Maastricht University

School of Business and Economics

Combining Professional and Survey Forecasts for Macroeconomic Data

Antoine Soetewey
I6083256

Thesis submitted to obtain the degree of

MASTER OF SCIENCE IN ECONOMETRICS AND OPERATIONS RESEARCH
Specialisation in Econometrics

Supervisor: Dr. Nalan Baştürk
Second Reader: Dr. Stephan Smeekes

Academic year: 2015-2016

Abstract

This thesis investigates the combination of survey forecasts and uses data on US GDP growth to determine whether we can benefit from combining forecasts. Two main findings arise from the analysis. First, the results show that the sole combination of survey forecasts outperforms the combination of survey forecasts with more conventional time series models forecasts. Second, we find that combining the Survey of Professional Forecasters and the Greenbook survey forecasts yields lower RMSE at all but one horizon from nowcasts to four quarters ahead predictions. In particular, we show that the Bayesian model averaging combination is preferred for nowcasts. The simple equal-weighted average combination dominates for two and three quarters ahead predictions. Lastly, the predictive least square combination is superior for four quarters ahead forecasts.

Acknowledgments

I owe this Master's thesis to the help and support of many kind people around me, to only some of whom it is possible to give particular mention here.

Above all, this thesis would not have been possible without the support of my principal supervisor, Dr. Nalan Baştürk. I thank her for her trust and giving me the opportunity to write about this topic. Furthermore, I thank her for her time through numerous meetings throughout the year. I am also grateful for her thoughtful and valuable comments on the penultimate version. She considerably improved the quality of the final version.

I express my gratitude to the co-reader of this thesis, Dr. Stephan Smeekes.

I thank my father; Christian Soetewey, and my friend; Floriane Dierckx for reviewing the penultimate version.

I hope to return all these favors someday.

For any errors or inadequacies that may remain in this work, of course, the responsibility is entirely my own.

Antoine Soetewey

Maastricht, The Netherlands

August 22, 2016

Contents

Abstract **iii**

Acknowledgments **iv**

Contents **v**

1 Introduction **1**

2 Literature review **6**

2.1 Hypotheses 13

3 US GDP data **15**

3.1 Samples selection 15

4 Models **25**

4.1 Evaluation tool 27

5 Forecast combination methods **29**

5.1 Equal-weighted average and median forecasts 29

5.1.1 Equal-weighted average 29
5.1.2 Median 31
5.2 Inverse MSE-weighted average forecast 31
5.3 Predictive least squares and adaptive expectations 32
5.3.1 Predictive least squares 32
5.3.2 Adaptive expectations 33
5.4 Bayesian model averaging 35

6 Results **38**
6.1 Equal-weighted average and median forecasts 38
6.2 Inverse MSE-weighted average forecast 42
6.3 Predictive least squares and adaptive expectations 44
6.3.1 Predictive least squares 44
6.3.2 Adaptive expectations 46
6.4 Bayesian model averaging 48
6.5 Comparison of combination methods 49

7 Conclusion **57**

References **59**

Appendix **69**

Section 1

Introduction

Prediction is of great interest in economics. Predicting future outcomes helps to take better decisions, to tackle threats and to reduce risks before issues even arise, or it can also simply reassure people by making them feel less uncertain about the future. Indeed, a considerable amount of information for any economic agent is found in expectations about future economic developments (Garcia, 2003). In this sense, several institutions provide forecasts of macroeconomic indicators based on professional forecasters and opinion surveys, referred as professional or survey forecasts. In order to issue these kind of forecasts, the institutions send a questionnaire every quarter to all panel members, and forecasters actively participate in the forecasting survey by returning the questionnaire. The questionnaire includes the main macroeconomic variables such as real GDP and its components, nominal GDP, measures of inflation, unemployment, industrial production, housing, etc. A set of the most recent data of the main economic variables is attached to the questionnaire so that all participants have access to the same information. After the questionnaires are collected and tabulated, the results and a summary of the economic outlook are released to

the press and published (Su and Su, 1975). These survey forecasts are found to be useful for forecasting, particularly for US inflation, see *e.g.* Faust and Wright (2013) and Baştürk et al. (2014).

Although the provided forecasts from different institutions are substantially different, forecast accuracy is of fundamental importance. Forecast users rely on forecast accuracy as forecasts will guide their decisions. Forecast producers' reputation is built upon the accuracy of their forecasts. Economists who are interested in differentiating forecasters with competing hypotheses reckon on forecast accuracy as well (Diebold and Mariano, 2012). The purpose of this research is to effectively combine the different forecasts of a macroeconomic variable and illustrate whether forecast combinations can provide useful and unique information. The topic is related to the superior performance of forecast and model combinations in different contexts. The analysis is done by proposing a rigorous empirical analysis showing properties of combined forecasts.

The present thesis intends to answer the following questions: Are different sources of GDP survey forecasts comparable? Do the survey forecasts of GDP outperform conventional time series models in forecasting? Can we benefit from combining GDP forecasts from different sources, such as the survey forecasts and model based forecasts? Which combination methods can be used to combine GDP forecasts from different sources, and how do different combination methods perform in forecasting GDP growth?

The macroeconomic variable that has been selected for this analysis is the growth of the US Gross Domestic Product (GDP).[1] The choice of this variable is not arbitrary. First, the GDP is probably among the most documented and most important

[1] As detailed in Section 3, the choice was limited by data availability.

variables when classifying countries. Second, the growth-transformation of a series removes the trend, which is said to be easier to forecast. Third, studies in GDP forecast combinations is rather rare in the literature relative to studies in inflation forecast combinations. This is a major contribution of the thesis.

Results of the analysis are (potentially)[2] interesting for central banks, public institutions, policy makers, privates businesses, investors, individual households, and any parties interested in GDP forecasting. More precisely, central banks, public institutions and policy makers take into account expectations regarding macroeconomic variables for decisions related to interest rates, monetary policy and price stability. Any decisions concerning interest rates hinge on a scrupulous assessment of the macroeconomic outlook. Monetary policy decisions can affect the economy for a long time and the impact on the economy in general and on price developments in particular is uncertain. For this purpose, policy makers who rely on measures of economic activity require reliable forecasts. With more reliable forecasts, policy makers can adapt more accurately their policy decisions, and thus more easily meet their objectives (Tkacz, 2001). Similarly, managers' choice to expand or recruit more employees is partly based on expectations of future demand for their products and services. Individual households and investors depend on the economic outlook to determine their preferences concerning savings, consumptions and investments. Future economic developments have actually a significant impact on a majority of economic decisions (Garcia, 2003).

The goal of the present thesis is twofold. I will first use available historical data on US GDP growth to provide my own set of forecasts with simple autoregressive integrated moving average (ARIMA) models. I will then evaluate my forecasts against the individual forecasts provided by the Survey of Professional Forecasters (SPF)

[2]It could be that interested parties are not interested in having accurate estimates, but are

and Greenbook data sets.[3]

In a second time - this will actually be the main part of the thesis -, I will combine the different data sets (*i.e.*, (i) own forecasts from ARIMA models, (ii) SPF and (iii) Greenbook) and evaluate the combined forecasts against the true observations to measure whether forecast combinations is prone to more accurate estimates than the ones obtained for each of the individual underlying forecasts.

In this thesis, two types of forecasts will be analyzed; "nowcasts" and pure forecasts. A nowcast can be seen as an estimate of the current situation, or a prediction of the present (or sometimes very near future), while a forecast is the usual prediction of an occurrence or event multiple periods in advance. Nowcasts are used to assess the current state of an economy and they are particularly interesting for GDP as it is often determined after a long delay[4] and even subject to many revisions. They are important for policy makers as well as financial institutions as they allow to estimate the direction of change in GDP before GDP itself is published. Moreover, it has recently become a popular measure for investors seeking to interpret the *latest* fluctuations and for governments and central banks seeking to quantify the expectations about the economic activity in the major economies and take appropriate monetary policy decisions. Importance of nowcasts are presented in Baffigi et al. (2002), Banbura et al. (2013) and Giannone et al. (2008). The usual components of nowcasts are, among others, figures about employment, trade balance, industrial production, expenditures. Throughout this thesis, nowcasts correspond to forecasts at horizon zero; $h = 0$.

---

rather interested in observing people's *expectations*. In the latter case, they would not care about results on forecast combination.

[3]See Section 3 for a detailed description of the data sets used for this analysis.

[4]Data on GDP are available with a delay from 30 days after the end of the quarter in the USA to 70 days for European countries (Golinelli and Parigi, 2008, p.2).

The plan for the remainder of the present thesis is as follows. Section 2 contains a review of the background on forecast combination and hypotheses drawn from this non-exhaustive literature review. The data used for the analysis are presented in Section 3. Section 4 discusses the models and the type of forecast for the forecasting exercise. Section 5 describes the forecast combination methods, including the evaluation tools used to measure forecast accuracy. Section 6 presents and analyzes the results. In this section, different forecasts from surveys, ARIMA models, and combinations of these models using different methodologies are compared. In addition, nowcasts and forecasts for different horizons are considered. Finally, Section 7 concludes.

Section 2

Literature review

Merits of forecast combination were introduced by the seminal paper by Bates and Granger (1969). They showed that, when combining two separate sets of forecasts of airline passenger data to form a composite set of forecasts, the composite set can yield lower mean-squared error than either of the original forecasts. More generally, they argue that when more than one approach together with a record of their past performance are available, the weights attached to each individual forecast should be set so that most weight is granted to the forecast that has performed best in the past.

Since then, much work has been undertaken in this domain and the existing evidence suggests that "The results have been virtually unanimous: combining multiple forecasts leads to increased forecast accuracy [...]." (Clemen, 1989, p. 559). More recently, Makridakis and Hibon (2000) confirm Clemen's statement by concluding that (p. 458) "The accuracy of the combination of various methods outperforms, on average, the specific methods being combined and does well in comparison with other methods."

It seems therefore appropriate to start this thesis by a general review of forecast combinations, which clarifies the potential channels involved, and to look at information and predictions provided by the literature.

As Timmermann (2006) summarizes, the advantage of combining forecasts is fourfold.

First, given that the information set underlying the individual forecasts is usually unknown to the forecast user,[5] it is not possible to pool this underlying information and specify a super model that would nest all the information sets. For instance, suppose that we want to predict a variable Y and that we are given two forecasts, $\hat{Y}_a$ and $\hat{Y}_b$. Let the forecast $\hat{Y}_a$ be built upon the variables X_a, X_b and X_c, *i.e.*, $\hat{Y}_a = g_a(X_a, X_b, X_c)$, and the forecast $\hat{Y}_b$ be based on the variables X_x, X_y and X_z, *i.e.*, $\hat{Y}_b = g_b(X_x, X_y, X_z)$. If all underlying variables $\{X_a, X_b, X_c, X_x, X_y, X_z\}$ were observable, we would obviously build a forecast $\hat{Y}_c$ based on all six variables, $\hat{Y}_c = g_c(X_a, X_b, X_c, X_x, X_y, X_z)$. However, as mentioned above the underlying variables are unobserved while the forecasts are available to the forecast user. Hence, the only alternative is to disregard the underlying information sets and combine those available forecasts, that is, to draw a model of the type $\hat{Y}_c = g_c(\hat{Y}_a, \hat{Y}_b)$ (this argument was originally provided by Bates and Granger (1969)). In line with this argument, the higher the overlap in the underlying forecasters' information set, the less efficient

[5]The underlying information set is unobserved to the forecast user in the cases of subjective forecasts (also referred as judgmental forecasts, *i.e.*, forecasts based on judgments, qualitative information, etc.). On the other hand, in the cases of objective forecasts (based on quantitative data and using well-defined econometric models) the information set is most of the time observable to the forecast user. However, as the present thesis is limited to survey and professional forecasts which are by definition at least a mix of subjective and objective forecasts, we can reasonably assume that the underlying set of information is (partially) unobserved to the forecast users. See, for instance, Reifschneider et al. (1997) who employ a mix of models and judgment guided by information not available to models for the Fed staff's Greenbook forecast.

a forecast combination is likely to be (Clemen, 1987).

Second, as argued by, among others, Figlewski and Urich (1983), Kang (1986), Diebold and Pauly (1987), Makridakis (1989), Sessions and Chatterjee (1989), Winkler (1989), Clements and Hendry (2002) and Aiolfi and Timmermann (2006), non-stationarities may have different impacts on individual forecasts. Combining individual forecasts can thus significantly reduce forecasting errors. Indeed, an unexpected shift in a time series can have a very different impact among individual forecasts due to the degree of adaptability of the underlying models. Some models will adapt more quickly and be affected over a shorter period of time than others. Therefore, on the one hand, if the time spent since the most recent shift is rather long, the slow adapting models are expected to produce the best forecasting performance relative to fast adapting models as they are more stable. On the other hand, the shorter the time spent since the most recent shift, the better one might expect fast adapting models to outperform. Shifts in time series, also referred as structural breaks, are often caused by major events such as wars, institutional changes, political instabilities, epidemics, crises, technological developments, etc. Such structural breaks are typically difficult to distinguish *ex ante* or in real time and even harder to predict. Therefore, combining forecasts from models with different degrees of adaptability will, on average, outperform individual forecasts.

