage density combination approach. First, we analyze the importance of new information in terms of providing more accurate density nowcasts. Then, we evaluate the different densities and test whether they are well-calibrated.

5.1 Evaluation of results

We measure the forecasting performance in terms of evaluating both the root mean square forecasting error of the mean of the predictive densities and the log score of the predictive densities produced with every new data release during a quarter.

First, we study the impact of different data releases on the nowcasting/backcasting precision measured by RMSE. In Figure 3 we measure the performance by RMSE between the nowcast/backcast and subsequent realizations of GDP growth. The figure depicts both the RMSE from the forecasts of the four different model classes as well as the combined fore-

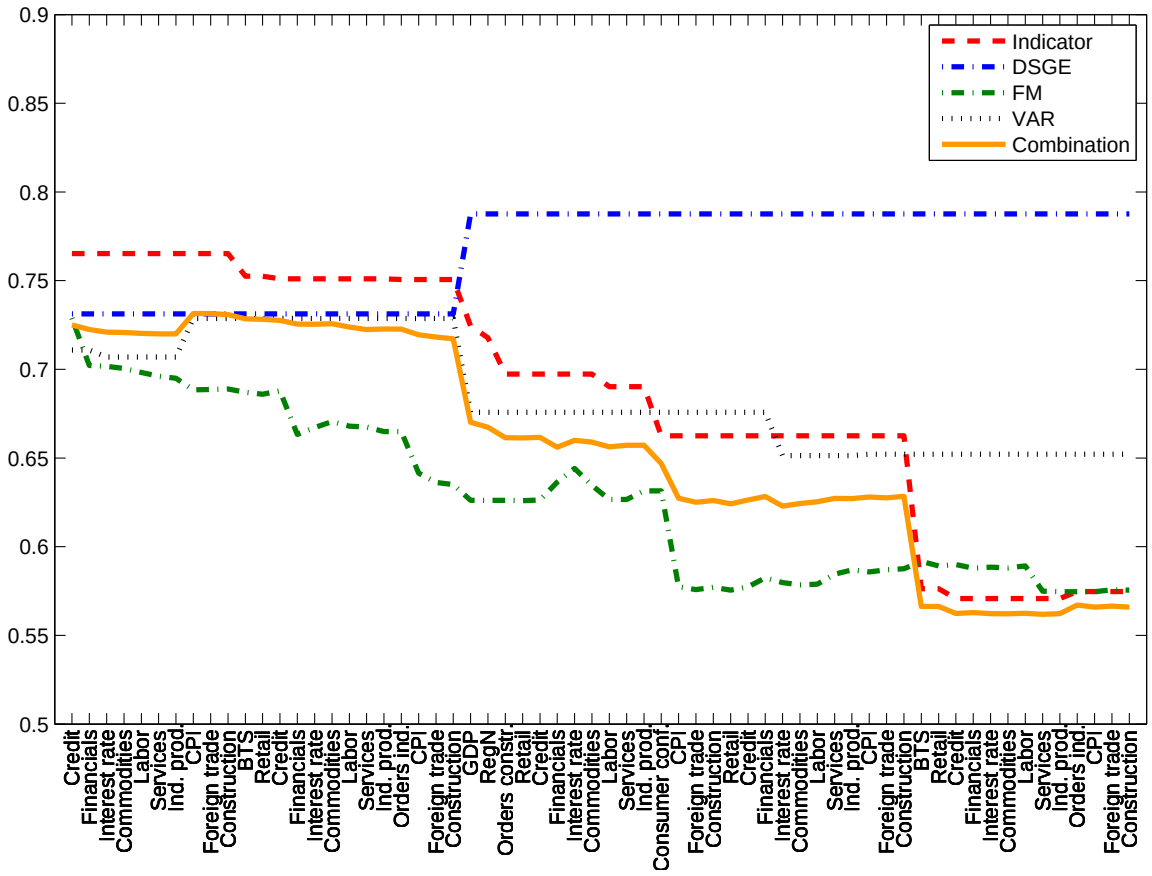


Figure 3. *Mean square forecasting errors made after different block releases. Evaluated against 5th release of data*

cast. The forecasting errors is steadily reduced for the combined model as new information becomes available throughout the quarter and until the first estimate of GDP is released. This is also the case for the forecasting errors from the VARs, the Factor models and the Leading indicator models. The DSGE model, on the other hand, seems to perform worse after GDP for the previous quarter is released. This indicates that in terms of RMSE, the 2-step ahead forecast from the DSGE model performs better than the 1-step ahead forecast. The factor model is clearly the best performing component model during the quarter and well into the next quarter. This is in line with previous findings, that factor models have very good nowcasting properties, see for instance Giannone, Reichlin, and Small (2008) and Aastveit and Trovik (2007). Note also that these models have an informational advantage compared to the other models, as they use monthly information. All other models only use

quarterly information. In particular asset prices from the Oslo Stock Exchange and disaggregated industrial production and consumer price series seem to improve the nowcasts early in the quarter. See Aastveit and Trovik (2007) for a more thorough discussion on this. Furthermore, the figure shows that leading indicator models are the worst performing model class early in the quarter. However, the performance is substantially improved during the quarter and especially after the business tendency survey is released approximately one month before the first estimate of GDP is released. A more detailed description of the importance of the different data releases for the factor models, VARs and leading indicator models are given in figure A.1 in the appendix. Finally, note that the combined forecast is performing considerably worse than the factor model for the first three months, but improves substantially after the business tendency survey is released. For the last weeks before GDP is released, the combined forecast is performing better than the forecast for all of the different model classes.

