NORGES HANDELSHØYSKOLE
Bergen, Spring 2012

How accurate are individual forecasters?

An assessment of the Survey of Professional Forecasters

Øyvind Steira

Advisor: Krisztina Molnár

Master thesis – Major in Financial Economics

NORGES HANDELSHØYSKOLE
Norwegian School of Economics

Abstract

This master thesis addresses the forecast accuracy of individual inflation forecasts from the Survey of Professional Forecasters. Based on a variety of accuracy statistics, there are five main findings to report of. First, I find that some individuals are able to accurately predict inflation over time, and that forecasters on average have improved their accuracy over time. Second, forecasting accuracy becomes worse during recessions compared to the average accuracy in the respective decades but accuracy have improved in newer recessions compared to old ones. Nonetheless, some individuals are able to outperform the mean and a random walk model. Third, I find no difference in accuracy among industries, but I find evidence for biased forecasts for the three and four quarter horizon. Fourth, I find evidence for bias in roughly one-third of the individuals for all forecasting horizons. These results improve slightly when only data from the last two decades are being analysed. Fifth, the majority of individuals perform significantly worse than a random walk model regardless of used time span.

I also find several problems with the database. These includes: missing values for the one-year-ahead forecast, irregularities in forecasters' response, reallocation of used ID's, changing base year and inconsistencies in individuals' forecasts.

Preface

This master thesis completes my five years of studies at the Norwegian School of Economics, and also my time as a resident in Bergen. As such, it represents much more than just a thesis; it stands for the end of an era, and the beginning of a new one.

Given that my specialization in the master's degree has been financial economics, it was natural for me to write a paper on a subject concerning macroeconomic perspectives. That being said, it was more a coincidence, or a twist of fate, that the thesis ended up with inflation expectations and the accuracy of inflation forecasts as a subject. Nevertheless, it was an incredible interesting topic which also entailed some work with statistical programming. This combination suited me perfectly. Hopefully, it also contributed with some insight to the subject at hand.

The paper is written as a part of a research program called crisis, restructuring and growth (KOV-project), which has been a motivating experience. I would like to express my appreciation for being selected as a participant in this project, and also thank all participants in this group for good feedback.

At times the work has been demanding, and I would like to thank my supervisor Krisztina Molnár for useful comments and advice. Erik Sørensen and Øyvind Anti Nilsen have been helpful with statistical problems which seemed unsolvable. Finally, I would like to thank my fiancée for valuable help and for always having my back.

Bergen, June 15th 2012

Øyvind Steira

Contents

Introduction

Over the years a large amount of studies on inflation expectations from surveys have accumulated. One of the oldest surveys available in the U.S. is the Survey of Professional Forecasters (SPF) which started in 1968 and is now conducted by the Federal Reserve Bank of Philadelphia.[1] Surveys have undergone extensive testing by economists and have undoubtedly participated greatly in the economic research the past 40 years. They have been used to test rational expectations theory, to analyze the formation of inflation expectations, in empirical research in macroeconomics, to investigate the formation and impact of monetary policy, and in a variety of other studies (Croushore 2009).

The importance of inflation expectations has been heftily debated even though it plays a crucial role in many economic agents' decisions (Elliott and Timmermann 2008; Mankiw *et al.* 2003). In an interview study of public attitudes towards inflation, Shiller (1997, cited in Shiller (2000)) showed that the general public pays a lot of attention to inflation, and it is widely believed that the inflation rate is a barometer of the economic and social health of a nation. He found that people had great feelings toward inflation, and perceived high inflation "as a sign of economic disarray, of a loss of basic values, and a disgrace to the nation, an embarrassment before foreigners" (Shiller 2000, p. 37). Consequently, it is reasonable to believe that people pay attention to the variable and is able to give realistic forecasts. Is it really so? Are individual forecasters able to accurately predict future inflation rates?

In this paper I will attempt to make an assessment of the forecast accuracy for the forecasts in the SPF database. In order to separate my thesis from the vast literature which already exists on the subject, I have made some choices. First, due to the large emphasis on longer forecasting horizons in previous studies, especially the one-year-ahead horizon, I have chosen to keep the main focus on shorter horizons. Second, I will mainly focus on individual inflation forecasts and not the consensus view which is more common. Third, I will use the GDP price index as my measure of inflation in contrast to the consumer price index which is more frequently used.

[1] Prior to 1992, when the Federal Bank of Philadelphia took over the survey, it was called the ASA-NBER Economic Outlook Survey. For simplicity I will only call it the SPF henceforth.

In this paper I find that some forecasters are able to predict inflation accurately over time, and it also seems as they are getting better at it over time. This conclusion does not hold for all forecasters, however. Next, forecast accuracy gets worse during recessions compared to the average accuracy during the respective decades. Some individuals, on the other hand, perform well in recessions and outperform a same change random walk model. Further, I find no difference in accuracy between industries but it seems as all three industries are biased for the three and four quarter horizons (at a five percent significance level). This is to my knowledge not documented before and thus represents an important finding. I also find evidence for some biased individuals, but the majority of forecasters are unbiased for all horizons (about 2/3 of the individuals). Last, most individuals, especially for the three longest horizons, do not add additional information to the forecast given by the same change random walk model. In other words, most of the forecasters fail to outperform the benchmark model. No newer papers have compared the random walk model and survey forecasts against the GDP price index. This result is still somewhat striking. Other studies find that survey forecasts outperform times series models when forecasting CPI inflation (Ang *et al.* 2007). Some do find, however, that the random walk model performs very well for some measures of inflation which could explain its good performance in this paper (Atkeson and Ohanian 2001).

The outline of this paper is as follows. First I will present a theoretical framework with terms used in the paper and arguments for and against survey data. The second part will consist of the methodology used to assess the forecast accuracy, followed by part three which presents some characteristics with the data. The fourth part concerns the SPF database, and includes an introduction and a section on potential problems and caveats with the dataset. Next, in the fifth part, I present the results of my analysis concerning forecast accuracy. Sixth and last, I will give my concluding remarks and give some direction for future research.

Literature review

There are two characteristics which are widespread in most of the literature on inflation expectations from surveys. First, studies have shown that pooling or combining data into a mean (often called “consensus” (Gregory *et al.* 2001)) creates a more consistent and accurate forecast (Batchelor and Dua 1995). Thus, most studies use the consensus forecast when studying

expectations. Second, due to the large revisions of the national income and product account (NIPA) variables (such as GDP) there can be problems if these are used as a measure of inflation (Croushore 2006). Consequently, most researchers studying the SPF have used a variety of the CPI as their inflation measure, after it was introduced in 1981q3[2].

One of the first studies on the SPF database was conducted by Su and Su (1975), who assessed the accuracy of forecasts using only a few years of data. They found that forecasts from the database were significantly better than autoregressive extrapolations. They also stated that the SPF forecasts are better at forecasting changes in the levels of the data than the levels themselves. Some years later, Hafer and Hein (1985) compared the accuracy of three different inflation forecasting procedures; an univariate time series model, an interest rate model and forecasts from the SPF. Their general conclusion was that the median survey forecasts of the implicit GNP deflator provided the most accurate ex ante inflation forecasts, even though they used data from the most volatile period in the whole survey time span.[3] Nevertheless, their results were in line with those of Su and Su, namely that SPF forecasts outperformed simple time series models. Another economist who has tested the SPF database extensively is Victor Zarnowitz. One of his first studies on the SPF forecast accuracy also included tests on an individual level for the first time. He argued that only using means or medians raised the possibility of aggregation errors such as differences among individuals and sampling variation (Zarnowitz 1984). Even though he acknowledged the importance of examining individual data, he still concluded that the consensus forecasts on average over time are more accurate than most individual forecasts and that this conclusion was valid for all variables and horizons. He also said that those individuals who did outperform the consensus had no common characteristics. Later, an even broader and more comprehensive study of the SPF database were conducted by Zarnowitz and Braun (1993). Here they provided a wealth of analysis on the database, with several important findings. First, they documented that forecast errors typically increase as the horizon increase, which is logical since there is more uncertainty associated with predicting development in macroeconomic variables further into the future. Second, they also stated that forecasters differed in many respects and therefore also their forecasts would differ.

[2] Here q denotes quarter, and this notation should thus be read as first quarter in 2010. It will be used throughout the paper.

[3] During the 1970s and early 1980s the U.S. experienced a high inflationary period, with severely high growth. See section 3.3.1.

Nevertheless, they found some common trends among the individuals and argued that this was due to common information sets and interaction and influence with fellow forecasters. Third, they found great differences in the extent to which macroeconomic variables can be forecasted. Variables with high autocorrelation (such as real GDP) are easier to predict than those which are highly random (e.g. business inventories). Fourth, they found no evidence for an improvement in forecasting over time, despite an improvement in computer technology and access to more modern economic theories. Fifth, they underlined the findings from earlier studies that group consensus outperform the majority of individuals and thus represents an accessible and inexpensive method for improving forecasts from individuals. Last, they demonstrated that consensus survey forecasts perform favorably in comparison to most simple time-series models.

There have also been conducted numerous studies testing the survey for bias, i.e. if the forecast errors are zero on average. Such tests are also imperative to prove rationality among individuals, i.e. if forecasters make repeated errors over time or not. The first tests on the Livingstone survey were not positive, as they suggested forecasters were biased and as a consequence not rational (Pearce 1979; Akhtar *et al.* 1983). However, in a study on the SPF database the conclusion was opposite: Zarnowitz (1985) concluded that 85 percent of individuals were unbiased. On those which were biased, half were forecasts of inflation. These results of biasedness and irrationality provided forecasters with a bad reputation, and many economists started to believe that forecasters in fact were irrational or that surveys were not representative for market agents` real inflation expectations (Croushore 1996, 2006). That being said, there were other reasons for the bad performance during these years. First, there were unexpected OPEC oil shocks in the early 1970s which drove up the rate of inflation. This was very hard to predict, which is why most forecasters performed badly and seemed biased during these times (Croushore 1993). Second, researchers were not aware of a problem economists today call the overlapping observations problem. When a shock hits the inflation variable, it affects not only one survey but several consecutive surveys. The reason for this is that the length of the forecast horizon normally is longer than the interval between the surveys, thus making the forecast errors correlated. By not taking this into consideration in their tests, researchers overstated the case against the surveys (Croushore 2009). In a later study on rationality, Keane and Runkle (1990) tested individuals from the SPF database and stated that much of the preceding literature on rationality were flawed for four reasons. First, the use of consensus forecasts was wrong because individuals

may have differing information sets. Second, they did not put enough effort in correct for revisions in the underlying data. Third, data from the Michigan survey were not trustworthy due to lack of incentive for the respondents to be rational in their responses, and fourth, past researchers failed to account for correlation in forecast errors across forecasters. When they dealt with all these previous mistakes they found that forecasters were unbiased and efficient, in contrast to most previous studies.

Newer research papers have chosen other interesting topics for their studies. Mehra (2002) argues, among others, that the predictive ability of a forecaster has more to it than just outperforming a simple naïve benchmark (which, according to Mehra, is what earlier studies have assumed). He uses the test of Granger-causality to determine if the survey contains additional information about the subsequently realized inflation values than the past values. His findings suggest that survey forecasts do in fact Granger-cause inflation, meaning they can help predict actual future inflation. He also concluded that forecasters from the SPF were biased. Another interesting study was conducted by Mankiw *et al.* (2003). They discovered substantial disagreement among forecasters, i.e. that forecasts given for the same variable and horizon can vary substantially among individuals. They believe that this oversight can be explained by the fact that standard theory does not open for disagreement. By using a sticky-information model, in which forecasters only periodically update their expectations due to high costs of collecting and processing information, they can explain much of the disagreement present in the data.

A fairly new study by Ang *et al.* (2007) compare and contrast four methods of predicting inflation: time-series models; regressions based on the Phillips curve using measures of economic activity; term-structure models derived from asset prices; and surveys. They conclude decisively that the survey-based measures yield the best results for forecasting CPI inflation, which seems to be in line with previous comparisons between surveys and time series models.

1. Theoretical framework

Before turning to the analysis of the data, it is important to have some insight on the formation of expectations and why it is so important. This section will provide a brief explanation of these questions, and will also present some arguments for why I have chosen individual data from the SPF as my data.

1.1 Expectations

1.1.1 Why are expectations important?

Expectations are very important for most people, even though many may be unaware of it. Almost everyone use expectations as a foundation for making everyday decisions, e.g. consumers, businesses, investors and authorities (Elliott and Timmermann 2008). Consumers alter their spending and saving based on the economic outlook, more specifically on factors such as future employment level and wage growth. Businesses use their expectations of future income and profitability to make investment decisions and to decide what strategy they are going to pursue. Investors use their expectations as a basis for decisions on what kind of assets to invest in, when to invest and much more. Most importantly, perhaps, is the importance of expectations for authorities` decisions. For example, it is crucial for central banks to take into consideration what expectations the consumers have when making policies, and it has a great deal of influence on wage negotiation (Thomas 1999). All of these decisions, which are based on expectations, will in turn affect the growth and inflation level in the economy.