A third argument for using forecast combinations referred to by Clemen (1989), Makridakis (1989), Diebold and Lopez (1996) and Stock and Watson (2001, 2004) is that individual forecasting models are likely to have a misspecification bias. Given that the data generating process is more complex than even the most advanced model specified by any forecaster and that forecasting models are only approximations, it is rather improbable that a forecast dominates all other forecasts *at all times*. The best model would be the one that can vary over time and adapt itself to the complex data,

something that is rather arduous to achieve based on past forecasting performance. Combining forecasts established with different models is a way to make the aggregate forecast robust to misspecifications and measurement errors present in the underlying data sets by ways of diversification.[6]

The last reason for combining forecasts is that individual forecasts might be subject to different loss functions (Zellner, 1986). For instance, assume that forecaster 1 wants to avoid large positive forecast errors and forecaster 2 wants to avoid large negative forecast errors. To reduce the likelihood of having adverse outcomes, forecaster 1 will tend to over-predict the variable of interest (so the forecast error distribution is centered on a negative value), while forecaster 2 will tend to under-predict it (so the forecast error distribution is centered on a positive value). If the direction of the asymmetry was known to the forecast user, these biases could easily be removed by adding a constant in the combination equation. This is, however, often not the case. In the latter case, a forecast user looking for a more symmetric forecast error distribution would be better off combining results from the two forecasters.

We expect these four advantages of forecast combination to exist specifically for GDP. Regarding the first argument, since the underlying information used by professional forecasters to build GDP forecasts (*e.g.*, expectations on GDP components, mix of private and public information, association of beliefs and knowledge about the economic activity, etc.) is unobservable, it is indeed not possible to specify a super model nesting directly all the underlying information sets. The only solution is to combine GDP forecasts made by professional forecasters, which indirectly contain and reflect the primary source of information. Concerning the second argument, GDP is expected to endure structural breaks and non-stationarities over time. Al-

[6]Note the similitude with the portfolio diversification strategy that will, on average, yield higher returns and lower risk than any individual investment found within the portfolio.

though it is hard to know precisely when it will occur, shifts in an economy which produce shifts in GDP growth can happen. Therefore, combining GDP forecasts build upon models with different degrees of adaptability are expected to outperform individual forecasts. The third argument is not an exception; GDP forecasting models are only approximations of the complex data generating process and are thus not immune to misspecification bias. Combining several GDP forecasting models increases the robustness of the aggregate forecast through diversification. Finally, the last argument for combining forecasts is also expected to exist specifically for GDP. Indeed, it might be the case that some forecasters prefer to under-predict GDP to be prudent with their expectations or to consider the worst case scenario, and that others would rather over-predict GDP in order to suggest a confident growth in GDP and convey a bullish sentiment about the level of the overall economy. This will most likely be unknown so combining GDP forecasts will be favorable to forecast users looking for accurate estimates.

Although combining forecasts seems a promising avenue to decrease forecast errors and obtain predictions of the variable of interest closer to its actual value, there also exists, as highlighted by Timmermann (2006), arguments against this technique.

The first and main criticism of combining forecasts follows from the first advantage detailed *supra*. It is true that if the information set underlying the individual forecasts is private and unobservable to the forecast user, combining forecasts could definitely add value (as it still allows to gather a variety of information sets when they are unobserved). Nonetheless, in situations where all information sets underlying the individual forecasts are observable, searching for a single best super model embedding all information sets could prove to be more efficient and accurate than using a combination of forecasts (Chong and Hendry, 1986; Diebold, 1989).

Second, Diebold and Pauly (1987), Elliott (2004) and Yang (2004) point out that estimations errors often lead to incorrect weights assigned to each of the individual forecasts when using a combination strategy.

The last disadvantage, referred to by Clemen and Winkler (1986), Diebold and Pauly (1987), Figlewski and Urich (1983), Kang (1986) and Palm and Zellner (1992) follows from the second advantage detailed *supra*. Although non-stationarities in the underlying data generating process can actually be the reason to combine forecasts, it is also known to lead to mediocre results due to instabilities in the combination weights.

The arguments against combining forecasts have not, however, dampened the empirical research, as illustrated by the considerable literature that has accumulated over the years. Summarizing theoretical and applied contributions from the forecasting, psychology, statistics, and management science fields, Clemen (1989) reviews the entire literature on the topic up to the date of the article. Applied results from forecast combinations and related to the subject of this thesis are presented.

Agnew (1985) showed that, using forecasts from the Blue Chip Economic Indicators, the sequential Bayesian weighting combination outperformed either the average or the median of the underlying forecasts. Armstrong (1984) reviewed 25 years of research (1960-1984) and found that sophisticated extrapolation methods[7] have had a negligible payoff for accuracy in forecasting. Indeed, out of 39 empirical studies, 28 showed evidence that sophisticated techniques did not perform better than simpler techniques. For example, Aiolfi et al. (2010) analyzed 14 macroeconomic variables including real GDP and suggested that a simple equal-weighted average of survey

[7]Extrapolation methods are defined as forecasting techniques that rely solely on historical data from the series to forecast future outcomes.

forecasts outperforms the best model-based forecasts for a majority of them. Along the same lines, Clemen and Winkler (1986), Sessions and Chatterjee (1989) and Zarnowitz et al. (1967) analyzed real and nominal GNP data and reported that the equal-weighted average performed well. Winkler et al. (1977) found the same results and added the conclusion that the more forecasts included in the consensus, the better the forecasting performance. Furthermore, Armstrong (1986) stated that, based on eight empirical applications, combining forecasts yields a reduction in forecasting errors from 0% to 23.4%. Gunter and Aksu (1989) went even further as they introduced the idea of combining different types of forecast combinations. Using GNP data, the combination of combined forecasts yielded a slight improvement in forecasting performance. Last but not least, Su and Su (1975) evaluated the first seven years of the ASA-NBER survey forecasts (former name for SPF; before it was taken over by the Federal Reserve Bank of Philadelphia in 1990) relative to econometric and extrapolation forecasts. They concluded that the survey forecasts significantly outperformed the extrapolation forecasts. Moreover, also using the consensus forecasts from the ASA-NBER surveys, Zarnowitz (1984) found substantial performance improvements in combined forecasts.

On the other hand, Bischoff (1989) and Bohara et al. (1987) are among the few authors who showed that one of the individual forecasts outperformed the combined forecasts. Bischoff (1989) combined ARIMA and econometric forecasts of several macroeconomic variables and concluded that one of the individual forecasts was on average superior to the combined forecasts. Similarly, Bohara et al. (1987) found that a combination of forecasts can perform worse than an individual forecast when one forecast is much more precise than the others.

As one can see, the literature on forecast combination is vast. This is particularly the case for inflation forecasts which are analyzed in detail in the literature. However,

substantially less work has been undertaken in applying forecast combination to professional and survey forecasts, and studies in GDP forecast combinations is rather rare. A major contribution of the thesis is to fill this gap in the literature.

2.1 Hypotheses

Based on the literature review, we expect that combining forecasts from SPF and Greenbook will produce better estimates of the US GDP for all horizons. Furthermore, Ang et al. (2007); Croushore (2010); Faust and Wright (2009) and Faust and Wright (2013) argue that, when forecasting inflation, subjective forecasts perform best. However, it would be interesting to test if purely judgmental forecasts are indeed at the frontier of our forecasting ability when dealing with GDP, or if including objective forecasts reduces the forecasting errors. Although the ARIMA(p, d, q) models used for the forecasting exercise may not be the most advanced models, they usually perform reasonably well so we expect that the combination of survey/professional forecasts and objective forecasts will outperform the simple combination of survey/professional forecasts.

Furthermore, the literature seems to prone rather simple combination methods and states quite surprisingly that advanced combination techniques do not necessarily lead to better results. As a researcher, I however intend to verify this statement by applying several combination techniques, some rather simple and some more complex, to the data. Limiting ourselves to simple techniques solely because it yielded satisfying results in the past would indeed be too restrictive, and more importantly, would remove all incentives for good statisticians and forecasters to go beyond and develop the forecasting science.

One potential reason if these hypotheses are not confirmed is that historical survey

forecasts provided by institutions are already a combination (often mean, median or mode) of forecasts from a panel of experts. The real improvements of forecast combination relative to individual forecasts could in fact be undermined by this prior averaging of respondents' estimates. Reductions in prediction errors are, however, still to be expected since more information and several combination techniques are considered.

Section 3

US GDP data

3.1 Samples selection

Among the institutions that provide predictions of macroeconomic indicators from professional forecasters and opinion surveys, only a few of them make their forecast data publicly and freely available. Furthermore, not all institutions provide historical forecast data which is necessary to evaluate *ex post* the forecasts made by professional forecasters. Indeed, many institutions provide nowcasts and forecasts for subsequent horizons but they do not provide historical forecasts that were made in previous years. Therefore, the choice of professional forecasts for performance evaluation and evaluation of forecast combinations are constrained by data availability.

For the purpose of forecast combination, two different sets of data have been selected. First, we will use the data on US real GDP growth from the Survey of Professional Forecasters (SPF).[8] This data set is the oldest quarterly survey of macroeconomic forecasts in the United States. The survey began in the last quarter of 1968

[8]The full database is available via www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters.

and the last survey available at the time of writing the present thesis is dated the first quarter of 2016. The survey shows quarter-over-quarter growth rate forecasts for the quarter in which the survey is conducted and the following four quarters.

Second, the projections from the Greenbooks of the Federal Reserve Board of Governors will also be used for our analysis.[9] These projections are released to the public with a lag of five years, and this is the reason that the data ends in 2010. Unlike the SPF, the Greenbook is produced more than once per quarter. Furthermore, these forecasts are often incomplete as not all horizons are covered. Therefore, to avoid missing values as much as possible we retain only the projections from the last Greenbook of each quarter. The survey began in the first quarter of 1967 and the last survey available is dated the last quarter of 2010. For each Greenbook, the data set provides values for the nowcast quarter, and (at most) the following nine quarters. Note that we use the mean of the individual responses for both data sets. Other measures such as median or mode could potentially be an avenue for future research.

The two sources of survey forecasts have a rather similar structure, and the questionnaires sent each quarter to the panel members are globally asking for the same measures. The only slight difference lies in the respondents; SPF respondents are mostly economists, consultants, researchers and professors, while Greenbook respondents are the Research staff of the Federal Reserve Board of Governors. The two sources of survey forecasts are thus expected to produce similar results, while not being exactly the same as they reflect a mix of different methods/models and beliefs about the future from different people. Wang and Lee (2014) examine rationality of forecasts of the Greenbook and the SPF under asymmetric loss functions and

[9]The data set is accessible via www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data.

find that all their results similarly hold in the Greenbook and the SPF forecasts. One explanation for their findings is that SPF try to be in line with Greenbook; "as choosing to be in line with Fed's asymmetry preference will allow them [SPF forecasters who come largely from the business world or Wall Street] to set up their portfolio which will be benefited from Fed's monetary policy, while going against the Fed's asymmetry preference may result in a debacle in financial market due to an "opposite" policy" (p.15). Given this statement, it is even more expected to find similarities between SPF and Greenbook.

Although the choice of the two sources of survey forecasts (SPF and Greenbook) was limited due to data availability, these two selected sources are among the most common and most important survey forecasts in US regarding macroeconomic variables. The differences between US and Europe concerning survey forecasts also naturally led to analyzing US data as survey forecasts are still recent in Europe; the European Central Bank started to issues survey forecasts for the economy of Europe in 1999, compared to 1968 for SPF.

As one can see, the two data sets have a quite similar structure in the sense that they both cover approximately the same time period and provide quarterly forecasts. In order to push the comparison even further, I decided to narrow the samples based on the lower and upper constraints so to have two identical samples. Concretely, the lower bound of Greenbook is increased to 1968:4 while the upper limit of the SPF forecast is reduced to 2010:4. Our baseline samples therefore span from 1968:4 to 2010:4. With this slight adjustment in samples selection, both data sets have now the exact same time period. Similarly, in order to match the time horizons, only the nowcasts and the forecasts for the subsequent four quarters have been considered. This narrowing does not seem to undermine our analysis as only a very small portion of the initial samples is dismissed from the data and more importantly, we expect

that all kind of dissimilarities among experts due to forecasting at different times are removed. To put it another way, if forecasting performances were found to be significantly better for one individual set over the other, it would only come from better predictive powers and not due to time varying specifications.

As stated above, the present thesis has two objectives. First, my own forecasts will be analyzed and evaluated against forecasts from SPF and Greenbook. Their accuracies will be measured against the real observations collected and maintained by the US Bureau of Economic Analysis.[10] Second, historical forecasts from SPF and Greenbook will be combined and this combination together with the two individual forecasts will be evaluated against the actual observations. This second step will shed light on forecast combinations and its potential usefulness as a forecasting technique in itself.

We have just presented the data for the combination exercise. Regarding the forecasting task, let us first define two important concepts in forecast evaluation. Assume we are interested in how good the model forecasts future observations. For this purpose, sample is often split into an *estimation sample* and an *forecast sample*. The estimation sample is used for estimating parameters, while the forecast sample is used for comparing model forecasts and observed values. From an econometric point of view, estimation and forecast samples should be similar since information from the estimation sample will be used to forecast observations in the forecast sample. Nonetheless, as the task of forecasting in this case is to compare them with forecasts made by experts, we set the estimation sample equal to the observations that was available to them at the time of their predictions, and the forecast sample equal to

[10]Data on US GDP from 1947 are available on www.bea.gov. These real data will serve as true observations upon which our analysis will be built.

their forecasts.

It is worth highlighting that the estimation sample must correspond to the observations available *at the time when the predictions were made*. For example, if we were to evaluate a forecast made in 1970, we must use the data that were available to the forecaster in 1970 and not the data available today. Data available at a particular moment in time are referred as "vintage data". The reason why one must not use the latest available data is that data of macroeconomic variables are often revised after being published. It is evident that forecasters build their predictions for the following quarters based on the data that are available to them at that time and they can not use revised data since the revisions are made after the release of their forecasts. Using revised data (which are by definition closer to the real observations)[11] to compare SPF, Greenbook and our forecasts against real observations would obviously gives our forecasts a significant advantage over the two others. By using vintage data we exclude the possibilities of obtaining smaller forecasting errors thanks to a data set that gives more insights into the expected outcome (and this, independently of the models or assumptions defined).