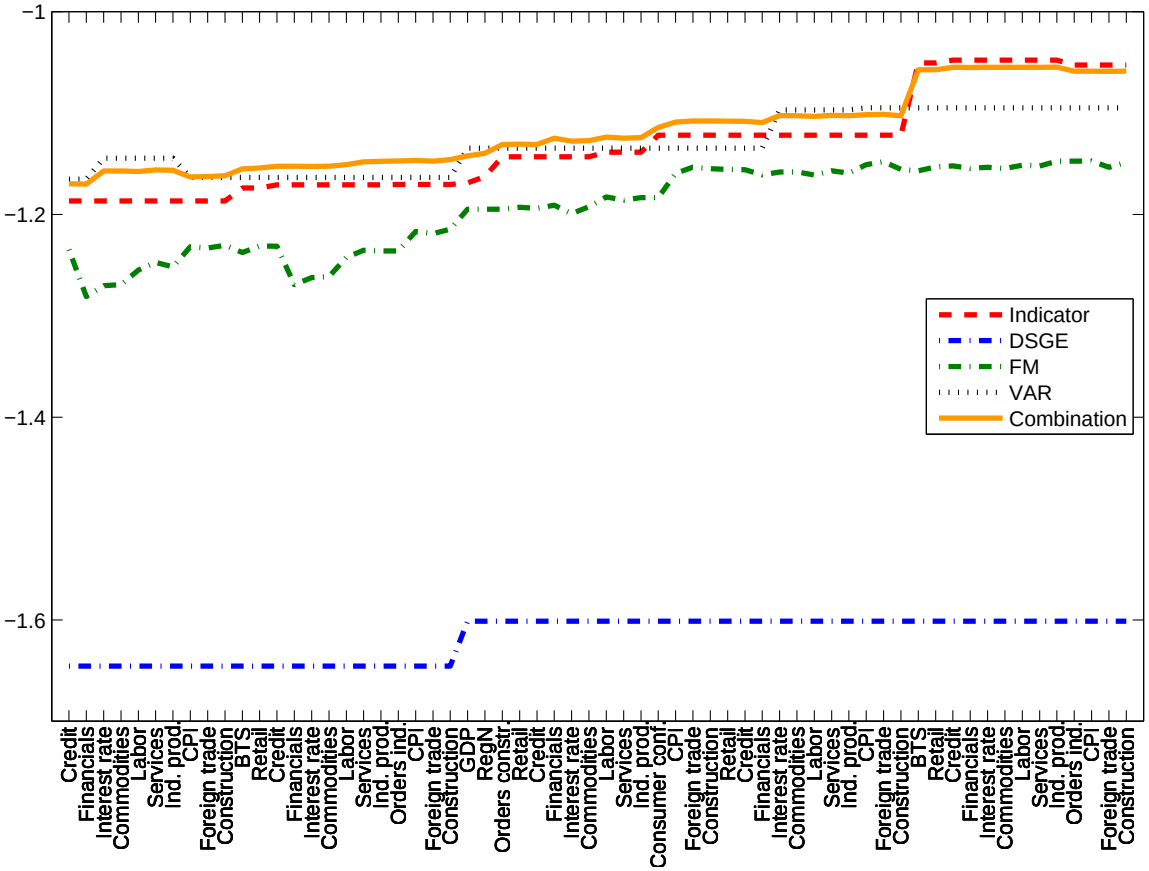


Figure 4. *Average log scores for forecasts after different block releases. Evaluated against 5th release of data*

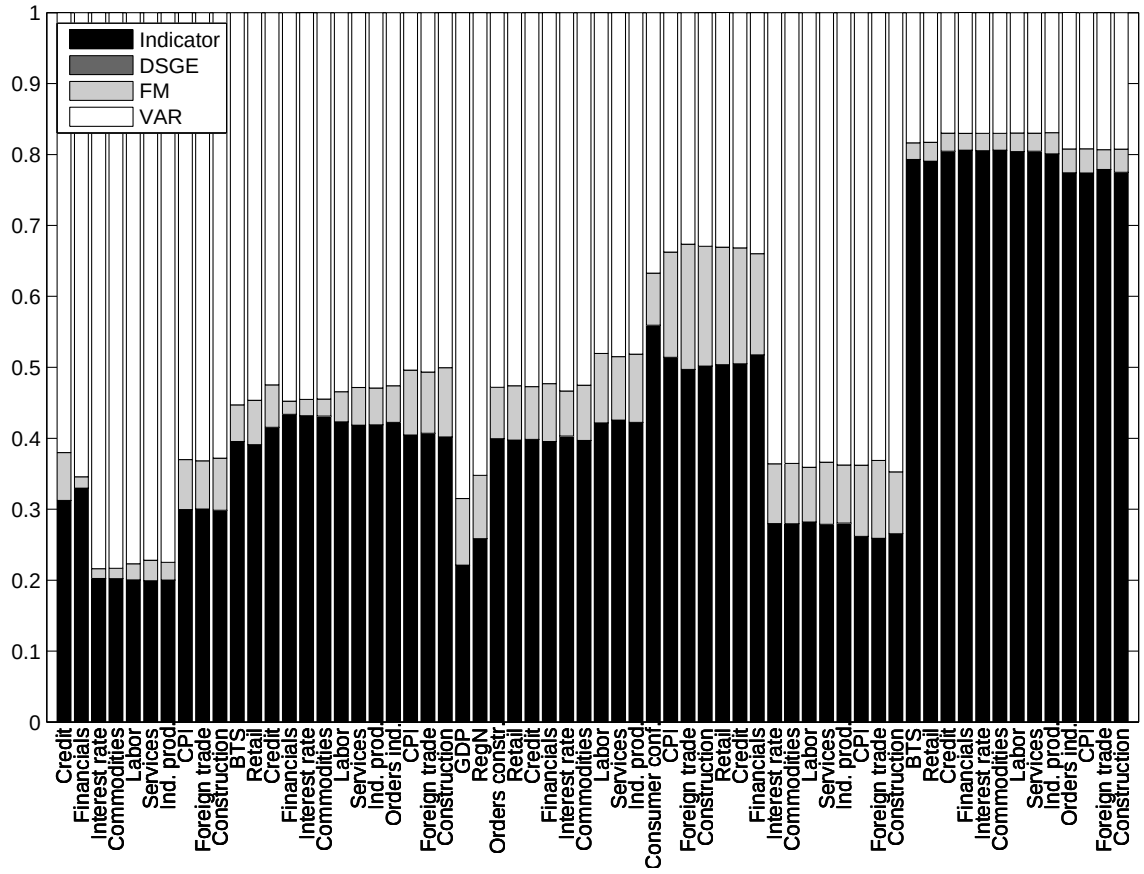


Figure 5. *Weights attached to the different model classes after different block releases. Evaluated against 5th release of data*