Inflation expectations have been an especially popular topic among researchers. The reason for this is probably because of the central banks` introduction of inflation targeting. It then became vital to check what people actually think and reveal if they use the information they should in order to make correct forecasts. Kershoff and Smit (2002) stated that almost every central bank with an inflation target studies inflation expectations surveys when forming monetary policies. This even goes for countries without a formal inflation target, like the U.S. If people do not make accurate inflation forecasts it means that they do not manage to make use of all the relevant information in order to predict the future. Thus, it becomes equally difficult for the

authorities to alter the peoples` expectations in order to affect the economy. A relevant example can be drawn from the recent economic crisis. As the economic growth plundered, and the employment level surged, people were starting to expect harder times and therefore started to cut their spending and increase their saving. The housing market bubble also left a lot of people without savings, thus reducing their spending even more. All of this affected the economy negatively and reduced growth further. This negative spiral would have continued if the authorities did not implement policies which altered people`s expectations and incentives to increase their spending. Low interest rates is one example of a policy which was meant to convince people that measures were being made to save the economy, and thus reducing the negative expectations people had.

1.1.2 The formation of expectations

Discovering how economic agents form their expectations is critical to our understanding of many economic outcomes. In the earlier years, when the term expectation was introduced, the common belief was that expectations were formed solely by looking at historical values of the variable; a so-called adaptive formation. Today, this thought is rejected by most researchers and is viewed to be too simplistic. A new theory arrived and argued that agents are rational, which simply means that forecasters employ all available information when forming future expectations (Akhtar *et al.* 1983). The underlying principle behind this change of thought was that agents are intelligent, and thus are able to correct for mistakes made in the past when predicting the future. This leads to the first of two characteristics of a rational forecaster: they do not systematically make errors. An important test in this regard is a test for bias, i.e. a test to check if the average forecasting error is equal to zero. The second characteristic concerns the issue of efficiency. In order for a forecaster to be efficient, he/she have to make use of all relevant information when forming their expectations. In this paper I will only focus on the first of these two characteristics, namely the forecasting accuracy.

Even though rational expectations have been widely accepted as the best way of describing the formation of expectations, there have been some critics which proclaim that the rational expectation formation theory was too easily accepted. Chow (2011), for instance, argues that there was insufficient empirical evidence for accepting the rational expectations hypothesis and

gives strong econometric evidence supporting the adaptive expectations hypothesis. It is probably more reasonable to believe that the formation follows a path which lies between the two extremes of adaptive and rational formation (Roberts 1998).

1.2 The Survey

In this paper all analysis will be based on survey data from the SPF.[4] In this section I will present some general arguments for and against using this kind of data material based on earlier literature.

1.2.1 Why use survey data?

Surveys are a method for collecting data from a chosen sample of the public. The sample can then be used to make statistical inferences about the population. In this case a survey is used to ask a certain group of people about their predictions for the GDP price index variable, among others. The mean of the responses can be interpreted as a consensus for the expected inflation rate.

According to Galati *et al.* (2011) surveys are one of two major methods to get hold of inflation expectations if one wants to work with such data. The first method consists of extracting inflation expectations using financial market instruments linked to some measure of inflation, e.g. bonds. If combined with a nominal counterpart one can back out financial markets` inflation expectations. It comes with a drawback, however, because it can be a bit technical to do the calculations in order to extract the expectations. What is more, one also has to take into consideration inflation-risk premium and liquidity premium (among others) which also increases the difficulty of this method. The second method is to use survey data, i.e. ask participants in the market what they believe (or expect) future inflation will be. This approach entails less knowledge about technical procedures, is easy to interpret and there are several surveys being conducted for several countries which are ready for use. However, as Galati *et al.* (2011) points out, it comes with some shortcomings. First, most surveys have a low frequency on their data

[4] See section 4 for more on the SPF database.

making them less suited for analysis concerning existence and timing of breaks in formations over short horizons. Second, they question the reliability of respondents as there is no way to make sure that they actually live up to their predictions. Hopefully, this problem will be less prominent when using professional forecasters (more on this in next section). Third, it is also pointed out that different surveys provide totally different results on inflation expectations. In a study undertaken by Mankiw *et al.* (2003), where they looked at over fifty years of data on inflation expectations in the U.S., they found substantial disagreement among both consumers and professional economists about expected future inflation. Nevertheless, due to its simplicity and easy access, survey data seems like the most reasonable choice.

1.2.2 Why use professional forecasters?

Surveys can be conducted on many different types of groups. In the U.S., for example, they have surveys asking household consumers (the Michigan Survey), businesses (the Livingstone Survey) and professionals (the SPF).

As stated by Gerberding (2006), participants in household surveys are more likely to have an opinion on the expected direction of future inflation than they are to give a precise predicted change for different horizons. In other words, she presents an argument in favour of qualitative data. In order to do an empirical analysis on such data, however, one needs to do a transformation to quantitative data which will inevitably bring along some uncertainty in the data. This is not likely to be a problem when using surveys with professional forecasters. They produce forecasts in their daily jobs, and should therefore be qualified to do a quantitative response to the survey. What is more, they also have a strong incentive to do a proper analysis before they turn in their answers as wrong answers may create some stigma in their professional life. This cannot be said of household or business (to some extent) surveys as they do not have to defend their answers in the same way. The same argument is underlined by Keane and Runkle (1990) who argue that professional forecasters predict the same expectations which they sell in the market and thus have an incentive to be accurate. What is more, others, e.g. Mestre (2007) and Ang *et al.* (2007), also conclude that professional forecasters outperform other agents on inflation expectations. Choosing professional forecasters as a source of data thus seems as a reasonable choice.

1.2.3 Why use individual data?

Most of the literature on inflation expectations in surveys makes use of the mean or median forecasts in their studies. No wonder, since almost all articles on the matter conclude that consensus forecasts are superior to individual forecasts. A study by Bates and Granger (1969) was one of the first studies which concluded that a combined set of forecasters can result in a higher accuracy than either of the original forecasts. Further, in a review of the literature on combining forecasts where over 200 articles were studied, Clemen (1989) found that forecast accuracy can be substantially improved through the combination of multiple individual forecasts. Newer research has reached similar conclusions. Batchelor and Dua (1995), for example, stated in their paper that individual responses may contain behavioural biases which could be removed if pooled together (in Batchelor (2000)).

There are those who argue for using individual data. Zarnowitz (1984) studied the accuracy of individual and group forecasts, acknowledging the importance to study both sides. Nonetheless, he concluded that "the group mean forecasts [...] are on average over time more accurate than most of the corresponding sets of individual predictions. This is a strong conclusion [...]" (Zarnowitz 1984, p. 15). Keane and Runkle (1990), in their study on rationality of individuals, gave a sharp critique of earlier studies on the subject. One of their arguments is that averaging individual forecasts will mask individual deviations from the consensus. If one group of people consistently make positive errors while another consistently make negative errors, the mean will become unbiased. They argue that the information given by the deviating groups are too important to loose in averaging all forecasts.

Lately, there have been very few papers analyzing individual data. This makes it intriguing to investigate what affects individual forecasting accuracy under different scenarios or during different time periods. This paper will therefore focus on the forecast accuracy of individual forecasts.

2. Methodology

This section will provide an outline of the methods used to assess the performance of the SPF database. This includes different measures for analyzing the forecast accuracy, and some statistical tests for comparing the performance between two different sources of forecasts.

2.1 Forecast accuracy

When talking about "the best" forecasting method one often interprets this as the forecasting method which is most accurate, i.e. result in the smallest error. There are several methods for evaluating a forecasts` accuracy, but most of them are calculated by comparing the values of the forecast against the actual (real) values of the same series.[5] The forecast error is therefore defined as

$$e_t = A_t - F_t$$

where A_t is the actual (real) value of the variable in question at time t, and F_t is the forecasted value at time t.

I will use four different forecasting horizons in my analysis. They will range from a one quarter horizon to a four quarter horizon. The actual forecasts are calculated as follows:

$$forecast\ 1q = 100 * (pgdp3/pgdp2 - 1)$$

$$forecast\ 2q = 100 * (pgdp4/pgdp2 - 1)$$

$$forecast\ 3q = 100 * (pgdp5/pgdp2 - 1)$$

$$forecast\ 4q = 100 * (pgdp6/pgdp2 - 1)$$

Where pgdp1-pgdp6 are the actual level forecasts given by the respondents in time *t-1* to *t+4* (i.e. pgdp1 is the inflation level for last quarter, pgdp2 is the inflation level for the current quarter, pgdp3 forecast for next quarter etc.).

[5] See for example Batchelor (2000), Mehra (2002) and Zarnowitz and Braun (1993)

In this section I will discuss different ways of measuring forecast accuracy. More specifically, I am going to make use of three different measures: 1) mean absolute error, 2) mean prediction error and last, 3) root mean squared error. When comparing the individual forecasts with those from the time series models, I will use Theil's U-statistic and a forecast comparison regression.

2.1.1 Mean absolute error

The first measure discussed is mean absolute error (MAE):

$$MAE = \sum_{t=0}^{N} \frac{|e_t|}{N}$$

where N is the total number of observations and t denotes time. This measure is preferred if one think the error is linear, rather than quadratic, because it ignores the sign of the error. This implies that a forecast error which is one too low represents just as much as a forecast error which is one too high. The closer MAE is to zero, the more accurate the forecast is.

2.1.2 Mean prediction error

The second measure of forecast accuracy used in this paper is the mean prediction error (MPE):

$$MPE = \sum_{t=0}^{N} ({}^{e_t}/_{N})$$

This measure is a simple average of the forecasting errors and hence should be close to zero over a time period in order for a forecast to be unbiased. A positive value indicates that the forecaster have underestimated actual values, while a negative MPE indicates that forecasters have overestimated actual values.

2.1.3 Root mean squared error

The third, and last, measure discussed in this section is the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=0}^{N} (e_t^2)}$$

This measure is computed by squaring all errors, thus removing the sign of the error. The average of all errors are calculated (producing mean squared errors, or MSE), and as the name suggests, RMSE is the square root of MSE. The main difference between MAE and RMSE is the assumption of the characteristic of the error. In contrast to the MAE, RMSE assumes a quadratic error. This implies that an error of two percent is treated four times (2^2) as serious as an error of one percent (in contrast to MAE where a two percent error is treated as twice as serious as a one percent error, because of assumed linearity in the error). Therefore, the RMSE put a larger penalty on forecasters who make a few large errors, relative to forecasters who make a larger number of small errors (Batchelor 2000). The forecast accuracy improves as the RMSE moves closer to zero.

2.1.4 Theil's U-statistic

Theil's U-statistic is a simple measure on how well a model performes compared to a naive time series model. The idea behind the rule is that if a forecast is to be taken seriously, it should be more accurate than the forecast given by a simple benchmark. The measure compares the RMSE of the two models, as the definition shows:

$$Theil`s\ U = \frac{RMSE\ of\ forecast}{RMSE\ of\ benchmark\ model}$$

In such a model, a value equal to one means that the two models have identical RMSE and thus are equally accurate. A value above (under) unity implies that the forecast (benchmark model) have a higher RMSE, and thus have performed worse (i.e. been less accurate).

2.1.5 Benchmark model

To assess the performance of the forecasts it is not enough to just look at accuracy statistics. Even if the forecast accuracy is terrible, it could still be characterized as a good forecast if no other forecasting methods are able to perform better. Thus, bad accuracy may still imply a decent performance relative to other methods. A common approach to account for this is to compare the survey`s forecast to a benchmark model. In this paper I will use a simple random walk model (RW) as a benchmark. According to this model the forecast for this quarter`s change in inflation for a given horizon will simply be the change experienced in last quarter for

the same horizon:

$$RW: F_{t,x} = A_{t-1,x}$$

where F_t is the forecast for the current quarter at time t for horizon x, and A_{t-1} is the actual value from last quarter for horizon x. Since the actual data is the percentage change in inflation the random walk model will represent a "same change model", i.e. the forecast will be equal to the change in last quarter (in contrast to a "no change model", where the forecast will represent no change in the level data).

Previous studies have proven that this model performs reasonably well when forecasting inflation, as it even outperforms other more sophisticated time series models for some inflation measures (Ang *et al.* 2007; Atkeson and Ohanian 2001). It therefore seems as a legitimate choice to use this as a comparison to the survey forecasts.

An expected advantage for this model is that it will be good at predicting turning points. While forecasters will have to analyse the economic situation based on numerous variables in order to precisely predict and time the actual turning point, the random walk model will automatically predict the turning point one quarter after it happened since it only bases it prediction on the first lagged value. On the other hand, the model will perform badly if the inflation rate has abrupt changes between high and low, as the model then will be unsynchronized with the actual values.

2.2 Statistical tests

2.2.1 Forecast comparison regression

In order to statistically distinguish one forecasting model from another, one can perform a forecast comparison regression. The regression line in this situation will be:

$$\pi_{t,x} = \beta \cdot f_{t,x}^{SPF} + (1-\beta) \cdot f_{t,x}^{RW} + \varepsilon_{t,x}$$

where $f_{t,x}^{SPF}$ is the forecast of $\pi_{t,x}$ from the SPF database, $f_{t,x}^{RW}$ is the forecast from the naive benchmark model, and $\varepsilon_{t,x}$ denotes the forecast error associated with the combined forecast. Further, t denotes time and x represent the forecast horizon. If β=0, then forecasts from the SPF

database add nothing to the forecasts from the benchmark model, and we thus conclude that the naive model outperforms the SPF model. If β=1, then forecasts from the random walk model add nothing to the forecasts from the survey, and we then conclude that the SPF forecasts outperform the benchmark model. In order to indicate better performance by the SPF forecasts, I will test if the null of β equal to zero is rejected and thus conclude that β is significantly different from zero. This is in line with both Stock and Watson (1999) and Ang *et al.* (2007).