We consider the vintage datasets from the real-time database maintained by the Federal Reserve Bank of Philadelphia.[12] Following Corradi et al. (2009) we use first release data, that is, data released the next quarter (except for 1995:4 where there was no data released in 1996:1, hence we use data released in 1996:2). The vintage data set includes historical data on US real GDP in *levels* from 1947:1 to 2015:4. The entire time-series history for each vintage is seasonally adjusted. The data were

[11] By closer I do not necessarily mean more right but definitely more in line with the observed values.

[12] Available via www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/data-files.

transformed from levels to annualized percentage using:

$$g_q = \left[\left(\frac{X_q}{X_{q-1}} \right)^4 - 1 \right] \cdot 100 \tag{3.1}$$

where X_q and X_{q-1} are the values of the real GDP in quarters q and $q-1$, respectively, and g_q is the annualized percent change.

One can rightly argue that the period from 1968 to 2010 includes many major economic events that had significant impacts on the observed variable.[13] It is true that these events are mostly unpredictable and that forecasters are generally not able to take them into account in their predictions. Nonetheless, even without any major economic event going on, GDP is known to be hard to predict, or at least, harder than some other macroeconomic variables. As argued by Mankiw et al. (1995), this is the case because there are too many potential influences on GDP for forecasters to isolate cause and effect and too many things happen at once so it is difficult to determine which relationship is stronger such that it prevails over others. A second cause that makes GDP hard to forecast is the high volatility in oil price, which has become a major macroeconomic factor in GDP. The effect that oil price shocks have on an economy is so important that Adelman (1962, p.537) stated that "Oil is so significant in the international economy that forecasts of economic growth are routinely qualified with the caveat: 'provided there is no oil shock.'" Moreover, according to Panas and Ninni (2000), the oil market is the most volatile one after Nasdaq and depicts strong evidence of chaos. As any volatile market or variable is hard to predict, GDP growth is also uncertain and hard to forecast (Xie et al., 2006).

[13] Among others, the Great Moderation, the end of the Great Inflation and the recent Great Recession which all took place in the past 40 years.

In this regards, one might make the argument that the periods of extreme situations should be avoided when forecasting as it does not reflect the "normal path" and is *per se* hard to capture. I however believe that all available data should be included in the sample, both the steady-state and the extreme eras. The latter might indeed create noises in the sample, but it also helps to capture a portion of the future unexpected events. This creates a trade-off between less precise estimates during normal times and more correct results in case of unlikely events.

For a better overview of the selected samples, the different data sets are plotted in Figures 3.1 and 3.2. To save space, only the samples for the two extreme horizons have been plotted, that is, figures for nowcasts ($h = 0$) and forecasts four horizons ahead ($h = 4$). The graphs for the remaining horizons are presented in the appendix (see Figures 7.1, 7.2 and 7.3 for horizons 1, 2 and 3, respectively). As explained at the beginning of section 3.1, the reason that the data ends in 2010 is because Greenbook's predictions are released to the public with a lag of five years, so the present analysis covers the period until that time. The graphs show that the larger the horizon, the less accurate the estimations provided by the two institutions. This pattern makes sense since forecasters have more true information upon which they can build their forecasts for the next period than when predicting several periods ahead. Notice that, for forecasts of 3 and 4 quarters ahead, some values are missing in the data sets of the two survey forecasts. We handle these missing values by omitting the periods in which there are such missing values. This led to the removal of 13 quarterly forecasts within the covered period, providing 156 forecasts for each horizon. This way of dealing with missing value does not entail our analysis as a significant number of forecasts remains for the analysis.

The descriptive statistics are summarized in Table 3.1. It can be seen that, for all

horizons, the mean and median of the GDP growth forecasts are relatively similar and close to the observed data. The standard deviation is larger for the actual data than for the forecasts, with a standard deviation of 3.228 for the actual data compared to a standard deviation ranging from 0.935 for SPF forecasts at $h = 4$ to 2.858 for Greenbook forecasts at $h = 0$. The GDP growth forecasts range from -10.9 to 8.8 (both for Greenbook forecasts at $h = 0$), while the minimum and maximum of the observed data is -10.370 and 11.160. The first-order autocorrelation is the lowest for the actual data with an autocorrelation of 0.493 and largest for the SPF forecasts at $h = 3$ with an autocorrelation of 0.832. The actual distribution of the GDP growth appears skewed, with the sign of the asymmetry being negative, indicating that the tail on the left-hand side of the density function is longer or fatter relative to the tail on the right hand-side. Whereas Greenbook forecasts seem to capture this feature (indicated by the negative skewness at all horizons), SPF forecasts and the average of the two survey forecasts miss this characteristics at the two largest horizons. The kurtosis suggested by the survey forecasts and the observed data indicate light tails or lack of outliers and a distribution concentrated toward the mean in most of the sub-samples except for the Greenbook forecasts at $h = 0$ where the kurtosis is only slightly larger than the one implied by the normal distribution (*i.e.*, a kurtosis of 3).

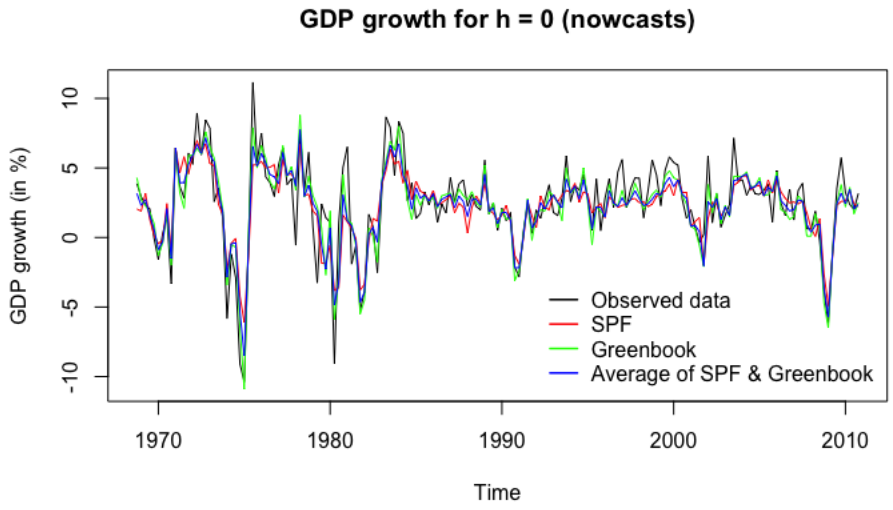

Figure 3.1: GDP growth forecasts provided by SPF, Greenbook, the average of the two for $h = 0$ (nowcasts) and observed data from 1968:4 to 2010:4

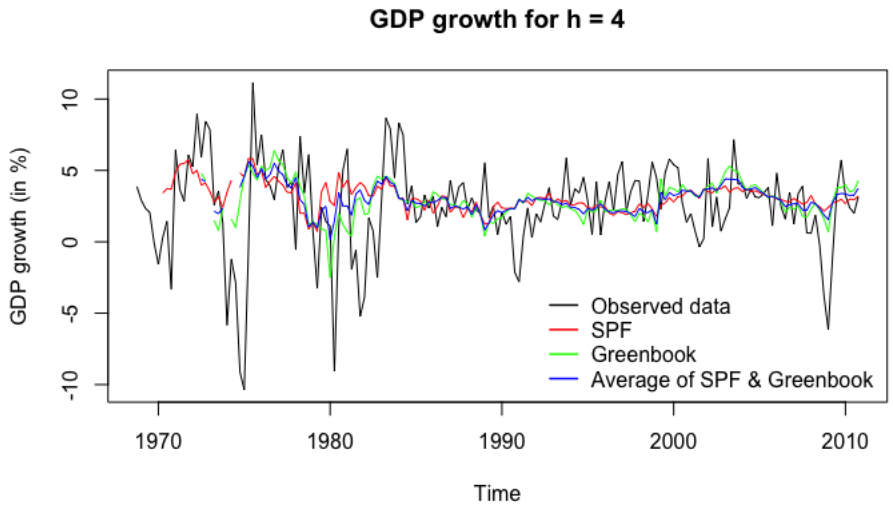

Figure 3.2: GDP growth forecasts provided by SPF, Greenbook, the average of the two for $h = 4$ and observed data from 1968:4 to 2010:4

Table 3.1: Descriptive statistics of GDP growth forecasts and observed data

Horizon	Mean	Median	Stdev.	Min	Max	Autocorr.	Skewness	Kurtosis
Observed data								
	2.459	2.650	3.228	-10.370	11.160	0.493	-0.967	2.671
SPF								
$h = 0$	2.286	2.523	2.314	-6.096	6.966	0.726	-0.906	1.384
$h = 1$	2.604	2.617	1.794	-3.751	6.447	0.789	-0.657	1.330
$h = 2$	2.850	2.826	1.395	-2.671	6.121	0.817	-0.598	2.242
$h = 3$	3.066	2.996	1.021	0.179	5.866	0.832	0.323	0.451
$h = 4$	3.132	3.016	0.935	0.726	5.824	0.777	0.467	0.528
Greenbook								
$h = 0$	2.307	2.600	2.858	-10.900	8.800	0.592	-1.201	3.152
$h = 1$	2.521	2.600	2.400	-5.000	7.900	0.726	-0.606	0.898
$h = 2$	2.741	2.700	1.926	-3.700	7.500	0.726	-0.553	1.391
$h = 3$	2.892	2.700	1.637	-3.800	7.500	0.695	-0.198	1.311
$h = 4$	2.914	2.900	1.331	-2.500	6.400	0.767	-0.117	0.969
Average of SPF & Greenbook								
$h = 0$	2.296	2.368	2.533	-8.498	7.744	0.675	-1.120	2.471
$h = 1$	2.563	2.627	2.040	-4.225	7.073	0.781	-0.602	1.093
$h = 2$	2.796	2.760	1.586	-2.794	6.710	0.800	-0.445	1.634
$h = 3$	2.971	2.856	1.247	-1.381	6.683	0.761	0.211	0.779
$h = 4$	2.997	2.896	0.994	0.159	6.074	0.763	0.409	0.423

Notes: The table shows the descriptive statistics of the GDP growth forecasts (provided by SPF, Greenbook and the average of the two) and the observed data from the full sample period 1968:4-2010:4, providing 169 quarterly data points for the horizons from 0 to 2, 166 observations for $h = 3$ and 156 observations for $h = 4$. "Stdev." denotes standard deviation and "Autocorr." the first-order autocorrelation of the series.

Section 4

Models

As mentioned *supra* we use ARIMA(p, d, q) (hereafter, ARIMA) models to construct our forecasts, which will serve as benchmark when evaluating combined forecasts. The full model can be written as

$$\Delta^d y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}, \quad (4.1)$$

where y_t and ϵ_t are the actual value and random error at time period t, respectively, ϕ_i $(i = 1, 2, ..., p)$ and θ_j $(j = 1, 2, ..., q)$ are model parameters, p and q are non-negative integers and correspond to orders of the model (Zhang, 2003).

We fit the best ARIMA model to each time series according to Akaike's Information Criterion (AIC) (Akaike, 1973) using the automated function `auto.arima()` in the software R (R Core Team, 2015) developed by Hyndman and Khandakar (2008). The function conducts a search over possible models and returns the best ARIMA model for each series on the basis of its past data, ultimately leading to more than just one specific ARIMA model used for the whole set of data. A quasi-realtime forecast for the current quarter (horizon $h = 0$) and the next 4 quarters $(h = 1, 2, 3, 4)$

is then made for each series. By quasi-realtime forecast we have in mind forecast for some point in time in the past based only on vintage data, *i.e.*, data that was available at that time (Faust and Wright, 2013).

Despite being very parsimonious, ARIMA models are the best univariate forecasting tool as shown by Stock and Watson (1998). Furthermore, using ARIMA models as benchmark is very simple and yet hard to beat. See for instance, Aiolfi et al. (2010); Ang et al. (2007); Faust and Wright (2013); Nelson (1984) and Stock and Watson (2004) who use such models as benchmark. Lastly, Baffigi et al. (2002), Guo-yong (2008) and Rünstler et al. (2003) use ARIMA models in GDP forecasting. This proves that we can compare professional forecasts with forecasts from more common time series models used in the literature for GDP.

Regarding the type of forecast used for this task, we want to find the forecasting approach that is the most similar to the task of the professional forecasters. The *dynamic* forecast approach uses past information to forecast at a specific point n horizons ahead, without updating even when new data becomes available. In contrast, the *static* forecast approach 'updates' its forecasts by using the actual value for each subsequent forecast. As professional forecasters provide their predictions for the following quarters at once, that is, their beliefs regarding the relevant macroeconomic variables for the current quarter and up to 9 quarters ahead, only the dynamic forecast technique is considered in our analysis.

4.1 Evaluation tool

The main evaluation tool used throughout the analysis is the root mean squared error (RMSE), given by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.2}$$

where y_i is the observed value for the ith observation and $\hat{y}_i$ is the predicted value.

The secondary evaluation tool is the mean absolute error (MAE), define as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{4.3}$$

where y_i is the observed value for the ith observation and $\hat{y}_i$ is the predicted value.