While figure 3 shows model performance in terms of point forecasts, figure 4 evaluate the density forecasts. The figure depicts the average log scores for the combined model as well as the four model classes from the different data blocks over the evaluation period. As for the RMSE, the forecasting performance improves when new information becomes available. The log score of the predictive densities for the combined model and all four model classes increases monotonically as new information arrives during the quarter. The large improvement for the density nowcasts due to the release of the business tendency survey towards the end of the period is evident from the dotted red line and the solid orange line. Further, we note that the factor model and VARs also contribute to improvements. However, the factor model performs substantially worse in terms of density forecasts than in terms on point forecast, see also figure A.2 in the appendix. In figure 5, we depict the weights attached to each model

class for the combined density forecast. Interestingly, the figure shows that at the start of the quarter the VARs is the model class that has the largest weight. As we move further into the quarter and more data are released, both the factor model and the leading indicator models increase their weight. Towards the end, the leading indictor models is the model class that has the highest weight. Note the remarkable change in the weights after the business tendency survey is released approximately one month before the first estimate of GDP is available. Furthermore, the performance of the DSGE model in terms of log score is poor. In fact, it ends up having no weight in the combined forecast. The poor performance of the DSGE model is consistent with findings in Bache, Jore, Mitchell, and Vahey (2009) and Amisano and Geweke (2009).

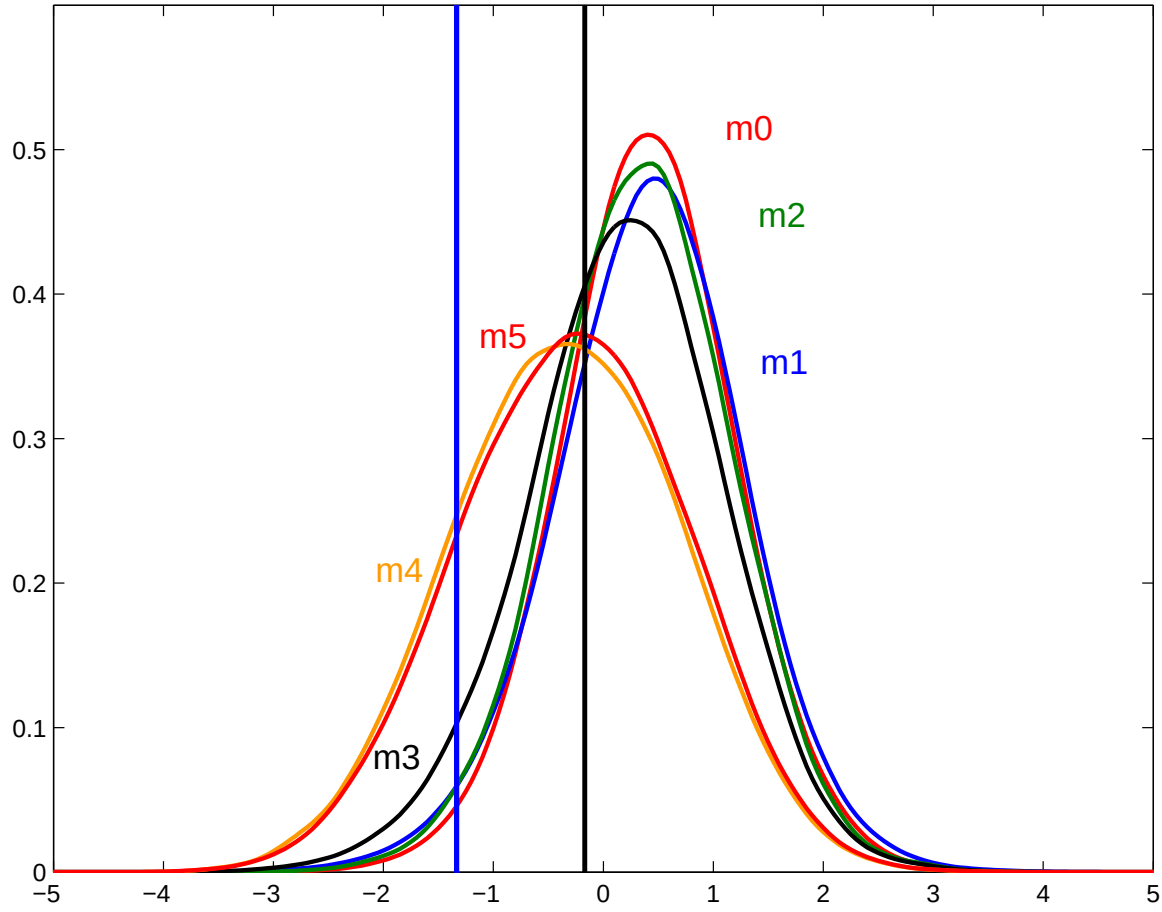


Figure 6. *Density forecast for Nowegian Mainland-GDP at different points in time. 2008Q4. Outturn 1st release (black line) and 5th release (blue line)*

Finally, our empirical exercise and the density forecast evaluation can be illustrated by

showing how the predictive density for the combined forecast changes as new information becomes available for the specific quarter of 2008q4. This is shown in figure 6. It is the first quarter of the financial crisis and is clearly a turning point in the Norwegian economy. $M0$ denotes the end of September 2008, $M1$ the end of October 2008, $M2$ the end November and so on. $M5$ denotes the day before the first release of GDP for 2008q4. The blue vertical line depicts the fifth release of GDP, while the black vertical line depicts the first release of GDP. The figure shows a clear improvement (in terms of increased log-score) for the predictive density at $M4$. This corresponds to the end of January 2009. This is related to the release of the quarterly business tendency survey. Note that there is also a clear improvement at $M3$. At this point, the factor model is already using some information from the two first months of the concurrent quarter.