To my knowledge it is not possible to restrict coefficients when performing a Newey-West regression in Stata. Thus, I had to rearrange the regression line in order to perform the analysis:

$$\left(\pi_{t,x} - f_{t,x}^{RW}\right) = \beta \cdot (f_{t,x}^{SPF} - f_{t,x}^{RW}) + \varepsilon_{t,x}$$

2.2.2 Bias

A bias test confirms if the forecast errors are centered on the correct value or if they systematically diverge from the real values of inflation. In other words: it tests if the forecasters systematically over- or underestimate inflation. A common approach to conduct such a test is to do a simple regression on the following equation:

$$A_t = \alpha + \beta F_t + \varepsilon_t$$

where A_t are the actual values of the inflation variable, α is the constant term, F_t is the forecast in question and ε_t the corresponding standard error. Subsequently, the null hypothesis of no bias is tested, i.e. if α=0 and β=1 holds. If these conditions are not rejected, it suggests an unbiased forecast. It has been shown, however, that the conditions α= 0 and $\beta = 1$ is not necessary for F to be an unbiased forecast of A. Holden and Peel (1990) show that by regressing forecast errors on a constant and test whether the constant can be restricted to zero, we get a condition that is both necessary and sufficient for unbiasedness. This method is also used by Mankiw *et al.* (2003), who re-arrange the original regression line above to the following:

$$A_t - F_t = \alpha + \varepsilon_t$$

Thus, the necessary condition in order for a forecast to be unbiased is the null hypothesis of α=0. If the null hypothesis is rejected the individual will be characterized as biased.

3. Data

This section will provide a brief introduction of the dataset from the SPF, followed by a section explaining the autocorrelation and heteroscedasticity in the data. Finally, a presentation of the actual data (i.e. the GDP variable) will be given.

3.1 Descriptive statistics

Due to the volatility in the GDP variable it is inevitable that some periods have been more turbulent than others when it comes to change rates in the inflation variable. This also leads to highly different standard deviations in the different time periods. The standard deviation has a useful purpose when assessing forecast accuracy, because it can be interpreted as a direct measure for the difficulty of forecasting in each period (McNees 1992). It will then be possible to compare forecasts given in different periods with different degrees of difficulty. The figures below illustrate the development of the standard deviation of inflation change over time for different forecast horizons. Figure 1 shows the standard deviation of the one-year-ahead forecasts from the SPF across time. From this figure, one can see that the inflation forecasts have become less erratic and volatile over time. Figure 2 shows the standard deviation for the real change in inflation for the four different horizons. It illustrates that the 1970s and 1980s were the most difficult periods to forecast in. After this there was a sharp decrease in the standard deviation, and inflation in the 1990s and 2000s ought to have been much easier to predict. For the two shortest horizons the 1980s have been most difficult to predict.

The number of participants who have responded to the survey have varied over its life time, as can be seen from figure 3. It also highlights the dwindling participation up to the closure of the survey, before the Federal Reserve Bank of Philadelphia took over responsibility for the execution. In the beginning the participation was very high, with a maximum of over 60 participants. After 1990, the number has become lower and it seems as the participation stabilized around the total mean of about 35.

Figure 1 and figure 2: Average standard deviation for the four quarter forecast from the SPF (left) and standard deviation per decade for real inflation change for all forecast horizons (right)

Figure 1:

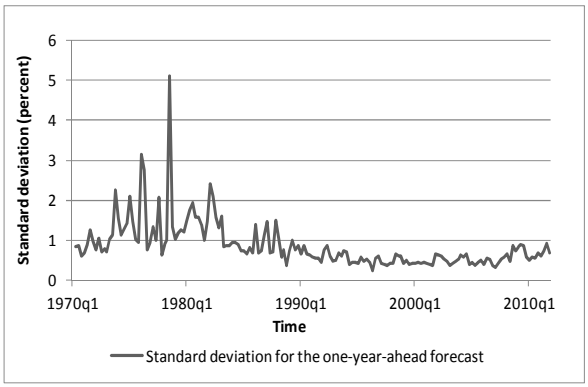


Figure 2:

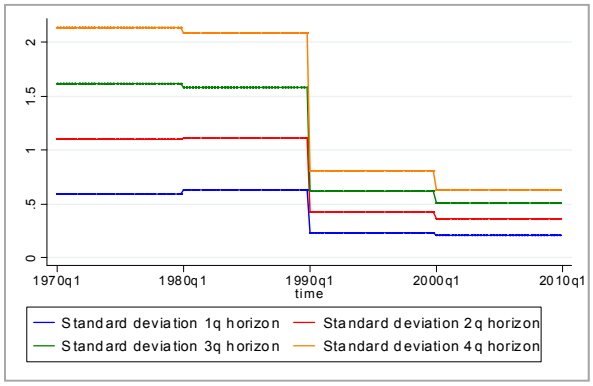
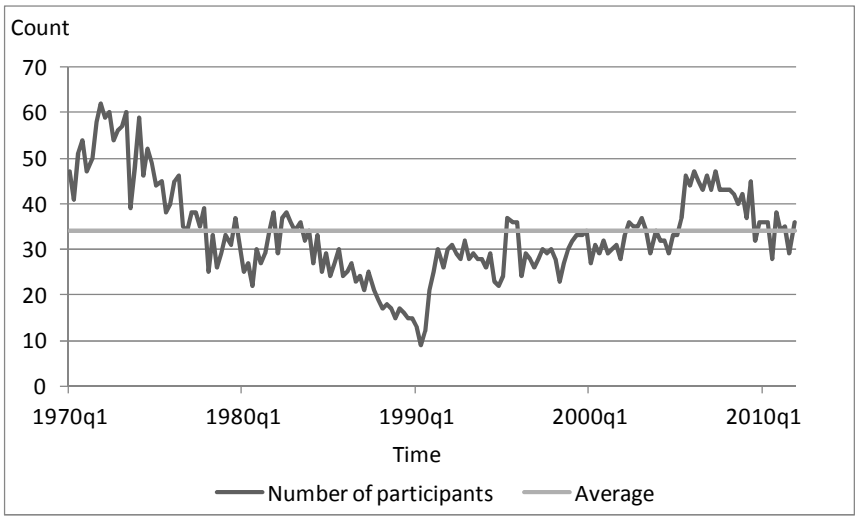


Figure 3: The number of participants in the survey across time



When doing an analysis on individuals it is best to have long uninterrupted series of responses to examine if the forecasters are able make accurate forecasts over time. It is possible for everyone to have a lucky guess a quarter or two in a row, but a forecaster who gives accurate forecasts quarter after quarter for a long time has much more credibility and a higher justification to be called accurate. Panel A in table 1 shows the five longest consecutive forecasting series, who they belong to and when it happened, while panel B gives some information on how many series which fall into different bins of varying length. As we can see, the longest series of consecutive responses is 49 quarters given from 1990q1 until 2002q2. Next individual has given 40 consecutive responses in the 2000`s, followed by two forecasters in the 1980`s and 1990`s with

36 consecutive responses. These will be well suited for analysis in the following sections. Panel B gives important information on how the situation is further down on this ranking. It shows, as we have seen in panel A, two series which are longer or equal to 40 quarters. Further, we have eight series which fall into the bin consisting of series between 30 and 39 responses, 24 series ranging from 20-29 responses and 127 series with length from 10-19 quarters. In other words, one can observe several individuals with an adequate amount of consecutive responses to test accuracy over time.

Table 1: The five longest consecutive series of responses

Panel A:				Panel B:	
ID	Maxrun	From	To	Number of responses	
65	49	1990q1	2002q2	>39	2
510	40	2001q4	2011q4	30-39	8
70	36	1980q3	1989q3	20-29	24
30	36	1981q2	1990q2	10-19	127
433	34	1990q4	1999q2	1-9	1244

Note: Panel A shows the longest series of consecutive responses (maxrun) given by a forecaster. Panel B shows number of consecutive series that fall into different bins of varying length.

3.2 Autocorrelation and heteroscedasticity

An inevitable characteristic of a survey like the SPF is the issue of overlapping observations (Croushore 2006; Grant and Thomas 1999).When testing forecasts over an equal or longer horizon than the sampling frequency of the data (e.g. one-year-ahead forecasts while the sampling frequency is quarterly) one need to take into consideration that a shock affects several of the underlying quarters. If an inflation shock affects actual data in 2010q1, this means that forecast errors from 2009q1 up until 2010q1 are all correlated.

Autocorrelation in the errors is a violation on the assumptions behind regular ordinary least square (OLS) regressions, making results from this kind of analysis spurious (Granger and Newbold 1974). OLS assumes that the errors in the regression are uncorrelated, normally distributed and have a constant variance (homoscedasticity). The last assumption is also most likely violated in our data set, because some individuals are more accurate than others and

because some periods are harder to predict than others. This implies heteroscedasticity in our data (in addition to autocorrelation).

A solution to this problem, taken from Croushore (2006), is to adjust the covariance matrix as shown by Newey and West (1987) and thus guarantee a positive definite covariance matrix. This will overcome the problem of heteroscedasticity and autocorrelation in the errors in the dataset. Practically, this will imply running a Newey-West regression with heteroscedasticity and autocorrelation consistent (HAC) standard errors when doing the forecast comparison regression and bias test mentioned in section 2.2.2 (p.19). This method will prevent any problems with autocorrelation and heteroscedasticity in the error terms of our data.[6]

3.3 Actual data

This paper will assess the predictive accuracy of inflation forecasts from individual forecasters. Inflation is known as an increase in the general price level for goods and services within a country over a certain time span. There are a number of diverse variables, which differs both in calculation method and content, which strive to describe the same phenomenon. A suitable question would then be which variable should one use?

In the SPF survey they have forecasts for two main inflation variables, namely the CPI index and the GDP price index[7]. Which variable to choose depends on the purpose of the study. Consumers would be best off using the CPI, as that measure gives the increase in price of a fixed basket of consumer goods. The GDP deflator, on the other hand, is more dynamic and can be used to show new expenditure patterns as it is based on all domestically produced goods in the country. I have chosen to use the GDP deflator as the inflation variable.

When using the GDP deflator one should be aware of that the variable undergoes severe revisions from the first initial calculations. This makes it hard to know which revision one is supposed to use as actual data. Studies have pointed out results demonstrating significant differences in accuracy between using the initial or revised data on GDP (Croushore and Stark

[6] Since our data have a quarterly frequency I will use a lag of four in the Newey-West regressions.

[7] Prior to 1996, GDP implicit deflator. Prior to 1992, GNP deflator (Federal Reserve Bank of Philadelphia 2011).

2001; Stark and Croushore 2002). They found that even though data revisions can have a large effect on its accuracy, it tend to not alter the relative accuracy between the survey and the benchmark projections (Stark 2010). McNees (1992) also concluded that forecasts are much more accurate when compared to preliminary data than they are compared to final revised data. He argues that if the aim is to measure how close forecasters come to what actually happened it is clear revised data is a better estimate of reality. This line of argumentation is perfectly logical; if forecasters cannot predict what actual revised inflation will be, but are only able to predict preliminary inflation, their forecasts are not much use for anyone. With this in mind, I will use the final revised data of the GDP deflator as actual values.

This choice may have some negative effects, considered that the random walk model is based on the final revised version of the GDP price index. It could be argued that this gives an advantage compared to the forecasters who only have knowledge of an preliminary version of the inflation rate when they make their forecasts. Thus, my use of revised data may bias the results against the individual forecasters.

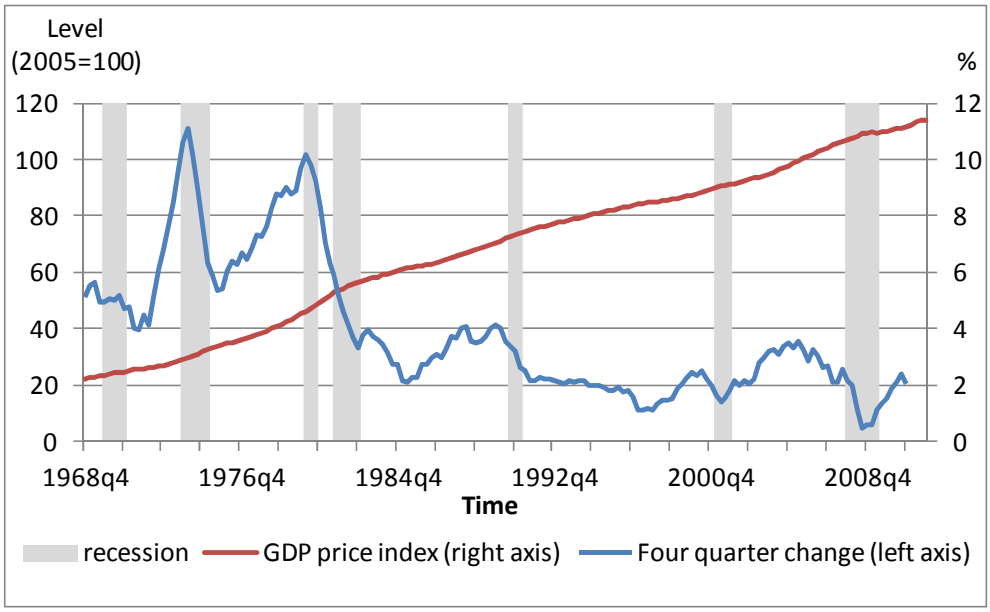
3.3.1 Historical development

In order to explain differences in forecast accuracy over time it is important to see how the inflation variable has developed over the time span of the survey. Over the past 40 years the real GDP deflator has had a striking linear growth, as it started at a level of 20 in 1968 and now has almost reached a level of 120 (see figure 4). By first sight, it seems as though it should cause no problems for forecasters to predict a simple linear trend. On the other hand, if we look at the four quarter ahead actual growth, it becomes more evident that forecasters need some skills in order to predict the actual change (which is what I am measuring in this paper). This highly inflationary period can mostly be explained by politics. In the late 1960s the U.S. was in a recession and it was an election year. To keep a low unemployment level President Nixon pressured the Federal Reserve to keep low interest rates, with the purpose of providing the public with a sense of recovery from the recession. This, however, turned out to be a short-lived satisfaction. In 1972-73 inflation started to rise sharply and it did not come under control until Paul Volcker became chairman of the Fed and introduced a tight monetary policy. This highly disturbing period has also led to the Federal Reserve keeping a more cautious and closer look at

the inflation. Later, the annual change in inflation has been stable in the low single digits, which leads to a more predictable variable.