Both metrics are regularly employed in forecast and model evaluation researches. Willmott and Matsuura (2005) and Willmott et al. (2009) have suggested that MAE is a better measure of average model performance. However, Chai and Draxler (2014) recently demonstrated that "the RMSE is more appropriate to represent model performance than the MAE when the error distribution is expected to be Gaussian" (p. 1247). In addition, the major difference between RMSE and MAE is that the former amplifies and punishes large errors more severely than the latter. RMSE has thus been selected over MAE as the primary evaluation tool. The choice remains subjective but through this performance evaluation measure we aim at evaluating forecast accuracy while strongly penalizing predictions that are far from the values actually observed, as large errors are particularly undesirable for forecast users and policymakers. For completeness, we present results of our analysis under both evaluation tools, but we will focus our conclusions based on findings from RMSEs and

we will mention the conclusions based on MAEs when deemed necessary.

For ease of interpretation, RMSEs and MAEs are usually presented relative to the benchmark in the denominator so that ratios lower than unity suggest that the alternative model or combination of models outperforms the benchmark.

We sometimes refer to the mean squared errors (MSE) in some forecast combinations, in particular when computing the weights assigned to the different individual forecasts. This evaluation tool is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{4.4}$$

which is basically the square form of equation (4.2).

Section 5

Forecast combination methods

Several combination methods are applied to our data sets and each of them is detailed in the next subsections. These techniques vary in the way they treat historical information to produce combined forecasts and in the extent to which the weight associated to an underlying forecast fluctuates over time. We start with the two simplest approaches: equal-weighted average and median of forecasts, the analysis is then extended to more complex combination methods.

5.1 Equal-weighted average and median forecasts

5.1.1 Equal-weighted average

Although the equal-weighted average combination scheme is simple, it has been shown that it performs relatively well and its performance in empirical applications is even sometimes puzzling when compared to more advanced weighted combinations. Clemen first addressed the issue when raising the following question: 'What is the explanation for the robustness of the simple average of forecasts?' (1989, p. 566).

Stock and Watson (1998, 2003, 2004) found similar results, that is, a simple average with equal weights often outperforms more sophisticated weighting schemes based on mean-squared forecast errors (MSFE). More recently, for GDP growth Genre et al. (2013) found that only a few of the forecast combination schemes tested are able to outperform the simple equal-weighted average forecast.[14] Furthermore, Smith and Wallis argue that "if the optimal combining weights are equal or close to equality, a simple average of competing forecasts is expected to be more accurate, in terms of MSFE, than a combination based on estimated weights" (2009, p. 351), due to the parameter estimation effect.

This forecast combination sets the weights given to the individual forecasts regardless of their historical performance. Two different approaches are derived from this equal-weighted combination scheme: (i) equal-weighted average of SPF and Greenbook forecasts, and (ii) equal-weighted average of SPF, Greenbook and time series forecasts (*i.e.*, forecasts with ARIMA models discussed in Section 4).

The combination of forecasts has the form

$$f_{t+h|t} = \sum_{i=1}^{n} w_{it} \hat{Y}^{h}_{i,t+h|t} \tag{5.1}$$

where $f_{t+h|t}$ is the combination forecast, w_{it} is the weight on the ith forecast in period t, $\hat{Y}^{h}_{i,t+h|t}$ is the ith individual forecast of $Y^{h}_{i,t+h|t}$, computed at date t and n is the number of forecasts in the panel (Stock and Watson, 2004). For equal-weighted average, w_{it} in equation (5.1) equals $1/n$, which yields

$$f_{t+h|t} = \sum_{i=1}^{n} \frac{1}{n} \hat{Y}^{h}_{i,t+h|t}. \tag{5.2}$$

[14] Combination techniques tested are combinations based on principal components analysis and trimmed means, performance-based weighting, least squares estimates of optimal weights as well as

5.1.2 Median

The median of the different individual forecasts as a combination method is also considered. Since the median of two numbers is equal to the mean (which is already covered by the equal-weighted average combination method), we only compute the RMSEs of the median of the time series (TS) forecasts, SPF and Greenbook forecasts.

Results of these two combination methods are provided in Subsection 6.1.

5.2 Inverse MSE-weighted average forecast

This combination method suggested by Bates and Granger (1969) and Diebold and Pauly (1987) computes the weights given to individual forecasts based on their historical performance, measured by the mean squared errors (MSEs). In particular, the weight of an individual forecast for the subsequent forecast combination depends inversely on its past MSE relative to MSEs of all individual forecasts. Specifically, this combination method has the form of equation (5.1) where the weight on the ith forecast is

$$w_{it} = \frac{MSE_{i,t}^{-1}}{\sum_{j=1}^{m}(MSE_{j,t}^{-1})} \tag{5.3}$$

where $MSE_{i,t}$ is the mean squared forecast error over the historical period over which forecasts have been computed (Chan et al., 1999), given by the square of the RMSE (equation (4.2))

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.4}$$

where y_i is the observed value for the ith observation and $\hat{y}_i$ is the predicted value.

The weights w_{it} have been determined from past known MSEs, except for the

Bayesian shrinkage.

weights for the first quarter of the sample since historical performance is still unknown at this point in time. For this first quarter the weight assigned to each individual forecast has been arbitrarily set to $1/n$ where n corresponds to the number of underlying forecasts considered.

Results of this combination method are provided in Subsection 6.2.

5.3 Predictive least squares and adaptive expectations

Predictive least squares and adaptive expectations are presented in the same section as they are similar in the sense that they assign 100% of the weight to one individual forecast, as opposed to the equal-weighted average and inverse MSE-weighted combinations which let the weight vary among the underlying forecasts. The two methods differ, however, in the way the selected individual forecast is chosen. Further details for each method are exhibited in the next two subsections.

5.3.1 Predictive least squares

The predictive least squares (PLS) approach, also referred as "most recently best" by Stock and Watson (2004) or sometimes "winner-takes-all" by others, places *all* weight on the underlying forecast that has produced the best forecasts as measured by the lowest average MSE. For completeness, we select the forecast that receives all weight based on two criteria: (i) lowest average MSE from the beginning of the sample up to the previous observation, and (ii) lowest average MSE over the previous

[15]The number of periods is in accordance with Stock and Watson (2004). Furthermore, this choice is motivated by the fact that four periods appear to give the best trade-off between flexibility, adaptability and persistence across historical performance of the survey forecasts.

four periods.[15] In other words, criterion (i) selects the best individual forecast for the current quarter based on the accuracy of its historical forecasts for the entire sample up to the previous quarter, while criterion (ii) chooses the best underlying forecast based on the accuracy of its last four forecasts.

The second criterion has the advantage that if a survey forecast consistently outperforms the other during a certain period of time within the entire sample, it is taken into account by the combination. Indeed, as an illustration, if SPF forecasts outperform Greenbook forecasts over 20 quarters in a row, the combination will handle this good performance by giving all weight to SPF forecasts for at least 16 quarters.[16]

Note that, as implemented in the inverse MSE-weighted average combination, weights for the first set of forecasts have been set to $1/n$ where n corresponds to the number of underlying forecasts considered since no information regarding past performance is known for the first observation. Moreover, for the second criterion, the observations that have less than four previous periods used all previous observations to compute the average MSE.

Results of this combination method are provided in Subsection 6.3.1.

5.3.2 Adaptive expectations

In this subsection, a new forecast combination technique based on the adaptive expectation model (Kmenta, 1986) and the concept of "regression to the mean" (Kahneman, 2011) is proposed to improve the accuracy of forecast combination. For simplicity, we will write only of the terms − adaptive expectation (AE) − to refer to this combination technique.

[16]This number may be larger depending on the magnitude of the difference in performance before and after these 20 quarters. Nonetheless, it is assured to last at least 16 quarters.

Before presenting the combination technique, we first define the two underlying concepts. The adaptive expectations theory states that individuals base their expectations of future outcomes on past events. For forecasters, the theory means that they will predict future outcomes based (partially) on their past forecasting errors. For instance, if a forecaster sees that her recent past prediction of the GDP growth is above the actual data, she will adjust her future predictions downwards. The opposite also holds; if another forecaster realizes that he underestimated the GDP growth at time $t-1$, he will adjust his forecast at time t by increasing it.

The concept of regression to (or towards) the mean, introduced by Galton in 1886, is to some extent similar to adaptive expectations but is however more general. As Kahneman (2011) explains in his brilliant book, performances of flight cadets tend to revert to the mean, which is due to the random fluctuations in the quality of performance.[17] This means that a flight cadet whose performance was better than average was probably just lucky on that specific attempt and thus likely to deteriorate on the next try. The opposite is true; a flight cadet who performed unusually bad on a particular attempt was probably unlucky and therefore likely to improve on the next try. More formally, in statistics, regression to the mean states that extreme observations of a variable tend to be followed by observations closer to the mean (Stigler, 1997). Put in another way, mean reversion means that observations tend to even out over time.

Assume that forecasters and flight cadets are alike, *i.e.*, human beings and thus not immune to making mistakes. Based on Kahneman (2011) summary statement that "poor performance was typically followed by improvement and good performance by deterioration" (p. 176), the choice of considering the worst historical forecast (measured again by MSEs) to predict the next period seems legitimate. More-

[17]Here, performance is measured by the ability to execute complex aerobic maneuvers.

over, incorporating the adaptive expectation model into fuzzy time series-models for the stock market,[18] Cheng et al. (2008) found that their model surpasses in accuracy the models that do not include Kmenta's (1986) model. Therefore, I have some reasons to believe that, although the approach of putting all weight on the worst survey forecast may seem counter-intuitive at first sight, this combination could actually perform at least as good as the other combinations detailed so far.

Since we want to capture the mean reversion phenomenon (that is, poor accuracy is followed by better accuracy in terms of forecasting ability) and avoid taking into consideration individual forecasts that perform badly in general (over a long time period) for unknown reasons, only the MSE of the previous period is considered to determine which forecast will receive 100% of the weight. This combination is similar to the predictive least squares combination except that instead of placing all weight on the best forecast measured by its previous MSE (computed up to date or over the last four periods), all weight is placed on the worst forecast measured by its previous MSE over the last period only.

Note that, as implemented in the inverse MSE-weighted average combination and the predictive least squares combination, weights for the first set of forecasts have been set to $1/n$ where n corresponds to the number of underlying forecasts considered since no information regarding past performance is known for the first observation.

Results of this combination method are provided in Subsection 6.3.2.

5.4 Bayesian model averaging

In the adaptive expectations combination, a model is selected and we proceed with this model as if it had generated the data. This approach, however, does not account

[18] Fuzzy time series, introduced by Song and Chissom (1993), refer to time series in which historical

for the uncertainty in model selection (Hoeting et al., 1999). On the contrary, the Bayesian model averaging (BMA) approach to forecast combination includes this model uncertainty. Indeed, the weights assigned to the individual forecasts when combining them can be based on the probabilities of each being the true model.

This method of combination received wide attention in the literature and has proved to be successful in the forecasting literature as well. In his paper, Wright (2009) considered Bayesian model averaging of US inflation, and finds that it generally gives more accurate forecasts than simple equal-weighted averaging. Koop and Potter (2004) apply these methods to the problem of forecasting GDP and inflation using quarterly U.S. data on 162 time series and report that the gains provided by using Bayesian model averaging over forecasting methods based on a single model are appreciable. Avramov (2002) uses Bayesian model averaging to analyze the sample evidence on return predictability in the presence of model uncertainty and shows that the out-of-sample performance of the Bayesian approach is superior to that of model selection criteria. Cremers (2002) analyses the stock return predictibility and finds that the out-of-sample results for the Bayesian average models show improved forecasts relative to the classical statistical model selection methods.

Bayes theorem states that when there are a set of models (M_1 and M_2), then the probability that model M_1 is true given the data (D) is

$$p(M_1|D) = \frac{p(D|M_1)p(M_1)}{p(D|M_1)p(M_1) + p(D|M_2)p(M_2)}, \tag{5.5}$$

where $p(D|M_k)$ is the probability of the data given M_k, and $p(M_k)$ is the prior probability of model M_k ($k = 1, 2$). A similar equation holds for $p(M_2|D)$ and

observations are linguistic values instead of real numbers. Weather conditions (good, bad, cold, hot, sunny, rainy, etc.) and the mood of a person (good, bad, etc.) over time are examples of fuzzy time series.

$p(M_1|D) + p(M_2|D) = 1$ (Raftery, 1995).

Furthermore, as argued by Raftery (1995) and Doppelhofer et al. (2004), Bayesian information criterion (BIC) (Schwarz, 1978) can be used as an accurate approximation to Bayes factors. Equation (5.5) is approximated by

$$p(M_k|D) \approx \frac{exp(-\frac{1}{2}BIC_k)}{\sum_{l=1}^{K} exp(-\frac{1}{2}BIC_l)}, \tag{5.6}$$

where these probabilities can be used as forecasting weights.

Note that, as implemented in the previous combinations, weights for the first set of forecasts have been set to $1/n$ where n corresponds to the number of underlying forecasts considered since no information regarding past performance is known for the first observation. Due to missing values at the beginning of the sample, the BIC could not be computed for the first set of forecasts (*i.e.*, 1969:4) at the largest horizon. Weights are therefore set to $1/n$ for this selection too.

Results of this combination method are provided in Subsection 6.4.

Section 6

Results

A discussion of the results using the different combination schemes detailed in Section 5 are presented below. For policy implications, the results are displayed by the length of the forecasting horizon rather than by the quarter in which the forecasts were made. This way, the interested reader is able to distinguish the most appropriate model or combination of models depending on the forecasting horizon considered. Indeed, it may be the case that a combination method produces the best forecasts for a given horizon, while another combination scheme yields more accurate predictions for another horizon. This is the aim of the present section.

6.1 Equal-weighted average and median forecasts

This subsection presents the results using the combination methods detailed in Subsection 5.1.1 and 5.1.2.