5.2 Testing the pits

We evaluate the predictive densities relative to the "true" but unobserved density using the pits of the realization of the variable with respect to the nowcast densities, see figure 7. Table 2 shows the p-values for four different test applied to all the four model classes and the combined forecast at six different points in time ($M0 - M5$), where P-values equal to or higher than 0.05 mean that we can not reject the hypothesis that the combination is correctly calibrated at a 95% significance level.

The predictive densities from the combined forecast passes all tests for horizons $M0$ and $M1$. This is the case where the nowcast corresponds to a two-step ahead forecast. Turning to the one-step ahead forecast ($M2 - M5$), the predictive densities from the combined forecast also seems to be well-calibrated. We cannot reject the null hypothesis that the combination is well-calibrated at a 95% significance level from the Ljung Box test, the Anderson-Darling test and the Person chi-squared test.[10] Interestingly, the predictive densities from the combined forecast seems to be better calibrated than the predictive densities from all of the four different model classes, except the VARs. Especially the DSGE model seems to be poorly calibrated. However, note that our evaluation period is very short.

[10]The null hypothesis in the Berkowitz test is rejected.

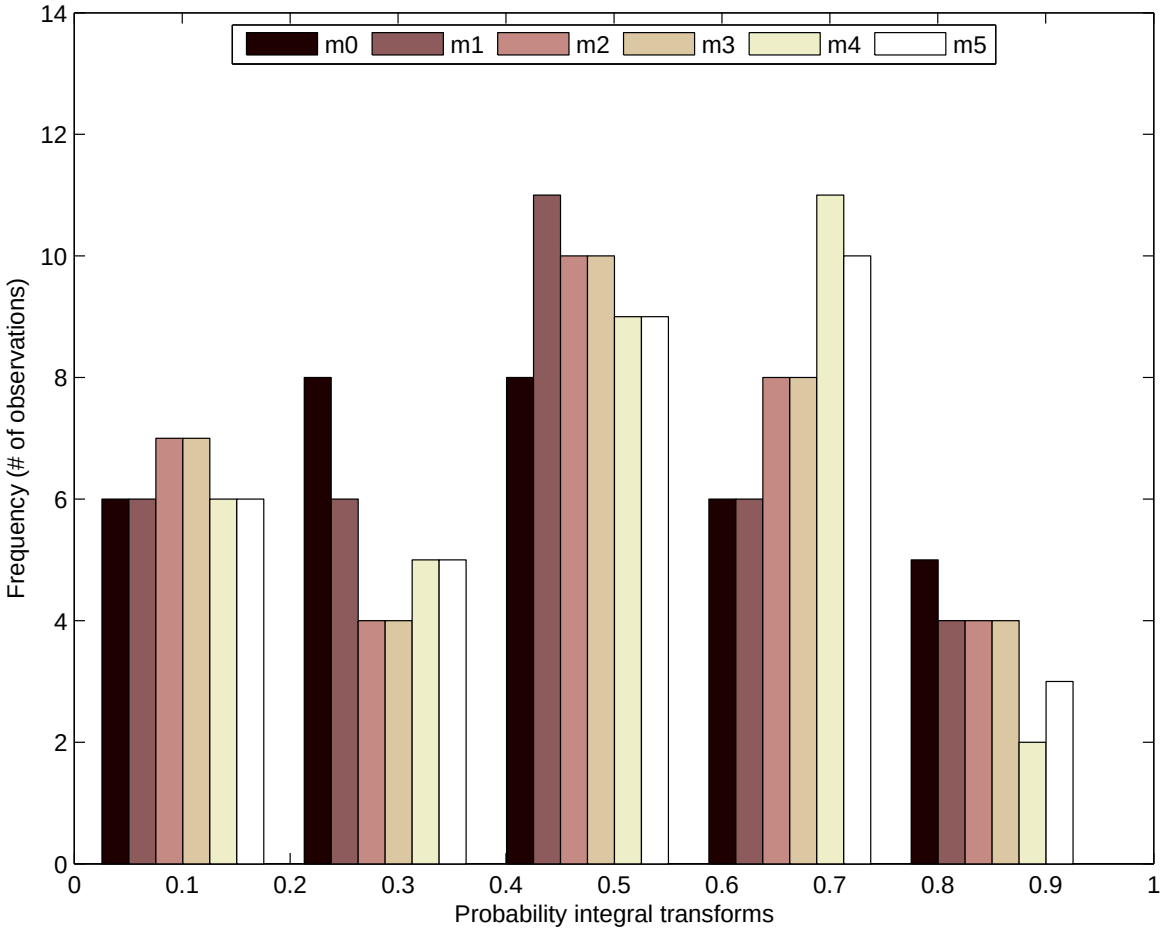


Figure 7. *The pits are the ex ante inverse predictive cumulative distribution evaluated at the ex post actual observations. The pits of a forecasting model should have a standard uniform distribution if the model is correctly specified.*

5.3 Robustness

Comment: See figure A.3 and figure A.4 in the appendix. More to be added

Table 2. Pits tests for evaluating density forecasts for GDP (p-values)