Figure 4 also depicts all U.S. recessions during the time span of the survey. According to the National Bureau of Economic Research (2010) there have occurred seven recessions, with four of them happening in the first 15 years of the survey.

Figure 4: Development for the GDP price index and U.S. recessions



4. The Survey of Professional Forecasters

This section will contain information regarding the SPF database.[8] I will first do an introduction and a brief summary of the database, before I provide a section concerning the timing of the survey explaining available information at the time forecasts were given. Finally, I will present our work relating to problems and caveats with the database.

4.1 Introduction

The SPF is a quarterly survey started in the fourth quarter of 1968, thus making it the oldest quarterly survey of macroeconomic forecasts in the U.S. It was started as a joint venture by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER), which led to its original name: the ASA-NBER economic outlook survey. Among the variables to be forecasted initially was the change in the GNP deflator, and horizons for 1-4 quarters ahead. They collected forecasts of the GNP deflator from 1968 to 1991, the GDP deflator from 1992 to 1995, and the GDP price index since 1996 (Federal Reserve Bank of Philadelphia 2011). This change in variable causes no severe problems, since the GNP deflator, GDP deflator and GDP price index behave quite similarly and there are no apparent breaks in the forecast series to be seen in either of the years where the change took place (Croushore 2006). The objectives of the survey were first stated in Mincer and Zarnowitz (1969), and later the performance in the first 22 years of operation was assessed in Zarnowitz and Braun (1993).

The survey was very popular in the early years with over 50 participants each quarter. However, as time passed the participation declined so much that it was decided to end the survey in first quarter of 1990. Later the same year it was decided that the survey should be resumed, now under control of the Federal Reserve Bank of Philadelphia. Measures were taken to ensure a higher level of participants and the timing of mailing and collecting the survey was improved in order to make them more consistent over time.

[8] More info and data can be found online at http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/

The respondents are, as the name of the survey implies, professional forecasters. This means the receivers of the survey have forecasting as a part of their job. This includes professors, bankers, consultants and other who have continuous interaction with macroeconomic data in their daily job. Participants are held anonymous in order to encourage people to provide their best forecasts without worrying about potential forecast errors or forecasts which may contradict with their jobs position (Croushore 1993). The survey is mailed to participants the day after the government release of quarterly data on national income and product accounts (NIPA). It asks for point forecasts for many different variables and time horizons.

4.2 Timing

When comparing different series of forecasts it is imperative to take into consideration the timing of the survey to ensure that all parties have the same starting point and the same information set when predicting. After the Federal Reserve of Philadelphia took over the survey in 1990q2 they made sure to maintain a consistent timing of the conduction of the survey (Federal Reserve Bank of Philadelphia 2011).[9] The survey is mailed to all participants just after the release of the NIPA advance report[10], which happens in the first month of the quarter. Included in the survey is a report on recent historical values of different variables from different sources, in order to make sure participants produce their forecasts on the same basis. The survey is due in the second or third week of the middle month each quarter. This implies that the participants can take advantage of information in the variables in question up until this time. There are no official data released during these weeks, however, so the last information the participants have knowledge of is from the preceding quarter. The results of the survey are released in the middle month, well before the next survey is sent to the participants. An example may enlighten the matter further: just after the advanced report is released in 2010q1 (January) the survey for the same quarter is sent to its participants. They now have knowledge to all historic values of the GDP price index up until 2009q4. They provide forecasts for 2010q1,

[9] Their first survey was officially in 1990q2. However, this survey was conducted after the fact because they had not yet received all the information from the NBER/ASA that they needed to conduct it in real time (Stark 2010).

[10] This contains preliminary results on the GDP deflator from the current quarter.

2010q2, 2010q3, 2010q4 and 2011q1. The one-year-ahead forecast comes as a result of their forecast of 2011q1 relative to 2010q1. As Croushore (2006) points out, even though this represents a four quarter period the end of the forecast horizon (2011q1) actually is five quarters after their latest known observation of the GDP price index (2009q4). The table below highlights the most important information regarding the timing of the survey.

When it comes to the timing before the Federal Reserve Bank of Philadelphia took over the survey in the second quarter of 1990 there is no certain documentation of the timing prior to this date (Croushore 2006). However, according to Stark (2010), a recent analysis of the timing when the ASA-NBER was in charge of the survey suggest that the schedule was very similar to the one we can observe now at the Federal Reserve Bank of Philadelphia.[11]

Table 2: Timing of the Survey of Professional Forecasters, from 1990q3 to present.
Source: Federal Reserve Bank of Philadelphia (2011)

Survey Name	*Questionnaires Sent to Panelists*	*Last Quarter of History in the Panelists' Information Sets*	*Date of Deadline for Submissions*[2]	*Results Released to the Public*
First Quarter	End of January (after NIPA advance report)	Q4	Middle of February (second to third week)	Middle to Late February (before NIPA second report)
Second Quarter	End of April (after NIPA advance report)	Q1	Middle of May (second to third week)	Middle to Late May (before NIPA second report)
Third Quarter	End of July (after NIPA advance report)	Q2	Middle of August (second to third week)	Middle to Late August (before NIPA second report)
Fourth Quarter	End of October (after NIPA advance report)	Q3	Middle of November (second to third week)	Middle to Late November (before NIPA second report)

[11] They compared the latest-available historical observation in the survey's data set with the values as they appear in the Philadelphia Fed's real-time data set and found a close correspondence, particularly since 1985.

4.3 Problems with the database

This section will give a presentation of potential problems and caveats one should be aware of when working with the SPF database.[12] It will also contain some proposed solutions which can be implemented to remove or minimize the problems at hand. This analysis is only conducted based on the one-year-ahead forecast horizon.

4.3.1 Irregular forecasters

An inevitable problem with individual data is the irregularity in respondents` participation. Since the survey is based on volunteer participation it is up to each individual to answer the survey or not. It is therefore unavoidable that most respondents at some time will give less priority to the survey if they are preoccupied with work or other business. This problem is also present in this dataset, where one often finds several gaps in an individual`s forecast series. This poses a challenge because it limits the number of long individual forecast series which can be used for analysis. The point with individual accuracy is to study if some individuals perform consistently accurate *over time.* Thus, it is not interesting to do analysis on a forecast series only a few quarters long. It will also make it harder to do statistical analysis as it may require longer data series to gain enough information.

The above-mentioned problem is also pointed out in most previous articles dealing with individual data in the SPF database. In Zarnowitz (1985, 1984) and Keane and Runkle (1990) they remove individuals which have less than 10, 12 and 20, respectively, responses from their dataset. This ensures that the problem becomes less prominent, and it removes the sporadic forecasters who give us no useful information. In order to see how these changes will affect the database, table 3 shows some descriptive statistics regarding the participation in the survey. Panel A shows the number of surveys per respondent, which is equal to the number of quarters the forecasters have responded to the survey. As we can see the average number of surveys increase as the irregular forecasters are removed, providing a dataset more eligible for analysis. The standard deviation of surveys per respondent decreases which is logical since individuals

[12] This section is the result of our work as research assistans for the crisis, restructuring and growth project. It was performed in cooperation with Karen Oftedal Eikill, and she therefore deserves some of the credit.

Table 3: Selected distributional statistics over whole sample and two sub-samples

Responses	All	>12	>20
Obs deleted	-	652	1336
Total obs	6568	5916	5232
Panel A:	Number of surveys per individual		
Total surveys	173	173	173
Mean	41.4	45.2	49.1
Std. Dev.	26.6	25.3	24.4
Min	1	12	20
Max	123	123	123
Panel B:	Number of individuals per survey		
Total respondents	313	176	131
Mean	42.2	41.8	41.3
Std. Dev	13.4	13.3	13.2
Min	9	9	9
Max	83	83	83

Note: Table shows descriptive statistics for different subsamples.

with few responses are removed. The highest number of surveys an individual responded to was 123 (but this does not imply 123 *consecutive* responses). Panel B shows the number of respondents per survey. When eliminating irregular forecasters the total number of unique forecasters decreases, along with the average number of forecasters per survey. This basically means that data on fewer respondents are included than would have been if the whole dataset were used. The changes to the data are not very large which suggests that removing irregular forecasters do not alter the database significantly. In my further analysis in this paper I will use data where those individuals with less than 12 responses are removed.

The descriptive statistics of key variables in the survey also changes when respondents are dropped. Table 4 shows the alteration of these variables when individuals with less than 12 responses are dropped. The mean of the forecasts for pgdp2 and pgdp6 both increase somewhat, from 138.54 to 138.93 for pgdp2 and from 144.21 to 144.59 for pgdp6. Thus, the one- year ahead forecasts also increases, from 3.68 percent to 3.71 percent. The standard deviations also increase slightly, while the minimum and maximum values are the exact same. All in all, the changes in the database due to dropped responses are minor and should have no impact on the coming analysis.

Table 4: Statistics for key variables before and after dropping irregular forecasters

Sample	All			>12 responses		
Statistic	pgdp2	pgdp6	forecast 1yr	pgdp2	pgdp6	forecast 1yr
Observations	6563	6134	6133	5912	5545	5544
Mean	138.54	144.21	3.68	138.93	144.59	3.71
Std	32.03	35.97	2.16	32.53	36.41	2.17
Min	104.41	105.7	-4.57	104.41	105.7	-4.57
Max	235	247	31.14	235	247	31.14

Note: table shows descriptive statistics for key variables, before and after removing irregular forecasters (i.e. with less than 12 responses).

Figure 5 shows a plot of the number of dropped individuals against the time variable. The figure illustrates that there are irregular forecasters over the whole time period, not only confined to a specific part of the survey. The maximum number of dropped individuals in one quarter is 16, which was in 1970q4. After year 2000 there have not been more than seven individuals with less than 12 responses per quarter, which could imply that forecasters have become more regular in recent times. Figure 6 shows how the median one-year-ahead forecast have changed after the removal of irregular forecasters. As one can see, the differences are small with the largest being 0.3 percent in 1990q4.

Figure 5 and figure 6: Number of dropped respondents per quarter (left) and the difference in the median inflation forecast before and after removal of irregular forecasters (right)

Figure 5:

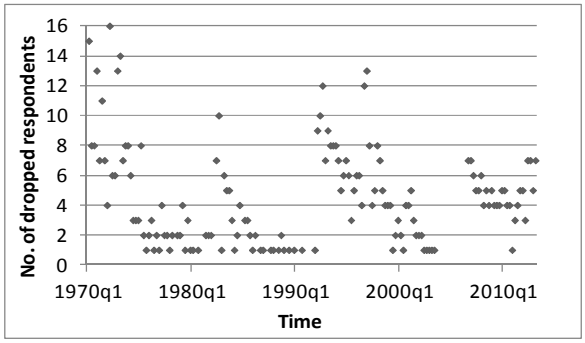
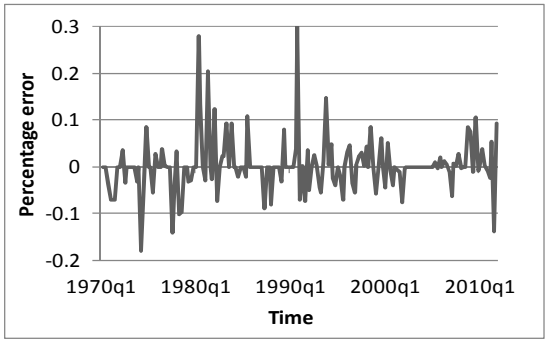


Figure 6:



4.3.2 Missing values

There are some problems with missing values in SPF database. In this context, missing values is not connected to the problem mentioned above with individuals not responding to the survey, but rather when individuals decide not to respond to all questions in the survey. This problem becomes particularly prominent when looking at the average value of the one-year-ahead forecast horizon. According to our inspection, there are five quarters which have no values at all: 1969q1-q3, 1970q1 and 1974q3. It should be noted that this issue does not concern the smaller forecast horizons. The reason for these missing values has to be due to missing values for variables pgdp2 and/or pgdp6.[13] However, examinations of the data material have shown that

[13] See calculation of one-year-ahead forecast in section 2, p. 15.

it is missing values for the pgdp6 variable that is causing the problem. There are two plausible solutions on this problem. First, one can start using the data after the last quarter with missing values for the four quarter forecast, i.e. drop all observations before 1974q4. This would ensure observations in all quarters for this forecast horizon and thus not bias our results. Another advantage with this method is that it demands very little effort to drop the unwanted observations, in contrast to other methods which may require more calculations. It should be noted, however, that this method may result in valuable information being lost. By removing almost 20 quarters with observations, one may get different answers and conclusions than before, especially when we take into consideration the volatility of the inflation variable during those years. Second, one can fill in values for the pgdp6 variable based on the other values the individuals reported on the respective surveys. As mentioned above, some forecasters chose to leave out some information in the surveys and as it turns out it was often only the pgdp6 value which was left out. One possible solution is to do a linear projection based on the other variables (pgdp2-pgdp5). A possible equation could be:

$$pgdp6 = pgdp5 * \left[\left(\frac{pgdp2}{pgdp1}\right) + \left(\frac{pgdp3}{pgdp2}\right) + \left(\frac{pgdp4}{pgdp3}\right) + \left(\frac{pgdp5}{pgdp4}\right)\right]/4$$

This method generates a value for pgdp6 by projecting pgdp5 based on the average growth rate in the other variables. This method entails filling in 396 (340) new values of pgdp6 when keeping (removing) individuals with less than 12 responses. As we can see from table 5, neither method change the basic statistics of the one-year-ahead forecast or pgdp6 much. It also fills in values for pgdp6 in all the quarters which previously had no observations. It therefore seems as a reasonable method to use.