Table 6.1 reports RMSEs of our forecasts, RMSE ratios (with respect to forecasts made with ARIMA models) of the survey forecasts and the combination of those

forecasts using the equal-weighted average and median schemes. Notice first the relatively high RMSEs for the TS forecasts at all horizons. One could argue that given the mixed performance of the benchmark, it is not particularly difficult to "beat" it. The benchmark does not, however, serve as an anchor to our analysis. Indeed, the aim of the present thesis is to test several combination techniques in order to see which one leads to better forecasts than the individual survey forecasts, that is, the forecasts provided by SPF and Greenbook (or if on the contrary individual forecasts are superior in different contexts). The relatively low predictability power given by our TS models does not entail our analysis as the benchmark can be seen as a way to gain some perspective on the magnitude of the benefits of model averaging, rather than a way to observe the benefits of model combinations relative to model selection (as opposed to Banternghansa and McCracken (2010)). Moreover, remember that TS forecasts are based on ARIMA models with different orders p and q selected by AIC. The RMSEs for the TS forecasts are thus not comparable across horizons and this is the reason why they do not increase as the horizons become larger. Indeed, we expect that the RMSEs increase (that is, the accuracy of the forecasts decreases) with the horizons as forecasts several quarters ahead are usually less precise than forecasts for the current or upcoming quarter. TS forecasts do not follow this pattern due to the different underlying models upon which the forecasts are built. Regardless of this, some conclusions can be drawn from the findings displayed in Table 6.1.

First, survey forecasts taken individually perform better than TS forecasts at all horizons, although the accuracy of the professional forecasters decreases as the horizon becomes larger (which in this case is in line with the expected increasing trend detailed above). For instance, whereas the professional forecasters are almost twice as accurate as the TS forecasts regarding nowcasts (RMSE ratios of 0.6 and 0.44 for forecasters from SPF and Greenbook, respectively), their accuracy decreases such

Table 6.1: Forecast performance of time series forecasts, individual survey forecasts and combinations of survey forecasts and time series models using equal-weighted average and median forecasts

	RMSE	RMSE ratios with respect to ARIMA				
	TS forecasts	Survey forecasts		Equal-weighted average		Median
Horizon	ARIMA	SPF	Greenbook	SPF_GR	ARIMA_SPF_GR	ARIMA_SPF_GR
$h = 0$	3.199	0.599	0.440	0.493	0.599	0.543
$h = 1$	3.384	0.743	0.701	0.702	0.747	0.729
$h = 2$	3.373	0.828	0.835	0.816	0.842	0.843
$h = 3$	3.337	0.906	0.897	0.887	0.900	0.907
$h = 4$	3.174	0.919	0.902	0.894	0.913	0.922
	MAE	MAE ratios with respect to ARIMA				
$h = 0$	2.176	0.676	0.457	0.528	0.639	0.593
$h = 1$	2.286	0.803	0.773	0.773	0.790	0.793
$h = 2$	2.374	0.841	0.861	0.837	0.851	0.866
$h = 3$	2.287	0.906	0.931	0.906	0.906	0.911
$h = 4$	2.246	0.907	0.923	0.900	0.910	0.913

Notes: For each horizon, the root mean squared error (RMSE) and mean absolute error (MAE) are calculated with 156 forecasts (quarterly forecasts covering the period from 1968:4 to 2010:4 where 13 quarterly forecasts were removed from the sample due to missing values in SPF and Greenbook data sets). Time series (TS) forecasts are based on ARIMA models selected by the AIC through an exhaustive search. Survey forecasts report the RMSE and MAE ratios of the predictions made by the experts from SPF and Greenbook. For the equal-weighted average combination, SPF_GR corresponds to the average of the SPF and Greenbook data sets, whereas ARIMA_SPF_GR refers to the average of the ARIMA, SPF and Greenbook forecasts. The median corresponds to the median of the TS, SPF and Greenbook forecasts. Survey forecasts, equal-weighted average and median present respective RMSEs and MAEs with respect to TS forecasts.

that RMSE ratios approach unity (ratios of 0.92 and 0.9 for SPF and Greenbook, respectively) when it comes to predicting GDP growth one year ahead ($h = 4$). Greenbook forecasts seems to be the most accurate among the two survey forecasts during the period covered since their predictions outperform SPF predictions for all horizons except for $h = 2$ where the advantage for SPF forecasts is relatively small (ratio of 0.83 compared to 0.84 for SPF and Greenbook, respectively). Improvements of Greenbook forecasts compared to SPF forecasts in terms of RMSE ratios with

respect to TS forecasts varies from 16% when $h = 0$ to 1% when $h = 3$.

Second, equal-weighted average of SPF and Greenbook forecasts (denoted as SPF_GR in Table 6.1) outperforms any of the two underlying forecasts for 3 horizons out of 5 (*i.e.*, $h = 2, 3$ and 4), performs similarly than the best of the two individual forecasts for another horizon ($h = 1$) and performs better than the least accurate of the two but performs worse than the most accurate of the two for the remaining horizon ($h = 0$). Although the improvements when considering the equal-weighted average is rather limited (improvement of 1% and 2 to 11% compared to the best and worst of the two forecasts in terms of RMSE ratio with respect to TS forecasts, respectively), the fact that the simple equal-weighted average is at least as accurate as the best individual forecast 80% of the time and more accurate 60% of the time already indicates that this method could be beneficial to forecast users.

Third, including ARIMA forecasts in the equal-weighted average combination scheme (denoted as ARIMA_SPF_GR in Table 6.1) decreases its forecastability power, in particular for the short term horizons. These results were expected given the relatively poor performance of the TS series forecasts. Furthermore, the principle made by Faust and Wright (2013) that subjective forecasts usually do best when forecasting inflation can now be extended to GDP growth, for the span covered in this analysis at least.

Fourth, the median of the three available forecasts (*i.e.*, TS, SPF and Greenbook forecasts) performs worse in terms of RMSEs than using the simple equal-weighted average of the two survey forecasts. These results are also expected given the performance of the TS models. The median combination, however, performs similarly than the equal-weighted average when all three forecasts are included. Further research taking into account more than two survey forecasts would shed light on whether combinations using median forecast estimates are superior to the equal-weighted average

association.

These conclusions do not significantly vary if the secondary evaluation tool (MAE) is considered instead of the RMSE.

6.2 Inverse MSE-weighted average forecast

This subsection presents the results using the combination method detailed in Subsection 5.2.

Results of the second method of forecast combination are depicted in Table 6.2. Given that the aim of the analysis is to see if an alternative option can produce better forecasts than the individual forecasts (at any given horizon), results of the inverse MSE-weighted average forecast combination are again presented besides the results of the underlying forecasts, that is, SPF and Greenbook predictions.

Table 6.2 shows some interesting results. First, combination of the two survey forecasts (denoted as SPF_Greenbook in Table 6.2) using the inverse MSE-weighted average outperforms the combination of the two survey forecasts and TS forecasts (denoted as ARIMA_SPF_Greenbook in Table 6.2) using the same method. Second, the RMSEs ratios of the former combination suggest that this method outperforms any of the two individual forecasts for predictions 2, 3 and 4 quarters ahead. Furthermore, whereas for one quarter ahead the inverse MSE-weighted average produces similar results than the most accurate individual forecast, for nowcasts the combination outruns only one of the two underlying forecasts.

This method also proved its potential as a combination technique since it is at least as accurate as the best individual forecast 80% of the time (4 horizons out of 5) and more accurate 60% of the time (3 horizons out of 5). However, for the three horizons where the combination beats the survey forecasts, the improvements are

Table 6.2: Forecast performance of inverse MSE-weighted average forecast compared to survey forecasts

	RMSE	RMSE ratios with respect to ARIMA			
	TS forecasts	Survey forecasts		Inverse MSE-weighted average	
Horizon	ARIMA	SPF	Greenbook	SPF_GR	ARIMA_SPF_GR
$h = 0$	3.199	0.599	0.440	0.471	0.486
$h = 1$	3.384	0.743	0.701	0.702	0.717
$h = 2$	3.373	0.828	0.835	0.817	0.829
$h = 3$	3.337	0.906	0.897	0.889	0.900
$h = 4$	3.174	0.919	0.902	0.895	0.908
	MAE	MAE ratios with respect to ARIMA			
$h = 0$	2.176	0.676	0.457	0.495	0.516
$h = 1$	2.286	0.803	0.773	0.775	0.784
$h = 2$	2.374	0.841	0.861	0.838	0.844
$h = 3$	2.287	0.906	0.931	0.908	0.907
$h = 4$	2.246	0.907	0.923	0.901	0.908

Notes: For each horizon, the RMSE and MAE are calculated with 156 forecasts. TS forecasts are based on ARIMA models selected by the AIC through an exhaustive search. Survey forecasts report the RMSE and MAE ratios of the predictions made by the experts from SPF and Greenbook. For the inverse MSE-weighted average combination, SPF_GR corresponds to the weighted average of the SPF and Greenbook data sets, whereas ARIMA_SPF_GR refers to the weighted average of the ARIMA, SPF and Greenbook forecasts. Both survey forecasts and inverse MSE-weighted average present respective RMSEs and MAEs with respect to TS forecasts.

bounded at 1%. I expected better results using this combination technique since it takes into account past performance of each survey forecast to assign weights for the subsequent periods. A forecast with more historical forecasting power (based on historical performance measured by RMSE) will thus receive more weight when combined with other forecasts. Surprisingly, the results indicate that the combination is useful for forecast users, but only to a small extent.

These conclusions do not significantly vary if the secondary evaluation tool (MAE) is considered instead of the RMSE.

6.3 Predictive least squares and adaptive expectations

6.3.1 Predictive least squares

This subsection presents the results using the combination method described in Subsection 5.3.1.

As detailed in Subsection 5.3.1, PLS places all weight on the best individual forecast, measured in two different ways: the (i) lowest average MSE from the beginning of the sample up to the previous observation, and the (ii) lowest average MSE over the previous four periods. Findings using this combination are reported in Table 6.3.

From Table 6.3, we see that the combination of the two survey forecasts outperforms the same combination when TS forecasts are included, regardless of the criterion. Second, the PLS combination of the two survey forecasts seems more efficient in terms of predictive power when the second criterion is applied, meaning that the combination gives better results when only the last four forecasts are considered to determine the best forecast for the next period. Indeed, the RMSE ratios with respect to ARIMA are lower for 4 horizons out of 5 (with the exception at $h = 0$) when comparing to the first criterion where the entire sample up to the previous quarter determines the best forecast for the current quarter. Third, when considering only the four last observations to determine the next best forecast, the PLS combination of the two survey forecasts is able to outperform the less accurate of the two individual survey forecasts for 40% of the time ($h = 0, 1$), perform equally well than the most accurate of the two at $h = 2$, perform worse than the two at $h = 3$ and outperform both survey forecasts at the largest horizon. Fourth, we see that across all combinations the RMSE ratios are lower at $h = 4$ than at $h = 3$. This is

Table 6.3: Forecast performance of the predictive least squares combination compared to survey forecasts

	RMSE	RMSE ratios with respect to ARIMA					
	TS forecasts	Survey forecasts		PLS			
				(i) up to date		(ii) last 4 forecasts	
Horizon	ARIMA	SPF	Greenbook	SPF_GR	+ARIMA	SPF_GR	+ARIMA
$h = 0$	3.199	0.599	0.440	0.442	0.443	0.470	0.599
$h = 1$	3.384	0.743	0.701	0.725	0.727	0.715	0.776
$h = 2$	3.373	0.828	0.835	0.856	0.857	0.828	0.843
$h = 3$	3.337	0.906	0.897	0.936	0.940	0.914	0.925
$h = 4$	3.174	0.919	0.902	0.906	0.908	0.888	0.878
	MAE	MAE ratios with respect to ARIMA					
$h = 0$	2.176	0.676	0.457	0.458	0.461	0.489	0.561
$h = 1$	2.286	0.803	0.773	0.801	0.805	0.811	0.860
$h = 2$	2.374	0.841	0.861	0.877	0.880	0.844	0.875
$h = 3$	2.287	0.906	0.931	0.962	0.972	0.929	0.952
$h = 4$	2.246	0.907	0.923	0.931	0.931	0.894	0.889

Notes: For each horizon, the RMSE and MAE are calculated with 156 forecasts. TS forecasts are based on ARIMA models selected by the AIC through an exhaustive search. Survey forecasts report the RMSE and MAE ratios of the predictions made by the experts from SPF and Greenbook. For the PLS combination, both selection criteria are reported; lowest average MSE (i) up to date and lowest average MSE over the (ii) last 4 forecasts. Moreover, SPF_GR corresponds to the weighted combination of the SPF and Greenbook data sets, whereas +ARIMA refers to the weighted combination of the ARIMA, SPF and Greenbook forecasts. Both survey forecasts and PLS combination present respective RMSEs and MAEs with respect to TS forecasts.

in contradiction with the RMSE ratios of the two survey forecasts and the findings of the previous combination techniques where the RMSE ratios constantly increase with the horizon, which means that this combination could be more beneficial for long term predictions.

These conclusions do not significantly vary if the secondary evaluation tool (MAE) is considered instead of the RMSE.

6.3.2 Adaptive expectations

This subsection presents the results using the combination method described in Subsection 5.3.2.

Table 6.4 exhibits results using the adaptive expectations combination. Remember that this combination puts all weight in the less accurate forecast measured by the MSE for the previous quarter. This combination is thus somewhat similar than the PLS combination since they both assign 100% of the weight to one individual forecast, but they differ in the sense that one places all weight in the best forecast while the other allocates all weight in the worst previous forecast.