		LogScore	Berkowitz	χ^2	Ljung-Box	Anderson-Darling
m0 nowcast	Indicator	-1.19	0.04	**0.17**	**0.44**	**0.19**
	DSGE	-1.65	0.00	0.00	**0.34**	0.01
	FM	-1.23	0.03	**0.11**	**0.24**	**0.26**
	VAR	-1.17	**0.24**	**0.54**	**0.39**	**0.41**
	Combination	-1.17	**0.20**	**0.59**	**0.38**	**0.42**
m1 nowcast	Indicator	-1.17	0.03	**0.12**	**0.46**	**0.20**
	DSGE	-1.65	0.00	0.00	**0.34**	0.01
	FM	-1.23	0.01	**0.17**	**0.24**	**0.21**
	VAR	-1.16	**0.30**	**0.82**	**0.40**	**0.41**
	Combination	-1.15	**0.20**	**0.59**	**0.39**	**0.43**
m2 nowcast	Indicator	-1.14	0.01	**0.18**	**0.41**	**0.14**
	DSGE	-1.60	0.00	0.00	**0.97**	0.01
	FM	-1.19	0.00	0.03	**0.21**	**0.14**
	VAR	-1.13	**0.07**	**0.49**	**0.42**	**0.23**
	Combination	-1.13	0.02	**0.20**	**0.41**	**0.19**
m3 nowcast	Indicator	-1.12	0.01	**0.44**	**0.37**	**0.13**
	DSGE	-1.60	0.00	0.00	**0.97**	0.01
	FM	-1.16	0.00	0.02	**0.34**	**0.08**
	VAR	-1.13	**0.07**	**0.49**	**0.42**	**0.22**
	Combination	-1.11	0.01	**0.20**	**0.40**	**0.16**
m4 backcast	Indicator	-1.05	0.00	**0.07**	**0.69**	**0.06**
	DSGE	-1.60	0.00	0.00	**0.97**	0.01
	FM	-1.15	0.00	0.04	**0.40**	**0.09**
	VAR	-1.09	0.04	**0.30**	**0.37**	**0.26**
	Combination	-1.05	0.00	**0.06**	**0.51**	**0.09**
m5 backcast	Indicator	-1.05	0.00	**0.07**	**0.67**	**0.06**
	DSGE	-1.60	0.00	0.00	**0.97**	0.01
	FM	-1.15	0.00	0.03	**0.44**	**0.07**
	VAR	-1.09	0.04	**0.30**	**0.37**	**0.27**
	Combination	-1.06	0.00	**0.08**	**0.49**	**0.09**

Note: The null hypothesis in the Berkowitz test is that the inverse normal cumulative distribution function transformed pits are identically, normally distributed, N(0,1). χ^2 is the Pearson chi-squared test suggested by Wallis (2003) of uniformity of the pits histogram in eigth equiprobable classes. Ljung-Box is a test for independence of the pits at lags greater than or equal to the horizon. The Anderson-Darling test is a test for uniformity of the pits, with the small-sample (simulated) p-values computed assuming independence of the pits.

6 Conclusion

In this paper we have used an expert combination framework to produce density combination nowcasts for Norwegian Mainland-GDP from a system four different model classes widely used at central banks; VARs, leading indicator models, factor models and a DSGE model. The density nowcasts are combined in a two-step procedure. In the first step, we group models into different model classes. The nowcasts for each model within a model class are combined using the logarithmic score. This yields a combined predictive density nowcast for each of the four different model classes. In a second step, these four predictive densities are combined into a new density nowcast using the logarithmic score.

The density nowcasts are updated for every new data release during a quarter and until the first release of GDP is available. Our recursive nowcasting exercise is applied to Norwegian real-time vintage data and evaluated on the period 2001q2-2009q1. First, we show that the logarithmic score for the predictive densities for Norwegian Mainland-GDP increases monotonically as new information arrives during the quarter. In particular, releases of indicator data are important and improve the predictive densities the most. Second, our results illustrates that the weights attached to the four different model classes change during the quarter in correspondence with new data releases. In this way, our combination procedure attaches a higher weight to models with new and relevant information. Finally, we show that the predictive densities from our density combination approach is well-calibrated throughout the quarter, while this is not the case for all of the four model classes.

References

Aastveit, K. A., and T. G. Trovik (2007): "Nowcasting Norwegian GDP: The Role of Asset Prices in a Small Open Economy," Working Paper 2007/9, Norges Bank.

Aastveit, K. A., and T. G. Trovik (2008): "Estimating the Output Gap in Real Time: A Factor Model Approach," Working Paper 2008/23, Norges Bank.

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

Amisano, G., and J. Geweke (2009): "Optimal Prediction Pools," Working Paper Series 1017, European Central Bank.

Bache, I. W., A. S. Jore, J. Mitchell, and S. P. Vahey (2009): "Combining VAR and DSGE forecast densities," Working Paper 2009/23, Norges Bank.

Banbura, M., and M. Modugno (2010): "Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data," Working Paper Series 1189, European Central Bank.

Banbura, M., and G. Rünstler (2010): "A look into the factor model black box - publication lags and the role of hard and soft data in forecasting GDP," *International Journal of Forecasting*, forthcoming.

Banerjee, A., and M. Marcellino (2006): "Are there any reliable leading indicators for US inflation and GDP growth?," *International Journal of Forecasting*, 22(1), 137–151.

Banerjee, A., M. Marcellino, and I. Masten (2005): "Leading Indicators for Euro-area Inflation and GDP Growth," *Oxford Bulletin of Economics and Statistics*, 67(s1), 785–813.

Bates, J., and C. Granger (1969): "The combination of forecasts," *Operations Research Quarterly*, 20(4), 451–468.

Berkowitz, J. (2001): "Testing Density Forecasts, With Applications to Risk Management," *Journal of Business and Economic Statistics*, 19(4), 465–474.

Bjørnland, H. C., K. Gerdrup, A. S. Jore, C. Smith, and L. A. Thorsrud (2010): "Weights and Pools for a Norwegian Density Combination," *North American Journal of Economic and Finance*, forthcoming.

Brekke, H., and K. W. Halvorsen (2009): "Norges Bank's regional network: fresh and useful information," Economic Bulletin 2, Norges Bank.

Brubakk, L., T. A. Husebø, J. Maih, K. Olsen, and M. Østnor (2006): "Finding NEMO: Documentation of the Norwegian economy model," Staff Memo 2006/6, Norges Bank.

Clark, T. E., and M. W. McCracken (2010): "Averaging forecasts from VARs with uncertain instabilities," *Journal of Applied Econometrics*, 25(1), 5–29.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39(4), 863–83.

Evans, M. D. (2005): "Where Are We Now? Real-Time Estimates of the Macro Economy," *International Journal of Central Banking*, 1(2), 127–175.

Frale, C., M. Marcellino, G. L. Mazzi, and T. Proietti (2010): "Survey data as coincident or leading indicators," *Journal of Forecasting*, 29(1-2), 109–131.