Another method for inspecting if these changes alter the database too much is to compare the median value in the individual database with the median database found on the SPF website. Initially, these two should be, and are, perfectly equal. After filling in values and dropping respondents, however, there are some differences to be seen. Figure 7 shows the difference between the aforementioned series when values are filled in but before individuals with less than twelve responses are removed. It shows minor errors, with the biggest ones being just above 0.1 percent and the errors being dispersed over the whole time span. Figure 8 shows the same as

figure 7, only now we have removed the irregular forecasters. The errors are now larger and more frequent, with the largest errors at about 0.3 percent. Even though we see a worsening picture, it cannot be said to be all that bad. It seems as though filling in values by linear interpolation does not alter the characteristics of the database in a too large extent. Since I am doing analysis on multiple horizons, I have chosen not to implement any of the above-mentioned suggestions.

Table 5: Basic statistics before (top) and after (bottom) removing individuals with less than 12 responses

Variable	Obs	Mean	Std. Dev.	Min	Max
forecast1y~w	5884	3.717201	2.158057	-4.568528	31.13736
forecast1yr	5544	3.708568	2.169319	-4.568528	31.13736
pgdp6new	5885	144.5502	35.98329	105.7	247
pgdp6	5545	144.5942	36.41208	105.7	247

Variable	Obs	Mean	Std. Dev.	Min	Max
forecast1y~w	6529	3.689335	2.144917	-4.568528	31.13736
forecast1yr	6133	3.682137	2.15636	-4.568528	31.13736
pgdp6new	6530	144.1155	35.48895	105.7	247
pgdp6	6134	144.2106	35.96883	105.7	247

Note: Variables ending with new (or ~w) represents new variables after filling in values with above-mentioned procedure.

Figure 7 and figure 8: Difference between the median one-year-ahead forecast from the individual database and the median database from the SPF website, before (left) and after (right) removing individuals with less than 12 responses

Figure 7:

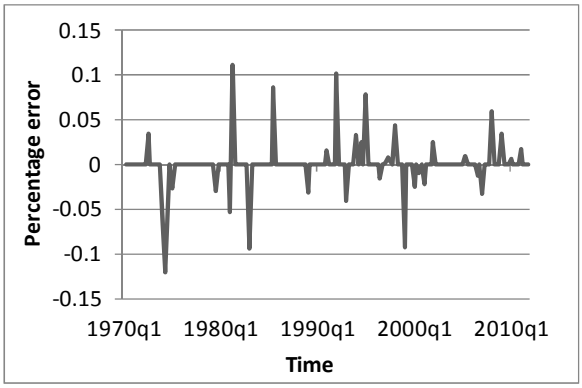
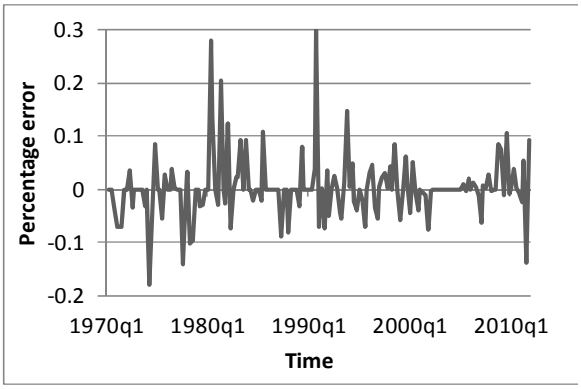


Figure 8:



4.3.3 Reallocation of ID numbers

Another problem with the database relates to the identification (ID) numbers, given to each respondent in order to secure anonymity of that person. Unfortunately, as the documentation paper from Federal Reserve Bank of Philadelphia (2011) suggests, there could be some problems with the allocation of these ID numbers. This entails some negative consequences and one should use some caution when interpreting the identifiers associated with each respondent.

First, it has been discovered occurrences where the same ID number could represent different individuals over the time span of the dataset. In some occurrences a respondent drops out and then reenters several periods later, suggesting that a new individual may have been given an already used number. This is only a problem in the data from when the NBER-ASA was in charge of the survey, i.e. from 1968q4-1990q1. The lack of hard-copy historical records prevents anyone from looking into this problem, thus making it hard to come up with any solutions to the problem. It is guaranteed that this is not a problem after the Federal Reserve Bank of Philadelphia took over the survey in 1990q2. Second, a new problem arose when the Federal Reserve Bank of Philadelphia took over the survey and introduced the industry variable. The question was, if a respondent change jobs (i.e. the industry variable changes) should the ID number follow the respondent or the initial firm? Being aware of the issue, the Federal Reserve Bank of Philadelphia decided on a guideline for how to deal with it. If a forecast seems more associated with the firm than the individual, the ID number stays with the firm and the respondent gets a new ID (and vice versa). Unfortunately, they do not give any information on which ID`s this concerns. When analysing individuals this could be a problem if the identification number is more connected to a firm than to an individual, as tests on ID's in reality will concern the associated firm and not necessarily one specific individual.

Even though there is a vast literature concerning the SPF database, there is to our knowledge not anyone using individual data who discusses the above-mentioned problems.[14] One reason for this is that they mostly use the old data from NBER-ASA and hence were not aware of the problem. Other researchers use the mean or median forecasts in their studies, thus ignoring the

[14] See for example Zarnowitz (1984), Zarnowitz and Braun (1993) and Keane and Runkle (1990)

problems connected with individuals. There has been done little research on individual data in the recent years, which also implies a lack of an up-to-date solution to the abovementioned problem.

One possible solution to this problem could be to divide individuals where large gaps in responses occur into two or more ID`s. The problem with this solution is that the Federal Reserve Bank of Philadelphia is not absolutely sure if ID`s really were re-used, they only know it could be a problem. A pause in responses from an individual does not necessarily mean it is a new person; it could be an individual who deliberately decided to stop responding to the survey for a time, and then started again at a later stage. Nevertheless, if the gap is big enough it could be argued that the respondent would have changed some over the years anyhow and calling her/him a new person would not be a terribly wrong act. When deciding how large the gap should be before the ID is divided up, we should also take into consideration that some respondents can be absent due to natural causes such as child birth and sick leave. Another thinkable scenario is where an individual changed jobs, and thus could not be called a "professional forecaster" anymore before he/she reentered the forecasting business at a later stage. The gap should therefore be large enough to consider such causes, e.g. 5 years (or 20 quarters) or more.

In table 6 we show some statistics concerning gaps in individuals' forecasting series. On average, when all forecasters are included, the average gap in a forecaster's series is 0.82 quarters with a corresponding standard deviation of 4.21. The largest gap is 73 quarters, which constitutes roughly 18 years. Based on this, it seems highly probable that reallocation is present. The table also points out that there are more forecasters without a gap in their response than there are forecasters with gaps (see column 1: "obs"). If only those with gaps are included the average size of the gap is 3.61 quarters with a high corresponding standard deviation of 8.24. One of the solutions mentioned above in section 4.3.2 was to start with data from 1974q4 in order to deal with missing values. This solution would also do some good when it comes to gaps in the forecasting series. As the last row illustrates, the average size on the gap and the corresponding standard deviation goes down when we start at a later time. This suggests that most of the gaps are positioned in the early stages of the survey. Table 6 also gives the same statistics after we have removed irregular forecasters. This measure will improve the gap

statistics for all of the different scenarios included in the table, which put emphasis on the correctness of implementing this measure.

Table 6: Gap statistics before and after removing irregular forecasters

	A: Before removing irregulars				B: After removing irregulars			
	Obs	Mean	Std	Max	Obs	Mean	Std	Max
All	6255	0.82	4.21	73	5740	0.66	3.37	51
With gap	1421	3.61	8.24	73	1229	3.08	6.76	51
No gap	4834	0.00	0.00	0	4511	0.00	0.00	0
From 1974q4	4844	0.48	1.95	46	4477	0.44	1.91	46

Note: Table shows statistics concerning gaps in the individuals` forecasting series, before and after removing irregular forecasters (i.e. less than 12 responses).

Going more in detail, it could be interesting to see how many individuals who in fact have large gaps in their forecasting series. Table 7 illustrates such numbers on gaps from five to twenty years, before and after we have corrected for irregular forecasters. It illustrates that fifty-eight individuals have a gap larger than five years, twenty-nine with gaps larger than ten year, four with a larger gap than fifteen years and none with gaps larger than twenty years. After removing irregular individuals, the numbers become better. There are forty individuals with a gap larger than five years, sixteen with a gap larger than ten years and none with larger gaps. Here we can also see that removing irregular forecasters improve some of the faults in the database.[15]

Table 7: Number of individuals with large gaps

Observations Gap\Time span	All >1968q4	>12 >1968q4	>12 >1974q3
>5 yr	58	40	21
>10 yr	29	16	2
>15 yr	4	0	0
>20 yr	0	0	0

Note: table shows the number of individuals with gaps larger than a given number, for different sub samples.

We also want to examine if the gaps are more prominent in a certain time period, and hence expose if there is a pattern in their occurrences. By plotting the number of gaps larger than five and ten years, as seen in figure 9 and figure 10, we can see that most of the gaps are mainly located in the beginning of the survey. This coincides with the finding in table 6. In

[15] We have to be aware of the fact that there could be individuals with more than one gap. After a closer look, however, we find that there are no individuals who have more than one gap larger than five years (and thus no one with even larger gaps).

1970q3 there are five individuals with a gap larger than five years, which is the highest number of individuals in one quarter. Later there are only a couple of individuals just before 1995 with large gaps. This tells us that the problem is mostly present during the NBER-ASA period of the survey as only three gaps come from the Federal Reserve Bank of Philadelphia period. For gaps longer than ten years, all except two are in the beginning of the survey; just before 1975 and 1995. A possible solution to the problem could then be to start with data from 1974q4. If we look at this subsample, there are only twenty-one observations with gaps longer than five years, and only two with gaps longer than ten years (see table 5). In other words, it reduces the problem to a great extent.

Figure 9 and figure 10: Number of individuals with gaps larger than five (left) and ten years (right) in their forecasting series

Figure 9:

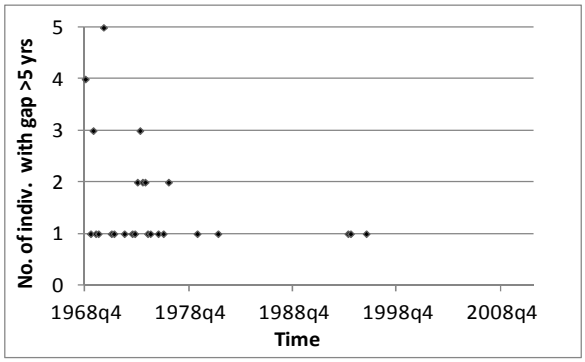
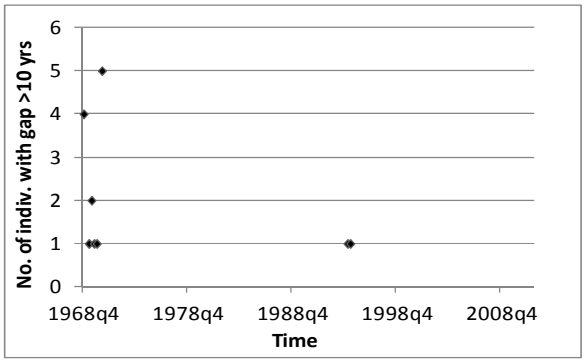


Figure 10:



4.3.4 Changing base year

When working with level data from the SPF database one should be aware of the fact that there are multiple changes in base year for several variables. Every fifth year, when there are benchmark revisions to the NIPA variables, the base year might change in addition to the data being revised. Since the forecasted levels in the data set have not been rescaled after the base year changes, the levels in the SPF dataset use the base year which was in effect when the questionnaire was sent to the forecasters. For the GDP deflator variable there have been eight changes in base year: 1976q1, 1986q1, 1992q1, 1996q1, 1999q4 and 2004q1 and 2005q1 (Federal Reserve Bank of Philadelphia 2011) They are listed in table 8 below, and shown in figure 11.