Table 6.4: Forecast performance of the adaptive expectations combination compared to survey forecasts

	RMSE	RMSE ratios with respect to ARIMA			
	TS forecasts	Survey forecasts		Adaptive expectations	
Horizon	ARIMA	SPF	Greenbook	SPF_GR	ARIMA_SPF_GR
$h = 0$	3.199	0.599	0.440	0.544	0.891
$h = 1$	3.384	0.743	0.701	0.715	0.840
$h = 2$	3.373	0.828	0.835	0.859	0.997
$h = 3$	3.337	0.906	0.897	0.961	1.004
$h = 4$	3.174	0.919	0.902	0.971	1.028
	MAE	MAE ratios with respect to ARIMA			
$h = 0$	2.176	0.676	0.457	0.594	0.856
$h = 1$	2.286	0.803	0.773	0.781	0.902
$h = 2$	2.374	0.841	0.861	0.873	0.985
$h = 3$	2.287	0.906	0.931	0.965	0.995
$h = 4$	2.246	0.907	0.923	0.967	1.028

Notes: For each horizon, the RMSE and MAE are calculated with 156 forecasts. TS forecasts are based on ARIMA models selected by the AIC through an exhaustive search. Survey forecasts report the RMSE and MAE ratios of the predictions made by the experts from SPF and Greenbook. For the adaptive expectations combination, SPF_GR corresponds to the weighted combination of the SPF and Greenbook data sets, whereas ARIMA_SPF_GR refers to the weighted combination of the ARIMA, SPF and Greenbook forecasts. Both survey forecasts and adaptive expectations combination present respective RMSEs and MAEs with respect to TS forecasts.

From Table 6.4, we see that including the forecasts based on the ARIMA models decreases the accuracy of the adaptive expectations combination. Second, if we consider the AE combination with the two individual survey forecasts only (denoted by SPF_GR in Table 6.4), the findings show that the combination performs worse than the two survey forecasts 60% of the time (at horizons $h = 2, 3, 4$) and performs slightly better than the less accurate of the two survey forecasts for the two smallest horizons ($h = 0, 1$). Although this combination had some merits from a theoretical point of view, the findings cast some doubts about its efficiency in forecast combination on an empirical ground.

These conclusions do not significantly vary if the secondary evaluation tool (MAE) is considered instead of the RMSE.

In order to test the robustness of this combination scheme, we also tested the combination when all previous forecasts were taken into consideration to determine the weight of the current quarter. Although the findings exhibit an improvement over the initial combination at the three largest horizons, the mixed evidence displayed by this combination does not allow us to advise this technique to forecast users. Furthermore, as mentioned above, the aim of this combination was to capture the mean reversion phenomenon. Therefore, it appears to me that part of the improvement with the broader selection criterion may be related to the specificity of the data sets, and perhaps even due to data snooping. Indeed, there is, in theory, no reason to believe that a forecaster who has performed badly in all previous year suddenly starts to constantly produce forecasts whose performances are above average. If it happens to be the case, it must be due to a change in the tools and methods used and not due to mean reversion or adaptive expectations, otherwise he would have revise his expectations (and thus performed less badly) earlier. This is the reason the results are presented in appendix (Table 7.1) and we draw the interest reader's

attention when interpreting this part of the adaptive expectations combination.

One potential reason that could explain these rather disappointing findings is that the panel of forecasters changes over time. Indeed, the data sets provided by SPF and Greenbook reflect the mean of the individual forecasts made by the experts. However, every quarter, some experts leave the panel of forecasters, while others join the panel. The underlying concept of adaptive expectations works best when forecasters can learn from their mistakes and adjust their predictions depending on their past forecasting errors. This variation in the panel of forecasters cancels part of the potential benefits of the combination. Further research with a non-varying set of forecasters would be needed to more appropriately test this combination technique.

6.4 Bayesian model averaging

This subsection presents the results using the combination method described in Subsection 5.4.

Table 6.5 shows the results of the last combination. As usual, we see that the combination with only the two survey forecasts outperforms the similar combination with the ARIMA forecasts included, although the difference is smaller in this approach compared to the previous ones. Findings combining the two survey forecasts (denoted by SPF_GR) demonstrates that the Bayesian model averaging combination outperforms both survey forecasts at $h = 0$ and 2, performs equally well than the best of the two at $h = 3$ and outperforms the least accurate of the two at the remaining horizons ($h = 1$ and 4).

If the secondary evaluation tool (MAE) is considered instead of the RMSE, the conclusions remain for the two first horizons. The combination, however, exhibits a poorer performance at the three largest horizons, and in particular at the two

Table 6.5: Forecast performance of the Bayesian model averaging combination compared to survey forecasts

	RMSE	RMSE ratios with respect to ARIMA			
	TS forecasts	Survey forecasts		Bayesian model averaging	
Horizon	ARIMA	SPF	Greenbook	SPF_GR	ARIMA_SPF_GR
$h = 0$	3.199	0.599	0.440	0.439	0.441
$h = 1$	3.384	0.743	0.701	0.717	0.719
$h = 2$	3.373	0.828	0.835	0.824	0.831
$h = 3$	3.337	0.906	0.897	0.897	0.897
$h = 4$	3.174	0.919	0.902	0.903	0.904
	MAE	MAE ratios with respect to ARIMA			
$h = 0$	2.176	0.676	0.457	0.455	0.458
$h = 1$	2.286	0.803	0.773	0.777	0.780
$h = 2$	2.374	0.841	0.861	0.845	0.855
$h = 3$	2.287	0.906	0.931	0.932	0.934
$h = 4$	2.246	0.907	0.923	0.924	0.923

Notes: For each horizon, the RMSE and MAE are calculated with 156 forecasts. TS forecasts are based on ARIMA models selected by the AIC through an exhaustive search. Survey forecasts report the RMSE and MAE ratios of the predictions made by the experts from SPF and Greenbook. For the Bayesian model averaging combination, SPF_GR corresponds to the weighted combination of the SPF and Greenbook data sets, whereas ARIMA_SPF_GR refers to the weighted combination of the ARIMA, SPF and Greenbook forecasts. Both survey forecasts and Bayesian model averaging combination present respective RMSEs and MAEs with respect to TS forecasts.

last horizons where the combination performs slightly worse than any of the two underlying forecasts.

6.5 Comparison of combination methods

So far, we have analyzed each forecasting combination vis-à-vis the survey forecasts and time series forecasts, without any link across the different methods. We now proceed to an overview of all the combinations covered in this thesis, where the performances of the combinations are compared to each other. This global comparison

will stress the benefits of combining forecasts, and will be useful to determine which combination technique should be put forward depending on the time horizon.

Since for all methods combining only the two professional forecasts yields better results than including the ARIMA forecasts, we compare the methods combining only the survey forecasts (denoted by SPF_GR in each table in subsection 6.1 to 6.4).[19] Furthermore, only the primary evaluation tool is considered and the secondary one is left out of the analysis as the conclusions do not substantially differ. Results of all combinations are depicted in table 6.6 and table 6.7. Note that the survey forecasts are excluded from the ranking presented in table 6.7, so to show the best and worst *combination* methods in comparison to each other and not in comparison with the underlying forecasts.

Table 6.6: Forecast performance of all combination methods compared to survey forecasts

	RMSE ratios with respect to ARIMA								
	Survey forecasts		EW	Median	IMSE	PLS		AE	BMA
Horizon	SPF	GR				(i)	(ii)		
$h = 0$	0.599	0.440	0.493	0.543	0.471	0.442	0.470	0.544	0.439
$h = 1$	0.743	0.701	0.702	0.729	0.702	0.725	0.715	0.715	0.717
$h = 2$	0.828	0.835	0.816	0.843	0.817	0.856	0.828	0.859	0.824
$h = 3$	0.906	0.897	0.887	0.907	0.889	0.936	0.914	0.961	0.897
$h = 4$	0.919	0.902	0.894	0.922	0.895	0.906	0.888	0.971	0.903

Abbreviations: GR = Greenbook, EW = equal-weighted average, IMSE = inverse MSE-weighted average, PLS = predictive least squares, (i) = up to date, (ii) = last 4 forecasts (see subsection 5.3.1 and 6.3.1 for further details), AE = adaptive expectations, BMA = Bayesian model averaging. Notes: For each horizon, the RMSE is calculated with 156 forecasts. All RMSE ratios are computed with respect to ARIMA forecasts provided in previous tables. Survey forecasts report the RMSE ratios of the predictions made by the experts from SPF and Greenbook.

[19] The only exception is for the median combination, which in this case combines all three set of forecasts, that is, ARIMA, SPF and Greenbook forecasts since the median of the two survey forecasts would be equal to the equal-weighted average.

Table 6.7: Forecast performance of all combination methods, presented as best (most accurate) and worst (least accurate) method by time horizon

Horizon	Best	(RMSE ratio)	Worst	(RMSE ratio)
$h = 0$	BMA	(0.439)	AE	(0.544)
$h = 1$	EW = IMSE	(0.702)	Median	(0.729)
$h = 2$	EW	(0.816)	AE	(0.859)
$h = 3$	EW	(0.887)	AE	(0.961)
$h = 4$	PLS (ii)	(0.888)	AE	(0.971)

Abbreviations: EW = equal-weighted average, IMSE = inverse MSE-weighted average, PLS = predictive least squares, (i) = up to date, (ii) = last 4 forecasts (see subsection 5.3.1 and 6.3.1 for further details), AE = adaptive expectations, BMA = Bayesian model averaging. Notes: For each horizon, the RMSE is calculated with 156 forecasts. All RMSE ratios are computed with respect to ARIMA forecasts provided in previous tables. Survey forecasts are excluded from this ranking.

Many points are interesting to note from tables 6.6 and 6.7, three of them are highlighted.

First, we observe from Table 6.6 that the IMSE yields quite similar results than the EW. On a theoretical ground, the IMSE is expected to perform better as the combination takes into consideration past performance to determine the future weights. In this case, a low-performing forecast will be less determinant in future predictions and vice-versa. Therefore, one can wonder how a very simple method such as the EW performs equally well when compared to a more complex combination like the IMSE. The answer lies in the weights assigned to each individual survey forecast for the IMSE. These weights are plotted in Figure 6.1 to 6.5, corresponding to $h = 0$ to $h = 4$. These weights correspond to the weights over time, depending on the accumulated historical performance of each survey forecast. For instance, Figure 6.1 shows that, from 1980 to 2010, the forecasts provided by SPF receive approximately 35% of the total weight, while the ones from Greenbook receive circa 65% of the weight. This difference means that forecasts from Greenbook were more accurate

than the ones from SPF, which is confirmed by the data. However, as we can see from the figures, the difference in the weight assigned to the two survey forecasts is always smaller at other horizons, and even becomes close to zero at $h = 2$ and 3 (see Figures 6.3 and 6.4 where both weights approach 50% for almost the entire sample). At these horizons, we observe that the RMSE ratios of both combination techniques are similar. Given these weights, the similarities between the performance of EW and IMSE are now more expected. A conclusion can however be drawn from this comparison; the more divergent the underlying forecasts, the more one should use the IMSE when combining forecasts. On the opposite, the more similar the survey forecasts, the more one should select EW.

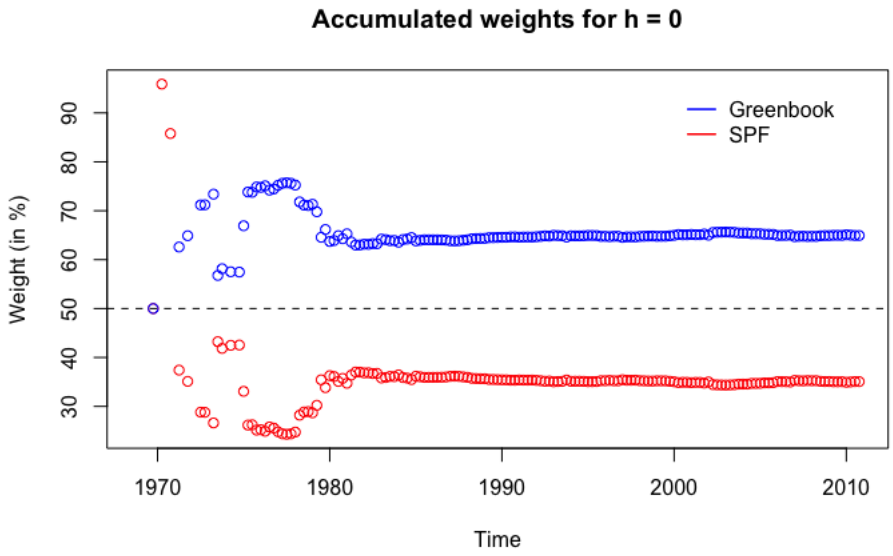

Figure 6.1: Weights assigned to the different survey forecasts in the inverse MSE-weighted average combination for $h = 0$, based on the accumulated MSEs

Second, Table 6.7 shows that no combination *constantly* outperforms all other combinations. Indeed, some are better at low horizons, while others are more accu-

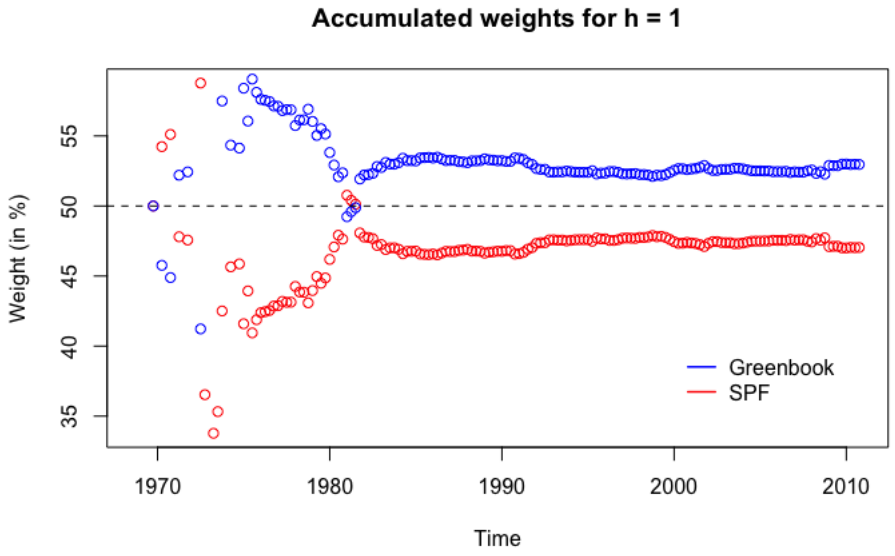

Figure 6.2: Weights assigned to the different survey forecasts in the inverse MSE-weighted average combination for $h = 1$, based on the accumulated MSEs

rate at large horizons. For instance, the BMA method seems to be the best combination at the lowest horizon, while the PLS is the best at the largest horizon, and the EW outperforms all other combinations in the medium term ($h = 1$, 2 and 3). The fact that the EW dominates the other combinations for three horizons out of five is in line with the literature, *i.e.*, that simple combination techniques often produce relatively good results and are even sometimes hard to beat. The best results are also achieved by the IMSE at $h = 1$. On the contrary, the AE is the least accurate combination for all but one horizon ($h = 1$), where the worst results are produced when considering the median. Although the AE seemed to be a promising avenue for combining forecasts, its low performance with this data set casts some serious doubts about its efficiency.