Gerdrup, K. R., A. S. Jore, C. Smith, and L. A. Thorsrud (2009): "Evaluating ensemble density combination - forecasting GDP and inflation," Working Paper 2009/19, Norges Bank.

Geweke, J., and G. Amisano (2010): "Comparing and evaluating Bayesian predictive distributions of asset returns," *International Journal of Forecasting*, 26(2), 216–230.

Giannone, D., L. Reichlin, and D. Small (2008): "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, 55(4), 665–676.

Hall, S. G., and J. Mitchell (2007): "Combining density forecasts," *International Journal of Forecasting*, 23(1), 1–13.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14(4), 382–417.

Jore, A. S., J. Mitchell, and S. P. Vahey (2010): "Combining forecast densities from VARs with uncertain instabilities," *Journal of Applied Econometrics*, 25(4), 621–634.

Kascha, C., and F. Ravazzolo (2010): "Combining inflation density forecasts," *Journal of Forecasting*, 29(1-2), 231–250.

Marcellino, M. (2006): "Leading Indicators," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann, vol. 1, pp. 879–960. Elsevier, Amsterdam.

Mitchell, J., and S. G. Hall (2005): "Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR 'Fan' Charts of Inflation," *Oxford Bulletin of Economics and Statistics*, 67(s1), 995–1033.

Mitchell, J., and K. Wallis (2010): "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness," *Journal of Applied Econometrics*, forthcoming.

Noceti, P., J. Smith, and S. Hodges (2003): "An Evaluation of Tests of Distributional Forecasts," *Journal of Forecasting*, 22(6-7), 447–455.

Schumacher, C., and J. Breitung (2008): "Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data," *International Journal of Forecasting*, 24(3), 386–398.

Stark, T., and D. Croushore (2002): "Forecasting with a real-time data set for macroeconomists," *Journal of Macroeconomics*, 24(4), 507–531.

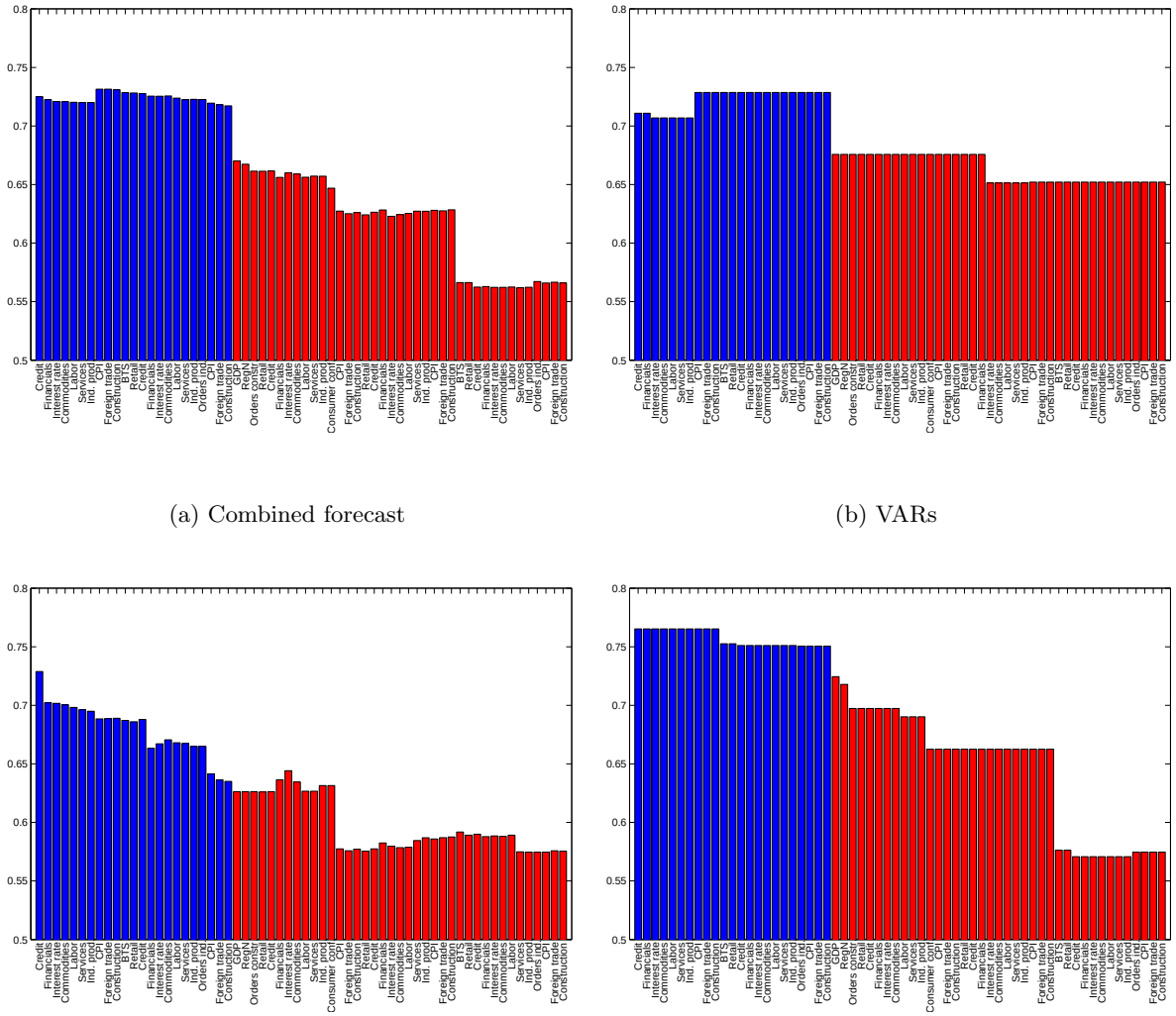
Stock, J. H., and M. W. Watson (2004): "Combining forecasts of output growth in seven-country data set," *Journal of Forecasting*, 23, 405–430.

Timmermann, A. (2006): "Forecast Combinations," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann, vol. 1, pp. 136–96. Elsevier, Amsterdam.