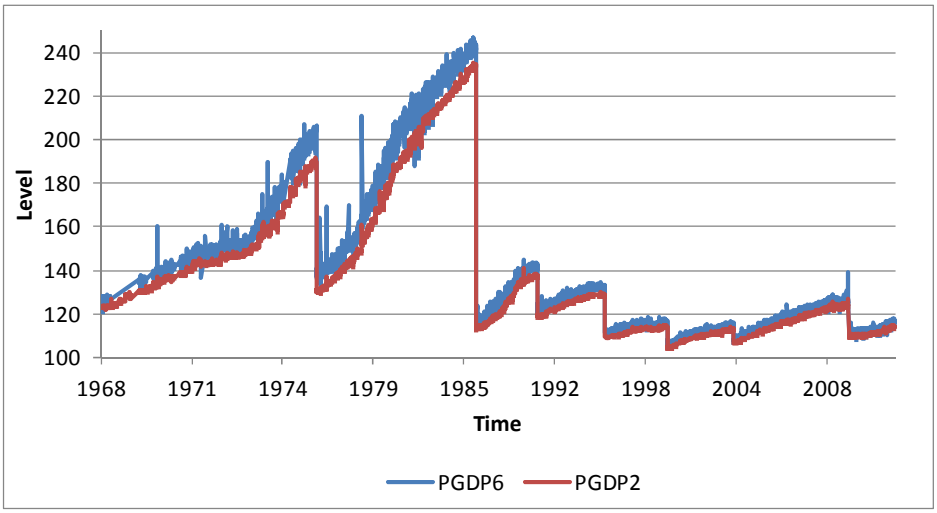
Table 8:Base year changes for NIPA variables, including the GDP deflator variable
Source: Federal Reserve Bank of Philadelphia (2011)

Range of Surveys	Base Year
1968:Q4 to 1975:Q4	1958
1976:Q1 to 1985:Q4	1972
1986:Q1 to 1991:Q4	1982
1992:Q1 to 1995:Q42	1987
1996:Q1 to 1999:Q33	1992
1999:Q4 to 2003:Q4	1996
2004:Q1 to 2009:Q2	2000
2009:Q3 to present	2005

Note: In the survey of 1992q1, the surveys measure of output swithces from GNP to GDP. In the survey of 1996q1, the surveys measure of NIPA prices and quantities switches to chain-weighted measures.

When working with percentage changes, the base year revisions do not have to be a problem, because the effect on the inflation rate is likely to be minor (Diebold *et al.* 1997; Clements 2006a). But, if we want to compare the quarterly levels of pgdp with the real data, a problem might occur. The survey may ask forecasters for predictions where the predicted horizon cross the time for the next base year change. One way to solve this could be to exclude all forecasts with horizons which extend beyond the date of systematic data revisions from the data (Keane and Runkle 1990). One could also use vintage data when comparing (Clements 2006b). Vintage data will always have the same base year as the forecasts, as vintage data are the data which were available at the time the forecast was made. It is also interesting to see if the forecasters managed to keep track of the base year changes in their forecasts. One way to inspect this is to plot the variable in question over time and see if there are any abnormalities. According to figure 11, which demonstrates the changing base year in the pgdp2 and pgdp6 variables, it seems reasonable to believe that the individuals were able to take into consideration the changes in base year when forecasting.

Figure 11: Base year changes for the pgdp2 and pgdp6 variables



4.3.5 Outliers and consistency of forecasts

A potential problem with most databases is biased results due to corrupt data. Thus, it is important that data used in the analysis are reliable and consistent. It is possible that errors from the survey questionnaires are transferred into the database (Giordani and Söderlind 2003). This section will discuss the possible event of extreme values and the consistency of the forecasts.

Extreme values ("outliers") can give biased results, which makes it important to locate them in order to assess their importance for our analysis. Figure 11 shows pgdp2 and pgdp6 over time, and gives a visual of any potentially problematic outliers. Pgdp2 seem to be relatively consistent, which makes sense since this is a "forecast" for the inflation value in the current quarter. Pgdp6, on the other hand, seems to have some potenital problematic outliers before the second base year change with the most serious being made in 1978q3. In the same quarter one can find similar outliers in pgdp4 and pgdp5 (not shown here), which could imply a forecaster who has made a mistake or made a forecast that deviates from the consensus. After some research, it seems as though these outliers are a result of a respondent giving an optimistic forecast (see table 9). This cannot be seen as an outlier beacause it seems as this was indeed the forecaster`s beliefs. Since this was the largest outlier, according to figure 11, it is reasonable to believe that other potential outliers are just an individual making a slightly more optimistic forecast than the consensus.

Table 9: "Outliers" from ID 47

Variable	pgdp1	pgdp2	pgdp3	pgdp4	pgdp5	pgdp6
Time	1978q2	1978q3	1978q4	1979q1	1979q2	1979q3
ID=47	150.7	160.9	172.5	185	197.7	211
Mean	150.7	153.7	156.8	160.1	163.1	165.8

If the forecasters are consistent, then the quarterly predicted pgdp levels should not be too different from the predicted annual levels (Smith and Yetman 2010). In the SPF database there seems to be a lack of consistency between the one-year-ahead forecasted inflation (pgdp6 when standing in quarter 1 or quarter 4, respectively) and the annual-average forecasts for the current year (pgdpa) or next year (pgdpb). This can be examined by plotting the abovementioned variables against each other. Figure 13 shows the first of these plots. It shows the difference between the forecasted inflation level one year ahead and the forecasted annual-average inflation

level for the current year, which should be close to zero. As we can see this is not the case, especially in the early 80`s, as the difference between the two variables varies from -3 percent to +10 percent. After the Federal Reserve Bank of Philadelphia took over the survey in the early 1990s, the problem almost ceases to exist. This could imply that there was better control over what was actually forecasted. The same comparison can be done for inflation forecasts for next year and the one-year-ahead forecast (i.e. pgdp6 if the forecaster is in the fourth quarter), and is shown in figure 12. The results are similar to what we have seen above, except for the fact that the errors are smaller. Nevertheless, we see a distinct improvement in the difference between the two variables over time. Another point worth mentioning is that the errors are more centered on zero than in figure 13.

Figure 12 and figure 13: Consistency in forecasts of one-year-ahead inflation and annual avareage for current year (left) and next year (right)

Figure 13:

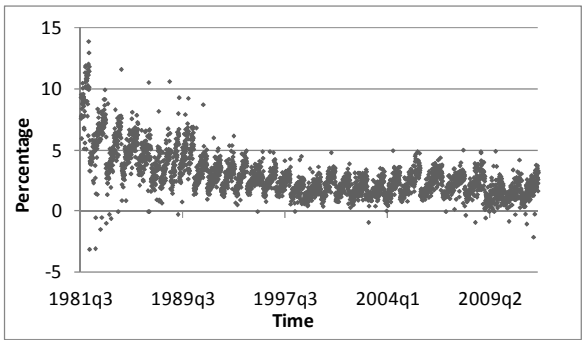
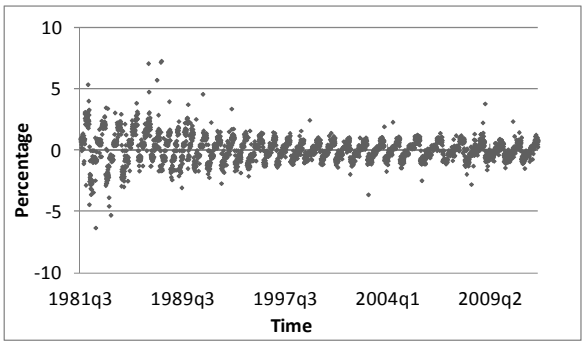


Figure 12:



One possible solution to this problem is to exclude values which are too extreme from the sample, e.g. all values that differs with more than a given percentage point (Smith and Yetman 2010). This would make the dataset more robust and less exposed to outliers. One problem, however, is that we can only perform this consistency check from when the survey started to include annual average forecasts in 1981q3. There is no way of checking the consistency for the earlier years, but according to the results in figure it is reasonable to suspect that they are not too good. Only removing values from the 80`s will not solve the problem which make this solution not desirable. Alternative solutions which seem more realistic are to use data from 1990 and onwards, or use sub-samples which start after the Federal Reserve Bank of Philadelphia took over the survey.

5. RESULTS

This section shows the results from the analysis of forecast accuracy among individual forecasters in the SPF. It is divided into several subsections which deal with different research questions. The purpose of dividing the analysis up in different parts is to illuminate and inspect the accuracy in diverse situations, and see how accuracy is affected.

5.1 How large are the forecast errors?

In order to get an overlook of the situation with regards to the forecast accuracy it can be useful to document how large and numerous the forecast errors really are. Table 10 gives an indication on the frequency and the size of the forecast errors for the individual forecasts. At first glance, it can be seen that the percentage of errors larger than the given levels increase as the forecast horizon becomes longer. This is a well documented and logical development, as longer horizons are harder to predict than shorter ones. One can also see the increased difficulty in forecasting longer horizons by looking at the accuracy statistics in the bottom section of the table. It is clear that the errors increase with the length of the horizon. Another result worth mentioning is the fact that the highest overestimations are much larger than the highest underestimations (see "range" variable). On all forecasting horizons, the largest overestimation is almost twice as large as the highest underestimation. This could have something to do with the sudden end of the high inflation period in the 1970s, where forecasters failed to regulate their forecasts in time.[16]

For the one-quarter-ahead forecast the errors do not seem to be very large, as expected (see column one). Almost ninety-six percent of all responses have an error below one percent, and only 0.4 percent has an error larger than two percent. When looking at the two-quarter-ahead forecast the frequency increase, especially for the "smaller" errors (column two). Now only eighty-one percent of all responses have an error below one percent, nineteen percent have an error above one percent and almost four percent of the responses have an error above two percent. The three-quarter-ahead forecasts have only sixty-three percent of responses under one

[16] See development of the GDP price index in section 3.3.1, p. 24.

percent error, while over ten percent miss the target with more than two percent (column three). The last forecast horizon, the one-year-ahead forecast, undoubtedly seems to be the hardest to predict of the horizons included (column four). Only fifty percent have an error less than one percent, while almost six percent of all responses have an error higher than four percent.

The MPE statistic in the bottom panel shows that the forecast error for the one-quarter-ahead horizon is fairly close to zero. This indicates that the under- and overestimations even out over time, making it an accurate forecast on average. For the other horizons, the MPE statistic is slightly more positive which indicates that forecasters on average tend to underestimate the actual change in inflation. It should be noted, however, that these numbers do not represent a large error. Nevertheless, based on this is could seem as forecasters are biased.

Table 10: Accuracy statistics for different forecast horizons

Horizon	1q	2q	3q	4q
Range	-5.6 to 3.0	-11.0 to 6.0	-16.4 to 8.1	-22.3 to 12.3
>1%	3.9 %	18.9 %	37.0 %	50.0 %
>2%	0.4 %	3.8 %	10.5 %	18.7 %
>3%	0.1 %	0.9 %	4.7 %	8.8 %
>4%	0.1 %	0.2 %	1.5 %	5.7 %
MPE	0.07	0.16	0.25	0.28
MAE	0.34	0.65	0.99	1.34
RMSE	0.48	0.90	1.39	1.90

Note: Range gives max and min error for each forecast. Numbers show how many percent of all responses which are above a certain absolute error, for 1-4 quarter horizon. MAE=mean absolute error, RMSE=root mean squared error, MPE=mean prediction error.

Going more into detail, table 11 illustrates the largest forecast errors for the different horizons, and the individual who gave the forecast. It documents two important points. First, all the largest forecast errors (top three for all horizons) are negative, i.e. the largest errors came from an overestimation of the actual inflation values. This result is somewhat contradicting when compared to table 10 where it was shown that forecasters on average underestimate inflation. This means that the number of positive forecast errors have to be larger than the number of negative errors in order for the mean error to be positive. Second, all the largest errors are almost exclusively from the mid 1970s and early 1980s. This leads to a conclusion that these

time periods were harder to predict, which is supported by the standard deviation in figure 2. It is no wonder why forecasters were given a bad reputation during this period (Croushore 2006).

Many of the individuals listed in the table recur several times in the table, across horizons. ID 117, for example, has eight entries in the table, while ID 125 has seven entries. These forecasters apparently had big problems dealing with the high inflation period. Thinking it would last longer than it did, they failed to moderate their forecasts for the upcoming inflation forecasts. Of the 40 entries in the table (top ten*four horizons) there are only 14 different forecasters which mean some optimistic forecasters stand for most of the largest errors.