Third, this ranking, together with the findings in Table 6.6 demonstrate, nonethe-

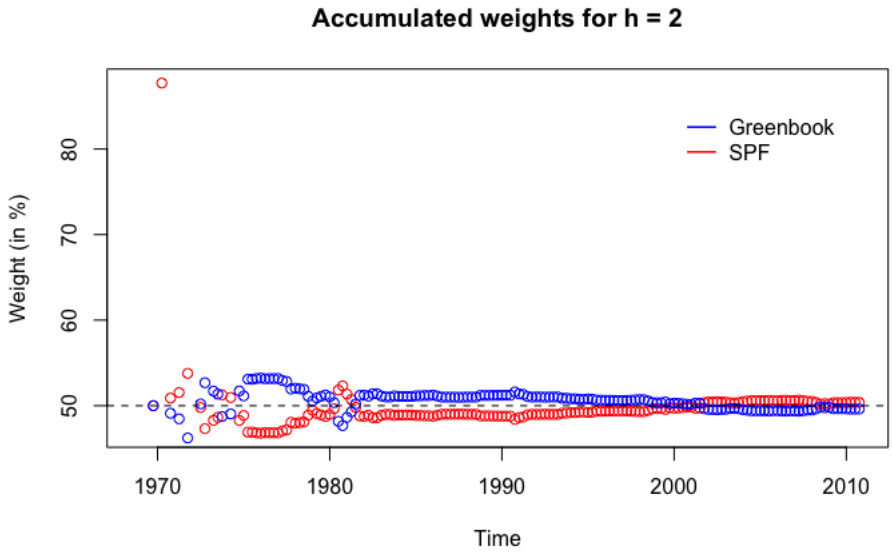

Figure 6.3: Weights assigned to the different survey forecasts in the inverse MSE-weighted average combination for $h = 2$, based on the accumulated MSEs

less, that some combinations are on average better than others, and that some combinations are able to sometimes outperforms both survey forecasts, or at least constantly outperforms the least accurate of the two survey forecasts. This is particularly the case with the EW and IMSE, and to a smaller extent, the BMA. Indeed, out of five horizons considered, the EW and IMSE combinations outperforms the least accurate of the two survey forecasts for two horizons ($h = 0$ and 1) and beat both survey forecasts for the three remaining horizons ($h = 2$, 3 and 4). The BMA is slightly less accurate, but still among the top combinations since the BMA outperforms the least accurate of the two survey forecasts for two horizons out of five ($h = 1$ and 4), performs equally well than the most accurate of the two survey forecasts ($h = 3$), and beats both survey forecasts for the two remaining horizons. These findings have major policy implications. Indeed, from an economic point of view, it is of high inter-

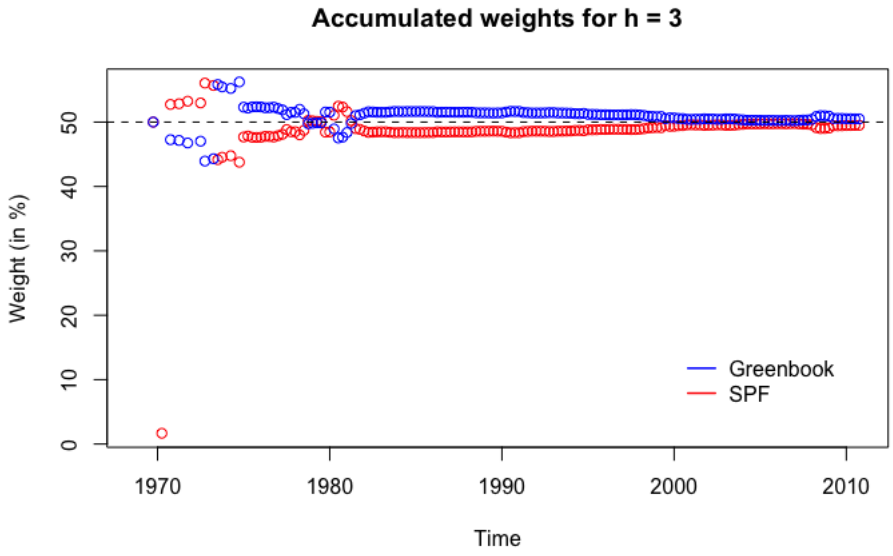

Figure 6.4: Weights assigned to the different survey forecasts in the inverse MSE-weighted average combination for $h = 3$, based on the accumulated MSEs

est to be able to limit losses due to major errors in predictions. Therefore, economic agents usually prefer to avoid risks and would thus prefer to rely on combinations that are always better than the worst underlying forecasts, rather than reckon on underlying forecasts that sometimes outperform all other predictions but otherwise produce the worst estimates. This makes even more sense if the forecast user is risk-averse or if it is preferred to be inaccurately right all the time than being sometimes precisely right and sometimes completely wrong.

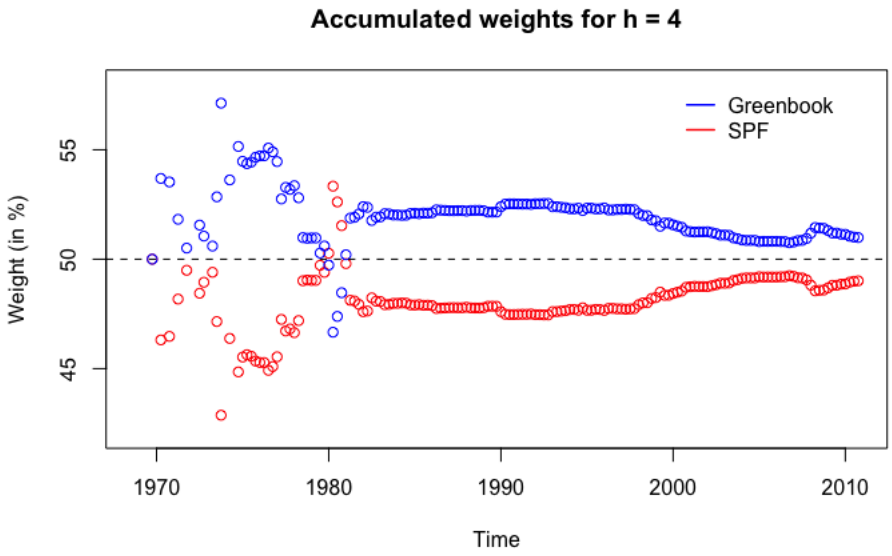

Figure 6.5: Weights assigned to the different survey forecasts in the inverse MSE-weighted average combination for $h = 4$, based on the accumulated MSEs

Section 7

Conclusion

This thesis attempted to shed light on forecast combination for macroeconomic variables and illustrate whether forecast combinations can provide useful and unique information.

Our results contribute to the growing literature regarding forecast combinations but more importantly, it shows that forecast combinations should not be narrowed to inflation forecast combination and that it should be extended to GDP forecast combinations. This major contribution is already enough to draw policy makers' attention on the topic of forecast combination when considering GDP expectations and forecasts in their decisions and targets. Prior research has shown that combining several forecasts usually produces lower forecasting errors and more accurate estimates, and that it can be achieved with rather simple combination methods. We find that it is indeed the case; combining SPF and Greenbook survey forecasts yields lower RMSE at all horizons from nowcasts to four quarters ahead predictions, except at $h = 1$ where the Greenbook forecasts (the most accurate of the two survey forecast) only very slightly outperforms a combination (the improvement equals 0.001 in

terms of RMSE ratios compared to EW and IMSE combinations). In particular, we find that (i) the BMA outperforms all other combinations for nowcasts, (ii) the EW combination dominates for two and three quarters ahead predictions, and that (iii) the PLS combination is superior for four quarters ahead forecasts.

In the introduction we asked ourselves the following questions: (1) Are different sources of GDP survey forecasts comparable? (2) Do the survey forecasts of GDP outperform conventional time series models in forecasting? (3) Can we benefit from combining GDP forecasts from different sources, such as the survey forecasts and model based forecasts? (4) Which combination methods can be used to combine GDP forecasts from different sources, and how do different combination methods perform in forecasting GDP growth?

For the first question, it has been shown that two different sources of GDP survey forecasts can be comparable. The survey forecasts provided by SPF and Greenbook were more or less similar, with a slight superiority for Greenbook in terms of forecasting accuracy for the period covered.

For the second question, it has been demonstrated that the survey forecasts of GDP outperform conventional time series models in forecasting. One explanation for this finding could be that forecasters are already using conventional time series model in their forecasts, and that they use judgments to improve their accuracy.

For the third question, which answers one of our hypotheses made in Section 2.1, the results show that the combination of survey and time series models does not actually outperform the sole combination of survey forecasts. The present analysis, nonetheless, shows that we can benefit to a relatively large extent from combining GDP forecasts from different survey forecasts.

For the fourth question, it has been shown that the BMA, EW, and PLS combinations can be used to combine GDP forecasts from different sources, each method

being most appropriate for a different horizon. Moreover, these best performing combinations decrease the RMSE ratios with respect to time series forecasts by 0.001 to 0.014 compared to the most accurate of the underlying survey forecast.

These results are based on a relatively simple framework which allows us to conservatively conclude that forecast combination has some merits. It would be interesting, nonetheless, to consider even more sophisticated approaches to evaluate several additional issues. I highlight two of them. First, more than two sets of survey forecasts could be used. Depending on the quality of the underlying professional forecasts, this number may vary. Indeed, there is a trade off between including more individual forecasts so to capture more valuable information and including too much information such that the noisy information cancels out the benefits of combining diversified forecasts. For this purpose, the main economic institutions should make their historical data more easily available, and perhaps try to standardize the structure of their data. For this thesis, data scarcity was the main obstacle of combining more than two sets of survey forecasts. The SPF and Greenbook survey forecasts are however the most important survey forecasts and are widely used in the forecasting literature. Second, this thesis intended to test several combination techniques with data on US GDP growth. Nonetheless, one may wonder whether this analysis could be extended to other macroeconomic variables,[20] and whether it could be applied globally (*i.e.*, across countries). These issues are left for future research.

[20]Other than inflation as much work has already been done in the forecasting literature regarding this measure.

Bibliography

Adelman, M. A. (1962). *The economics of petroleum supply: Papers by MA Adelman, 1962-1993*. MIT press.

Agnew, C. E. (1985). Bayesian consensus forecasts of macroeconomic variables. *Journal of Forecasting*, 4(4):363–376.

Aiolfi, M., Capistrán, C., and Timmermann, A. G. (2010). Forecast combinations. *CREATES research paper*, (2010-21).

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second international symposium on information theory*, pages 267–281.

Ang, A., Bekaert, G., and Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of monetary Economics*, 54(4):1163–1212.

Armstrong, J. S. (1984). Forecasting by extrapolation: Conclusions from 25 years of research. *Interfaces*, 14(6):52–66.

Armstrong, J. S. (1986). The ombudsman: research on forecasting: A quarter-century review, 1960-1984. *Interfaces*, 16(1):89–109.

Avramov, D. (2002). Stock return predictability and model uncertainty. *Journal of Financial Economics*, 64(3):423–458.

Baffigi, A., Golinelli, R., Parigi, G., et al. (2002). *Real-time GDP forecasting in the euro area*, volume 456. Citeseer.

Banbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the real-time data flow.

Banternghansa, C. and McCracken, M. W. (2010). Real-time forecast averaging with alfred. *FRB of St. Louis Working Paper No.*

Başẗurk, N., Çakmakli, C., Ceyhan, S. P., and Van Dijk, H. K. (2014). Posterior-predictive evidence on us inflation using extended new keynesian phillips curve models with non-filtered data. *Journal of Applied Econometrics*, 29(7):1164–1182.

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Or*, pages 451–468.

Bischoff, C. W. (1989). The combination of macroeconomic forecasts. *Journal of Forecasting*, 8(3):293–314.