Wallis, K. F. (2003): "Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts," *International Journal of Forecasting*, 19(3), 165–175.

——— (2010): "Combining forecasts - forty years later," *Applied Financial Economics*, forthcoming.

Appendix



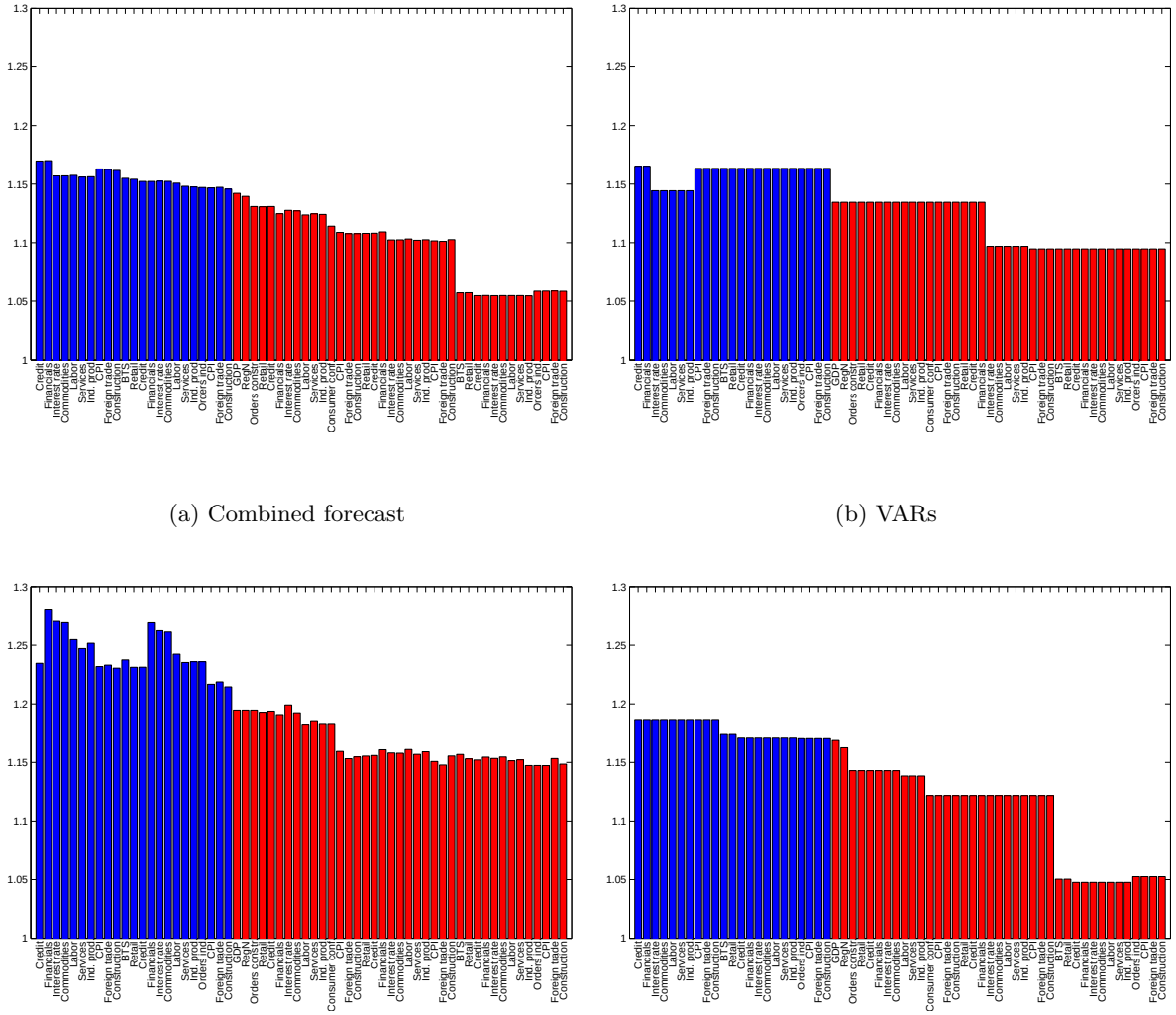
(a) Combined forecast

(b) VARs

(c) FM

(d) Indicator

Figure A.1. Mean square forecasting errors made after different block releases. Evaluated against 5th release of data.