Table 11: Largest forecast error for different forecast horizons

	Panel A: 1q horizon			Panel B: 2q horizon		
Number	ID	Time	Error	ID	Time	Error
1	117	1976q1	-5.64	47	1978q3	-11.04
2	41	1981q4	-5.11	117	1976q1	-9.45
3	47	1978q3	-5.10	117	1976q2	-7.01
4	13	1981q4	-4.04	127	1974q3	5.97
5	100	1987q3	-3.40	41	1981q4	-5.38
6	117	1976q2	-3.23	9	1982q1	5.19
7	127	1974q3	2.99	20	1968q4	4.78
8	8	1979q4	2.84	125	1977q3	-4.59
9	62	1982q3	-2.75	13	1981q4	-4.58
10	125	1977q3	-2.71	9	1982q2	4.54
	Panel C: 3q horizon			Panel D: 4q horizon		
	ID	Time	Error	ID	Time	Error
1	47	1978q3	-16.37	47	1978q3	-22.33
2	117	1976q1	-15.45	117	1976q1	-18.95
3	117	1976q2	-11.98	117	1976q2	-17.04
4	127	1974q3	8.13	125	1974q1	12.32
5	9	1982q1	7.32	125	1977q3	-9.89
6	148	1971q3	7.20	7	1974q1	9.22
7	20	1968q4	7.06	148	1973q4	9.19
8	125	1977q3	-6.82	9	1982q1	9.18
9	125	1974q1	6.72	125	1975q1	9.05
10	148	1973q4	6.55	22	1978q3	9.00

Note: The table shows the ten largest forecasting errors and the quarter the forecast was made. One panel for each forecasting horizon.

5.2 Accuracy for the longest individual forecasting series

An important objective for forecasters is to generate accurate forecasts over time. One way to investigate if the individuals in the SPF database possess this quality is to measure the accuracy for the most continuous individuals, i.e. those with the longest consecutive forecasting series which I presented earlier (see *table 1*, p.22). Were these individuals able to accurately predict inflation over time? Did they manage to outperform the benchmark model? The result from this analysis is shown below in table 12 for the one-quarter-ahead horizon.

Panel A illustrates how the different individuals performed when all of their responses are included (not only data for their longest consecutive series), and how their performance was compared to the random walk model. At first sight one can see that the errors are fairly similar across individuals, and that the number of individuals who over- or underestimate real change in inflation are about the same. The best performer, based on the MAE and RMSE statistics, is ID 510 with a MAE equal to 0.17 and RMSE equal to 0.22, while ID 483 comes close behind with MAE equal to 0.18 and RMSE equal to 0.23. A more striking point is that these two forecasters are the only ones who managed to beat the benchmark model, according to Theil's U-statistic. At the other end one can find ID 125, who had almost twice as high RMSE as the random walk model. Included in the table is also the coefficient from the forecast comparison regression, and the corresponding p-value for the null of β equal to zero. With a five percent significance level, β is significantly different from zero for all individuals but two. It thus seems as the individual forecasters included in this table indeed have something to add to the forecast of the benchmark model, even though they have worse RMSE`s.

Panel B shows the same statistics as panel A, but now only data comprised by the time span of the longest series are included. In addition, a column with RMSE/SD (RMSE normalized with the standard deviation) is included which makes it possible to take into consideration that some time periods are harder to predict than others. As panel B shows, ID 70 performed best in his/hers longest series given the difficulty in that period. His/hers performance was also approximately equal to the forecasts from the random walk model (Theil's U is just above unity). Again, only ID 510 and ID 483 managed to outperform the benchmark model even though they were not the top performers according to column five. This leads to a preliminary

conclusion that a simple random walk model is a good alternative forecast method for predicting the one-quarter-ahead change in inflation. It also seems as it was extremely hard to predict the change in inflation during the high inflation period, as forecasters perform badly in that period even though the difficulty has been taken into consideration. Statistically, the null hypothesis is rejected in all instances except two with a five percent level and all instances except five with a one percent significance level. According to these results, ID 30 and ID 48 do not statistically contribute to the forecast given by the benchmark. All other individuals have something to add to the forecast given by the random walk model. In total, it therefore seems as these individuals who are included in the table in fact are able to predict inflation accurately over time.

Table 12: Accuracy statistics for the most persistent individuals

					Panel A: Entire dataset						Panel B: Longest series						
#	Id	Length	Start	Stop	MAE	MPE	RMSE	T-U	β	P	MAE	MPE	RMSE	T-U	RMSE/SD	β	P
1	65	49	1990q2	2002q2	0.28	0.10	0.39	1.35	0.23	0.021	0.15	-0.01	0.18	1.11	1.02	0.43	0.000
2	510	40	2002q1	2011q4	0.17	0.04	0.22	0.85	0.72	0.000	0.17	0.03	0.22	0.81	0.81	0.77	0.000
3	70	36	1980q4	1989q3	0.34	0.05	0.43	1.13	0.38	0.006	0.27	-0.17	0.31	1.09	0.68	0.40	0.024
4	30	36	1981q3	1990q2	0.38	-0.11	0.46	1.84	0.11	0.135	0.41	-0.14	0.48	1.71	1.58	0.13	0.189
5	60	34	1985q2	1993q3	0.32	-0.12	0.39	1.16	0.34	0.005	0.26	-0.20	0.31	1.44	1.36	0.22	0.038
6	433	34	1991q1	1999q2	0.19	-0.02	0.24	1.26	0.34	0.000	0.17	-0.15	0.21	1.45	1.50	0.19	0.001
7	48	33	1968q4	1976q4	0.58	0.46	0.68	1.63	0.12	0.337	0.58	0.46	0.68	1.63	1.21	0.12	0.337
8	483	32	2004q1	2011q4	0.18	0.04	0.23	0.91	0.61	0.000	0.21	0.05	0.25	0.89	0.85	0.64	0.000
9	446	32	2004q1	2011q4	0.21	0.03	0.25	1.12	0.40	0.001	0.27	0.15	0.32	1.13	1.08	0.40	0.011
10	125	30	1974q1	1981q2	0.60	-0.21	0.85	1.89	0.19	0.001	0.67	-0.30	0.93	2.09	1.84	0.06	0.006

Note: Table shows accuracy statistics for the longest consecutive series of response to the survey,for the one-quarter-ahead forecast. MAE=mean absolute error, MPE=mean prediction error, RMSE=root mean squared error, T-U=Theils U-statistic, SD=standard deviation for the respective time periods the forecast was made, thus representing the forecasting difficulty that period. Beta comes from the forecast comparison regression and P denotes the P-value for the null of β=0.

5.3 Have forecast accuracy improved over time?

A desired attribute for forecasters is the ability to learn from their previous mistakes, and correct or it in future forecasts. If individuals do have such a characteristic one could expect the accuracy to improve over time, especially taking into consideration that the forecasting techniques and tools available have improved (Zarnowitz and Braun 1993). This section will try to examine if this is true or not.

Table 13 illustrates how forecast accuracy has developed over time, here represented by decades. The different panels correspond to different forecast horizons. Based on the RMSE/SD variable one can compare accuracy statistics for different time periods and consider different levels of difficulty. The MPE statistic for the different decades suggests that the 1980s and 1990s were overestimated for all horizons, and increasingly so for the longer horizons. In the 1970s and 2000s, on the other hand, inflation was underestimated for all forecast horizons. What is more, the errors seem to converge towards zero which suggests that accuracy has improved. For the one quarter horizon the 1990s have been the most accurate decade, based on the MAE and RMSE statistics. For the three longest horizons, on the other hand, accuracy has improved over time resulting in the 2000s being the most accurate decade. Alternatively, if the RMSE`s are normalized with the standard deviation for the respective time periods the 1980s become the most accurate decade for the two shortest horizons, while the 2000s are most accurate for the two longest horizons. The 1990s brought along the worst forecasting performance in all decades, given the difficulty in that period. Thus, even though the accuracy statistics have improved over time it seems as forecasters should have performed better given the lower difficulty in the later decades.

The most striking point one can draw from this table is how bad the performance have been compared to the benchmark model. For the one quarter horizon, Theil's U-statistic is fairly close to unity which indicates an equal performance of the forecasters and the random walk model. For all other periods and horizons the benchmark model performed better, and in the worst period the RMSE for the benchmark model was over four times lower than the mean forecast. This result is also supported by the forecast comparison regression test. For the shortest horizon the null of β equal to zero is rejected for all decades with a five percent significance level, suggesting that the survey forecasts have something to add to the forecast given by the benchmark model. For the other horizons, however, the picture is less promising. None of the β`s are significantly different from zero, implying that the mean forecast have nothing to add to the forecast from the benchmark model.

Going further, one can break down the average forecasts in each decade to show the best forecasters, and how they performed compared to the benchmark model. Table 14 shows this kind of analysis. According to panel A in the table, ID 145 performed best in the 1970s with ID

21 and ID 14 close behind. They also outperformed the benchmark model as most forecasters did, except in the 1990s where only one individual slightly outperformed the naïve model. In total (panel E) it seems as the random walk performs better than roughly half of the ten best forecasters, based on Theil's U. It also seems as forecasters have improved their forecasting over time based on a diminishing RMSE statistic.

Table 13: Accuracy statistics for different decades

Years	MPE	MAE	RMSE	SD	T-U	β	P	Years	MPE	MAE	RMSE	SD	T-U	β	P
Panel A: 1q								Panel C: 3q							
1970`s	0.26	0.38	0.48	0.87	1.06	0.43	0.046	1970`s	1.03	1.14	1.54	1.04	2.55	0.00	0.976
1980`s	-0.16	0.26	0.31	0.54	1.06	0.41	0.001	1980`s	-0.64	0.87	1.00	0.62	2.53	-0.05	0.463
1990`s	-0.13	0.16	0.19	0.92	1.21	0.32	0.004	1990`s	-0.50	0.55	0.60	1.20	3.37	-0.04	0.325
2000`s	0.07	0.20	0.24	0.91	0.91	0.60	0.000	2000`s	0.19	0.49	0.61	0.94	2.05	0.08	0.282
Panel B: 2q								Panel D: 4q							
1970`s	0.60	0.73	0.97	0.95	1.77	0.09	0.506	1970`s	1.55	1.67	2.26	1.16	3.49	-0.08	0.315
1980`s	-0.37	0.55	0.64	0.57	1.76	0.03	0.636	1980`s	-0.98	1.20	1.40	0.69	3.26	-0.07	0.308
1990`s	-0.30	0.34	0.38	1.04	2.13	0.06	0.315	1990`s	-0.71	0.79	0.85	1.35	4.43	-0.06	0.075
2000`s	0.14	0.34	0.42	0.92	1.51	0.21	0.076	2000`s	0.24	0.65	0.82	0.99	2.56	0.01	0.812

Note: Table shows average statistic per decade, and each panel shows a different forecast horizon. P shows p-value for the null of β=0. See table 12 for additional notes.

Those individuals who performed best in total (panel E), have made their forecasts exclusively in the 1990s and 2000s. This could indicate several things: it was easy decades to forecast inflation in, forecasting techniques have improved (better computers, newer theories etc.) or that forecasters have improved their skills by learning from previous mistakes. Which one of these reasons that have made the largest impact on the accuracy, or if other reasons were present, is hard to test. That being said, according to the results in this subsection it is reasonably to believe that the forecast accuracy in fact has improved and most likely it is a result of a combination of all the above-mentioned reasons.

An interesting point revealed in table 14 is that very few individuals recur in multiple decades. The most obvious explanation is that few participants have responded to the survey for so many years. Nevertheless, there are some individuals who have performed well in two consecutive decades. ID 94 was ranked sixth in the 1980s and first in the 1990s but did not make it into the list containing all responses. Another explanation for none recurring individuals could be that studies have shown that there are no forecasters who innately outperform the others (D'Agostino

et al. 2010). For that reason, it is not likely anyone will recur as top performers in several decades.

Table 14: Top ten forecasters per decade

	Panel A: 1970`s			Panel B: 1980`s			Panel C: 1990`s			Panel D: 2000`s			Panel E: All		
#	ID	RMSE	T-U	ID	RMSE	T-U	ID	RMSE	T-U	ID	RMSE	T-U	ID	RMSE	T-U
1	145	0.357	0.73	80	0.249	0.87	94	0.146	1.26	500	0.172	0.88	440	0.167	1.63
2	21	0.376	0.79	44	0.273	0.78	429	0.150	1.02	502	0.183	1.00	414	0.178	1.16
3	14	0.380	0.83	99	0.289	0.99	440	0.167	1.63	439	0.194	0.98	416	0.181	1.28
4	133	0.426	1.02	51	0.296	1.04	431	0.169	0.99	510	0.214	0.80	502	0.189	0.95
5	84	0.427	0.92	94	0.298	1.06	411	0.176	1.18	498	0.219	0.89	500	0.195	0.90
6	7	0.429	0.99	67	0.300	0.94	414	0.180	1.22	407	0.227	0.82	409	0.205	1.68
7	89	0.431	0.91	5	0.304	1.01	416	0.181	1.28	483	0.235	0.87	510	0.215	1.11
8	138	0.443	1.07	70	0.314	1.11	65	0.184	1.13	548	0.247	0.75	429	0.219	0.83
9	144	0.459	1.04	84	0.330	1.04	446	0.191	1.22	456	0.248	1.14	404	0.223	1.48
10	66	0.465	0.94	15	0.332	1.31	424	0.193	1.25	508	0.249	1.13	483	0.226	0.90

Note: The table shows the ten best forecasters in each decade based on RMSE for the one-quarter-ahead horizon. Individuals with less than ten responses in the respective decades have been removed in order to secure against a lucky guess and remove those whose series stop early in a decade. See table 12 for additional notes.

Nevertheless, it would be interesting to examine this question further. It is possible that some individuals are top performers in multiple quarters (in contrast to decades discussed above). If it turns out that some individuals are systematically outperforming the others in several quarters it could be an important finding. By learning what characterizes a top performer, researchers can gain valuable insight into well-functioning techniques and theories. If these findings could be generalized, others may learn from the findings and consequently people would get more correct inflation expectations.