Bohara, A., McNown, R., and Batts, J. T. (1987). A re-evaluation of the combination and adjustment of forecasts. *Applied Economics*, 19(4):437–445.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Chan, Y. L., Stock, J. H., and Watson, M. W. (1999). A dynamic factor model framework for forecast combination. *Spanish Economic Review*, 1(2):91–121.

Cheng, C.-H., Chen, T.-L., Teoh, H. J., and Chiang, C.-H. (2008). Fuzzy time-series based on adaptive expectation model for taiex forecasting. *Expert systems with applications*, 34(2):1126–1132.

Chong, Y. Y. and Hendry, D. F. (1986). Econometric evaluation of linear macro-economic models. *The Review of Economic Studies*, 53(4):671–690.

Clemen, R. T. (1987). Combining overlapping information. *Management Science*, 33(3):373–380.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583.

Clemen, R. T. and Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, 4(1):39–46.

Clements, M. and Hendry, D. F. (2002). *Pooling of forecasts*. Nuffield College.

Corradi, V., Fernandez, A., and Swanson, N. R. (2009). Information in the revision process of real-time datasets. *Journal of Business & Economic Statistics*, 27(4):455–467.

Cremers, K. M. (2002). Stock return predictability: A bayesian model selection perspective. *Review of Financial Studies*, 15(4):1223–1249.

Croushore, D. (2010). An evaluation of inflation forecasts from surveys using real-time data. *The BE Journal of Macroeconomics*, 10(1).

Diebold, F. X. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, 5(4):589–592.

Diebold, F. X. and Lopez, J. A. (1996). Forecast evaluation and combination.

Diebold, F. X. and Mariano, R. S. (2012). Comparing predictive accuracy. *Journal of Business & economic statistics*.

Diebold, F. X. and Pauly, P. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, 6(1):21–40.

Doppelhofer, G., Miller, R. I., et al. (2004). Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. *The American Economic Review*, 94(4):813–835.

Elliott, G. (2004). Forecast combination with many forecasts. Technical report, Mimeo, Department of Economics, University of California, San Diego.

Faust, J. and Wright, J. H. (2009). Comparing greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business & Economic Statistics*, 27(4):468–479.

Faust, J. and Wright, J. H. (2013). Forecasting inflation. *Handbook of economic forecasting*, 2(Part A):2–56.

Figlewski, S. and Urich, T. (1983). Optimal aggregation of money supply forecasts: Accuracy, profitability and market efficiency. *The Journal of Finance*, 38(3):695–710.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Garcia, J. A. (2003). An introduction to the ecb's survey of professional forecasters. *ECB Occasional Paper*, (8).

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

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

Golinelli, R. and Parigi, G. (2008). Real-time squared: A real-time data set for real-time gdp forecasting. *International Journal of Forecasting*, 24(3):368–385.

Gunter, S. I. and Aksu, C. (1989). N-step combinations of forecasts. *Journal of Forecasting*, 8(3):253–267.

Guo-yong, G. (2008). The application of arima model in forecasting gdp of shenzhen [j]. *Mathematics in Practice and Theory*, 4:011.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series for forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3).

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kang, H. (1986). Unstable weights in the combination of forecasts. *Management Science*, 32(6):683–695.

Kmenta, J. (1986). Elements of econometrics.

Koop, G. and Potter, S. (2004). Forecasting in dynamic factor models using bayesian model averaging. *The Econometrics Journal*, 7(2):550–565.

Makridakis, S. (1989). Why combining works? *International Journal of Forecasting*, 5(4):601–603.

Makridakis, S. and Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476.

Mankiw, N. G., Phelps, E. S., and Romer, P. M. (1995). The growth of nations. *Brookings papers on economic activity*, 1995(1):275–326.

Nelson, C. R. (1984). A benchmark for the accuracy of econometric forecasts of gnp. *Business Economics*, pages 52–58.

Palm, F. C. and Zellner, A. (1992). To combine or not to combine? issues of combining forecasts. *Journal of Forecasting*, 11(8):687–701.

Panas, E. and Ninni, V. (2000). Are oil markets chaotic? a non-linear dynamic analysis. *Energy economics*, 22(5):549–568.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, pages 111–163.

Reifschneider, D. L., Stockton, D. J., and Wilcox, D. W. (1997). Econometric models and the monetary policy process. In *Carnegie-Rochester Conference Series on Public Policy*, volume 47, pages 1–37. Elsevier.

Rünstler, G., Sédillot, F., et al. (2003). Short-term estimates of euro area real gdp by means of monthly data. Technical report.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Sessions, D. N. and Chatterjee, S. (1989). The combining of forecasts using recursive techniques with non-stationary weights. *Journal of Forecasting*, 8(3):239–251.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle*. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Song, Q. and Chissom, B. S. (1993). Fuzzy time series and its models. *Fuzzy sets and systems*, 54(3):269–277.

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical methods in medical research*, 6(2):103–114.

Stock, J. H. and Watson, M. W. (1998). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. Technical report, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2001). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. *Festschrift in Honour of Clive Granger*, pages 1–44.

Stock, J. H. and Watson, M. W. (2003). How did leading indicator forecasts perform during the 2001 recession? *FRB Richmond Economic Quarterly*, 89(3):71–90.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Su, V. and Su, J. (1975). An evaluation of asa/nber business outlook survey forecasts. In *Explorations in Economic Research, Volume 2, number 4*, pages 588–618. NBER.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, 1:135–196.

Tkacz, G. (2001). Neural network forecasting of canadian gdp growth. *International Journal of Forecasting*, 17(1):57–69.

Wang, Y. and Lee, T.-H. (2014). Asymmetric loss in the greenbook and the survey of professional forecasters. *International Journal of Forecasting*, 30(2):235–245.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79.

Willmott, C. J., Matsuura, K., and Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, 43(3):749–752.

Winkler, R., Murphy, A., and Katz, R. (1977). The consensus of subjective probability forecasts: Are two, three,, heads better than one. In *Preprint. 5th Conference on Probability and Statistics*, number 15-18, pages 57–62.

Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting*, 5(4):605–609.

Wright, J. H. (2009). Forecasting us inflation by bayesian model averaging. *Journal of Forecasting*, 28(2):131–144.

Xie, W., Yu, L., Xu, S., and Wang, S. (2006). A new method for crude oil price forecasting based on support vector machines. In *International Conference on Computational Science*, pages 444–451. Springer.

Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20(01):176–222.

Zarnowitz, V. (1984). The accuracy of individual and group forecasts from business outlook surveys. *Journal of Forecasting*, (3):11–26.

Zarnowitz, V. et al. (1967). An appraisal of short-term economic forecasts. *NBER Books*.

Zellner, A. (1986). Biased predictors, rationality and the evaluation of forecasts. *Economics Letters*, 21(1):45–48.

Zhang, G. P. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175.

Appendix

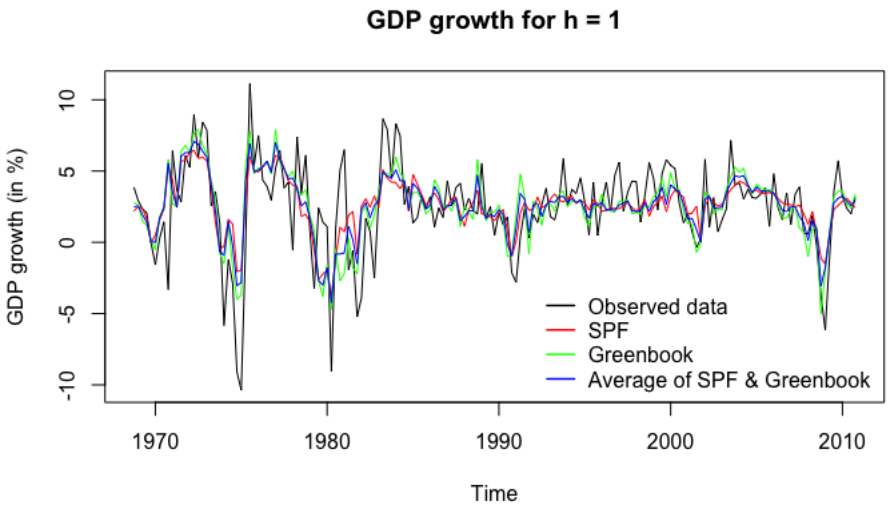

Figure 7.1: GDP growth forecasts provided by SPF, Greenbook, the average of the two for $h = 1$ and observed data from 1968:4 to 2010:4

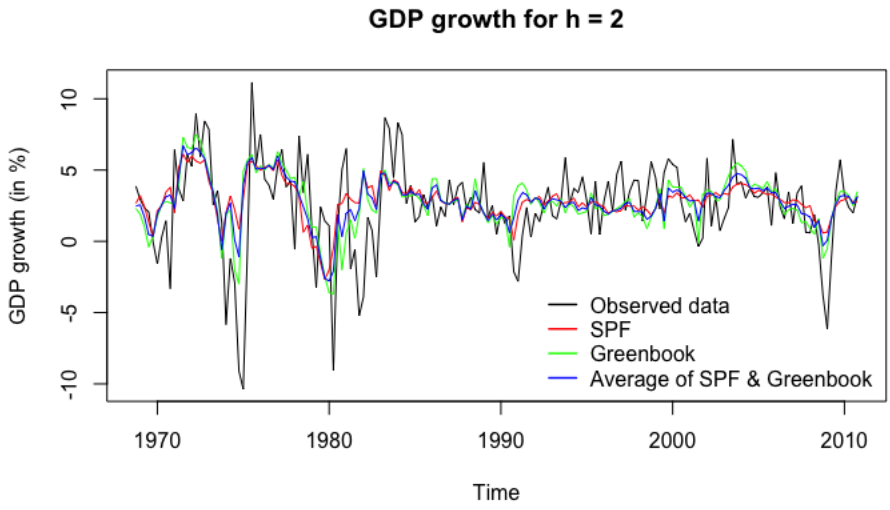

Figure 7.2: GDP growth forecasts provided by SPF, Greenbook, the average of the two for $h = 2$ and observed data from 1968:4 to 2010:4

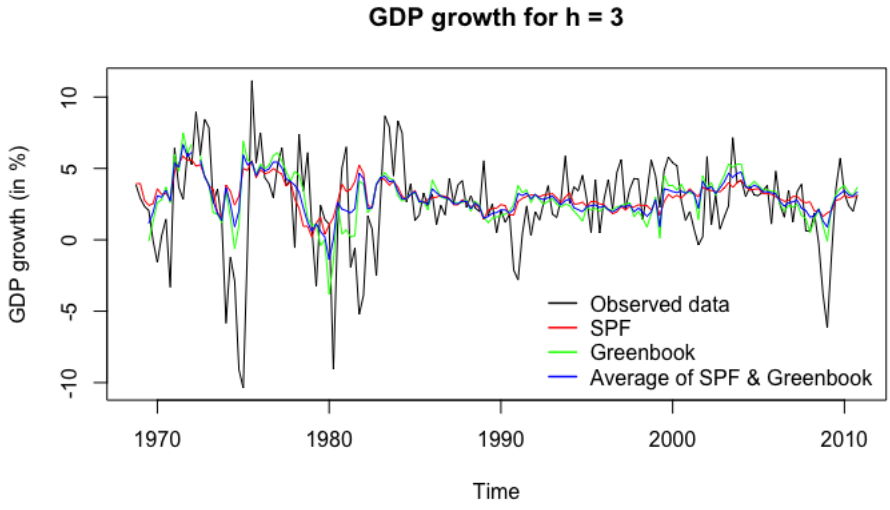

Figure 7.3: GDP growth forecasts provided by SPF, Greenbook, the average of the two for $h = 3$ and observed data from 1968:4 to 2010:4

Table 7.1: Forecast performance of the adaptive expectations combination compared to survey forecasts

	RMSE	RMSE ratios with respect to ARIMA					
	TS forecasts	Survey forecasts		Adaptive expectations			
				(i) last forecast		(ii) up to date	
Horizon	ARIMA	SPF	Greenbook	SPF_GR	+ARIMA	SPF_GR	+ARIMA
$h = 0$	3.199	0.599	0.440	0.544	0.891	0.597	0.997
$h = 1$	3.384	0.743	0.701	0.715	0.840	0.719	0.994
$h = 2$	3.373	0.828	0.835	0.859	0.997	0.807	0.996
$h = 3$	3.337	0.906	0.897	0.961	1.004	0.865	0.987
$h = 4$	3.174	0.919	0.902	0.971	1.028	0.914	0.998
	MAE	MAE ratios with respect to ARIMA					
$h = 0$	2.176	0.676	0.457	0.594	0.856	0.675	0.994
$h = 1$	2.286	0.803	0.773	0.781	0.902	0.774	0.993
$h = 2$	2.374	0.841	0.861	0.873	0.985	0.826	0.993
$h = 3$	2.287	0.906	0.931	0.965	0.995	0.874	0.984
$h = 4$	2.246	0.907	0.923	0.967	1.028	0.899	0.999

Notes: For each horizon, the RMSE and MAE are calculated with 156 forecasts. TS forecasts are based on ARIMA models selected by the AIC through an exhaustive search. Survey forecasts report the RMSE and MAE ratios of the predictions made by the experts from SPF and Greenbook. For the AE combination, both selection criteria are reported; highest MSE (i) for the last forecast, and highest MSE (ii) up to date. Moreover, SPF_GR corresponds to the weighted combination of the SPF and Greenbook data sets, whereas +ARIMA refers to the weighted combination of the ARIMA, SPF and Greenbook forecasts. Both survey forecasts and AE combination present respective RMSEs and MAEs with respect to TS forecasts.

Maastricht University

School of Business and Economics