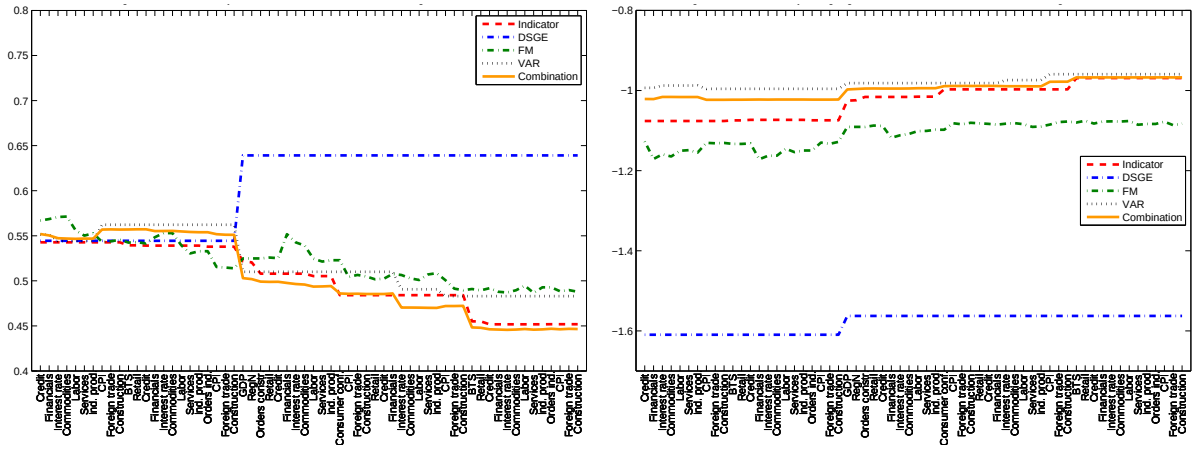
(a) Combined forecast

(b) VARs

(c) FM

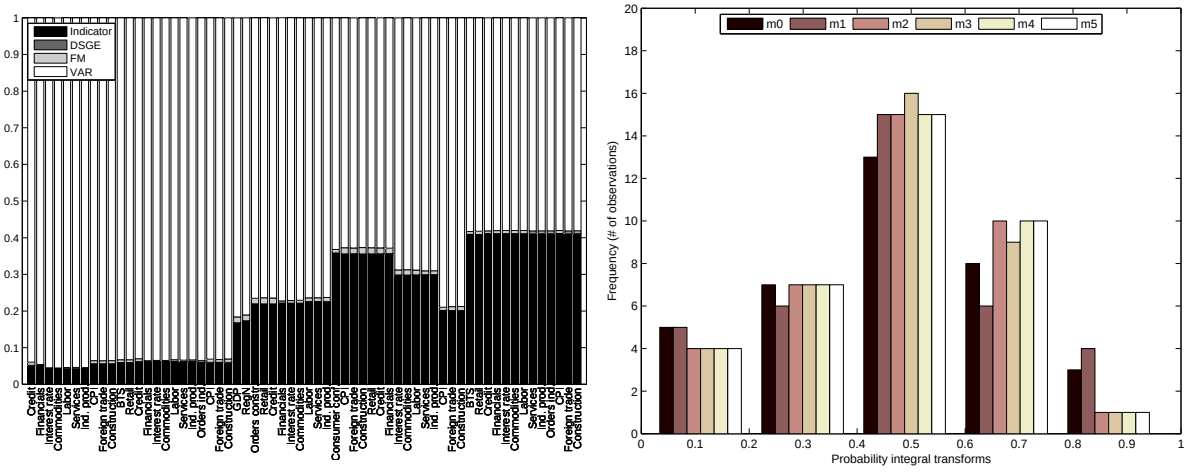
(d) Indicator

Figure A.2. Average log scores (inverted) for forecasts after different block releases. Evaluated against 5th release of data.



(a) RMSE

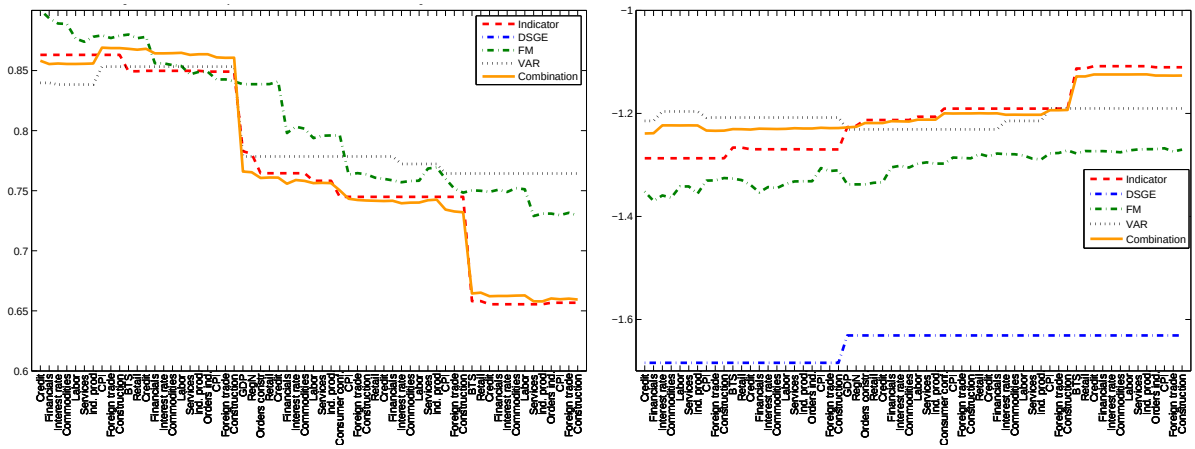
(b) Average log score



(c) Weights

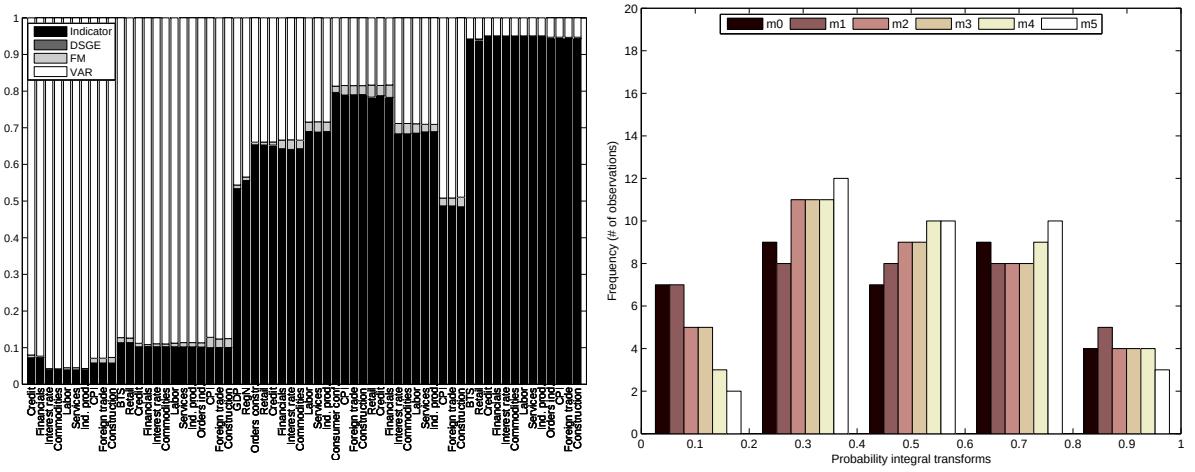
(d) Pits

Figure A.3. Robustness of results when evaluated against 1st release of data.



(a) RMSE

(b) Average log score



(c) Weights

(d) Pits

Figure A.4. Robustness of results when evaluated against final vintage of data.