Table 15 below illustrates this question. Every time an individual has given one of the best forecasts (top five) in a quarter, he/she is given one point. This can then be compared to the number of quarters the individual has participated in the survey (column four).The first noticeable point from table 15 is the low number of quarters the individuals in the table have been top performers compared to their total number of responses. It does not seem as any of the individuals have managed to consistently outperform the others. That being said, there are one individual who separates somewhat from the rest. ID 20 is listed in all panels in the table which indicates that this person was one of the best performers for all horizons. What is more, in the

four quarter horizon ID 20 have five more top rankings than next person on the list. Compared to the total number of responses it cannot be said that this is impressive, but it is still a signal that this individual performed better than the rest. The unanswered question is if this is enough to get some general insight in what characterizes a top performer.

This finding is in line with Batchelor (1990) who showed that there are no consistency in individuals' characteristics in accuracy rankings that can be used to pick the best individual forecasters. Zarnowitz (1984) also concluded that no single forecaster has been able to earn a consistent superior accuracy, as others have done before him (Zarnowitz 1967; McNees 1979, cited in (Zarnowitz 1984)).

Table 15: Individuals with the highest number of top five rankings

#	ID	Top	RMSE	Total	#	ID	Top	RMSE	Total
	Panel A: 1q					Panel C: 3q			
1	433	7	0.24	68	1	446	9	0.69	70
2	407	6	0.27	69	2	65	8	1.06	123
3	84	6	0.30	122	3	20	7	1.25	88
	Panel B: 2q					Panel D: 4q			
1	428	7	0.50	75	1	20	13	1.55	88
2	65	6	0.69	123	2	431	8	0.84	64
3	431	5	0.43	64	3	420	8	0.88	69

Note: The table shows individuals with the highest number of forecasts which wereranked best in a quarter and total responses to the survey. Panel A-D shows 1-4 quarter horizon, respectively. Panel A: ID 20 and 60 also have six top rankings but higher rmse. Panel B: ID 20, 72, 94 and 472 also have five top rankings but higher rmse. Panel D: ID 15, 30 and 411 also have eight top rankings but higher rmse.

5.4 Accuracy during recessions

According to the National Bureau of Economic Research (2010, p. 1), a recession is defined as "a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment, industrial production, and wholesale-retail sales". Such economic downturns produce uncertainty and increased volatility in macroeconomic variables and therefore it also represents one of the more difficult times to predict inflation. How will the central bank react? How will it affect consumption and saving? The outcome of inflation will depend on the answer of such questions, thus making it very

difficult to predict the actual inflation change. Previous findings seem to suggest that forecasters are not able to predict turning points, and most forecasters failed to predict the recessions in the early 1990s (Batchelor 2000). This section will examine how accurately individuals managed to predict inflation during recessions in the U.S.

The first panel in table 16 shows average statistics for the different recessions. Several of these results are worth mentioning. First, forecasters have been more accurate during the most recent recessions than the older ones, both when difficulty is included (fourth column) and not. In fact, if one looks at the RMSE statistic it seems to be a clear distinction between the RMSE before and after the fourth recession in the early 1980s. After this recession, forecasters seem to be better at predicting inflation during recessions. It could mean the later recessions have affected inflation in a less prominent way, making forecasting easier. On the other hand, it could also suggest that forecasters have learned from previous recessions and have become better at predicting inflation in difficult times. Second, the worst recession, in terms of forecast accuracy, was the second recession during the high inflation period in the mid 1970s. This conclusion holds both when looking at the RMSE and the MAE statistic. If the standard deviation is included, however, forecasters were least accurate in the first and third recessions. Third, forecasters underestimated inflation during the first three recessions but have overestimated inflation in the last four recessions. It seems as though newer recessions have brought lower inflation than what most forecasters predicted. One natural reason for this is the Federal Reserve, who has kept a much tighter leash on inflation after it was brought under control in the mid 1980s. Thus, it has also become easier to predict the outcome of GDP price index as it has been less volatile in the more recent recessions (see figure 2, p.21).

Panel B goes more into detail, and shows the best individual in each recession based on the RMSE statistic, and how they performed compared to the simple random walk model. As can be seen, there are no individuals who were top performer in several recessions. It also seems, as pointed out above, that forecasters have improved their predicting skills because both the RMSE and RMSE/SD variables are declining over time. When comparing panel B to table 14, which shows top performers by decade, an interesting point emerges: none of the ten best performers in each decade are the best performer in a recession. This may be due to several reasons. First, most forecasters did not complete the survey during all quarters the recession lasted which

means they are not included in the evaluation in panel B. Second, the best performers in table 14 may be biased if they did not predict inflation in "hard" quarters (i.e. quarters with a high standard deviation). This would lead to higher accuracy statistics as harder quarters often bring along higher forecast errors. Third, being a good forecaster involves predicting correct values of inflation but it also implies being good at predicting turning points in inflation, e.g. changes before and after a recession. Take ID 117 as an example. He/she was the best performer during the recession in the early 1970s (no. 2) but is not to be found on the list of top ten performers during the decade. The reason for this could be found in table 11. ID 117 has the highest forecast error ever given for the one-quarter-ahead forecast horizon which means that he/she totally failed in his/hers prediction when the high inflation growth ended. Thus, many of the best performers during the decade may have been better at predicting the turning point of the recession which kept their total performance close to the top.

Table 16: Forecast accuracy during contractions for the one-quarter-ahead horizon[17]

	Panel A: Average statistics				Panel B: Best individual					
Time of recession	MAE	MPE	RMSE	RMSE/SD	ID	MAE	MPE	RMSE	RMSE/SD	T-U
1: 1969q4 - 1970q4	0.49	0.38	0.56	2.53	86	0.35	0.22	0.41	1.86	1.22
2: 1973q4 - 1975q1	0.77	0.58	0.90	1.80	117	0.56	0.46	0.68	1.36	1.39
3: 1980q1 - 1980q3	0.40	0.26	0.61	2.51	98	0.19	0.19	0.23	0.95	0.90
4: 1981q3 - 1982q4	0.48	-0.27	0.72	2.18	70	0.39	-0.39	0.45	1.36	1.71
5: 1990q3 - 1991q1	0.24	-0.16	0.32	2.06	62	0.14	-0.14	0.17	1.10	0.65
6: 2001q1 - 2001q4	0.17	-0.06	0.22	1.28	428	0.11	-0.05	0.12	0.67	0.66
7: 2007q4 - 2009q2	0.26	-0.12	0.35	1.20	472	0.14	-0.13	0.19	0.65	0.59

Note: This table shows accuracy statistics for the one-quarter-ahead forecast horizon for different recessionary periods in the US economy, as defined ny the National Bureau of Economic Research (NBER). Average statistics (panel A) are calculated across the time period for the recession, the same goes for the standard deviation. The individual must have participated in all surveys which are included by the time period given by the NBER in order to be considered in panel B. See table 12 for additional notes.

[17] Due to the short time span of the recessions there are insufficient observations to do a forecast comparison regression test. Recession four (1981q3-1982q4) had only one individual who responded to all surveys throughout the recession. All the others had three or more to individuals to rank.

5.5 Which industry contains the best forecasters?

Even though previous literature has studied individual forecasts, there is to my knowledge none who have used the industry variable in their research. One obvious reason for this is that there have been no studies on individual data after the industry variable was introduced in 1990q2. Thus, an analysis of the accuracy for the different industries is very interesting and of current interest.

The Federal Reserve Bank of Philadelphia allocates an industry number based on the type of firm the individual currently works at. Industry one contains people working in a firm characterized as a financial provider (investment banking, hedge and mutual funds, asset management etc.), industry two contains non-financial providers (universities, forecasting firms, research firms etc) while industry three contains all people where they do not know the affiliation. Having the ability to predict the inflation correctly would be of great importance to all of these firms, as forecasting is a large part of their daily jobs.

Table 17 shows the forecast accuracy for the different industries for four different forecast horizons. Some of the results from this table are worth presenting. First, there seems to be slight differences in accuracy between the different industries. Based on the MPE statistic, industry 1 and 2 outperform industry 3 with about 1.5 percentage points. However, if we base performance on MAE and RMSE statistics, they are almost identical in performance and this goes for all forecast horizons. It seems as though none of the industries have an advantage over others when it comes to forecasting inflation. Second, a more striking result is that all industries across all forecast horizons seem to have overestimated inflation in their predictions. This finding is somewhat contradictory to the earlier findings where it was suggested a tendency towards underestimation (see table 10 and table 18). The explanation for this is simple and can be found in table 13. On average, forecasters have underestimated inflation in the 1970s and 1980s, while they have overestimated inflation in the 1990s and 2000s. Thus, since I only use data from 1990q2 in this analysis it will result in an overestimation from the industries. This also shows what happens to the forecasting accuracy when the turbulent 1970s and 1980s are removed from the sample.

With this in mind a new question arises: are the industries biased? After performing a bias test it seems as the answer depends on the forecast horizon (see table 17). For the two shortest horizons the null hypothesis of α equal to zero cannot be rejected, thus suggesting unbiasedness. For the two longest horizons, however, the results are opposite. On a five percent significance level all industries are biased. On the other hand, if we use a one percent significance level it is only industry one and two for the longest horizon which are biased.

Table 17: Forecast accuracy for different industries

Industry	MPE	MAE	RMSE	α	P	Industry	MPE	MAE	RMSE	α	P
Panel A: 1q						Panel C: 3q					
1	-0.007	0.226	0.290	-0.007	0.504	1	-0.070	0.611	0.751	-0.070	0.041
2	-0.009	0.221	0.278	-0.009	0.334	2	-0.062	0.597	0.740	-0.064	0.037
3	-0.027	0.222	0.302	-0.027	0.265	3	-0.175	0.574	0.751	-0.175	0.030
Panel B: 2q						Panel D: 4q					
1	-0.026	0.419	0.519	-0.027	0.233	1	-0.133	0.810	0.991	-0.141	0.003
2	-0.028	0.403	0.501	-0.028	0.147	2	-0.113	0.808	0.995	-0.118	0.006
3	-0.093	0.392	0.508	-0.093	0.070	3	-0.229	0.785	0.994	-0.234	0.041

Note: The table shows accuracy statistics for different industries, on average across time. The panels shows different forecasting horizons. Data from 1992q2. α=bias, p=p-value for null hypothesis α=0. See notes to table 12.

5.6 Overall performance

This section will provide a discussion concerning the overall performance of the individual forecasters. More specifically, it will examine the predictive accuracy and biasedness of individuals, and how forecasters perform compared to the random walk model. The analysis will be divided into two sub-samples; one containing all observations and one containing data from 1990q2 when the Federal Reserve Bank of Philadelphia took over the survey.

Table 10 indicated, based on the MPE statistic, that forecasters on average underestimate inflation. If this is true, it is a problematic result because if forecasters are systematically underestimating the forecasted variable it means that forecasters do not learn from their previous mistakes and therefore are biased. Table 18 examines this in more detail, and shows the percentage of responses that fall over or under the true change in real inflation for each of the four different forecast horizons (see panel A). It confirms the earlier findings, because individuals seem to persistently underestimate real change in inflation. On all horizons, more

than 50 percent of all responses are below the actual outcome. This seems to indicate some sort of biasedness among the individual forecasters. However, when testing this statistically the results are not that shocking. According to the results in the table below, roughly two-thirds of the individuals have an α significantly different from zero which mean they are biased. As the horizon increases, so does the percentage of biased forecasters. The increase cannot be said to be alarming, however, as it only rises slightly more than two percentage points from the one quarter horizon to the four quarter horizon. The same table also gives a nice summary with regards to the forecast performance of the individuals compared to the benchmark model. It shows the proportion of responses which perform better or worse than the benchmark. The results are a bit striking, as they indicate that the majority of responses from the survey are worse than the simple benchmark. For the four quarter horizon, for example, 86 percent of the responses were worse than the benchmark. Note, however, that the table says nothing about the difference between the forecasts. So if the individual forecast is 0.001 percent worse than the benchmark it is still considered to be worse, even though the difference is not significant. Therefore, I also conducted a forecast comparison regression for all individuals to see how many percent of the individuals who in fact have no information to add compared to the random walk model. For the one quarter horizon, almost 50 percent of the individuals have a comparison regression coefficient which is significantly different from zero implying they have something to add to the forecast from the benchmark model. For the longer horizons it looks worse, as only 12.5 percent, 9.1 percent and 16.6 percent add information to the forecast from the benchmark model in the two-four quarter horizons. This basically means that the majority of individuals fail to outperform a simple random walk model.

Table 18 also shows the same statistics when only data from 1990q2 and onwards are being analyzed (see panel B). This will illustrate how the forecasting performance for the individual forecasters is when the high inflation period and most of the problems discussed in section 4.3 are removed. The situation seems to be a bit better. According to the table, most forecasters are now overestimating inflation in contrast to panel A where underestimations were more frequent. Nevertheless, it still suggests biased forecasts and therefore the same bias test as above was conducted. For this sample, the test shows fewer individuals being biased for all horizons than for the sample including all horizons. This leads to an assumption that forecasters predicting in the last two decades are more accurate than forecasters predicting in the earlier years. It could

also suggest that forecasters thus have become more accurate over time. The bottom section of panel B illustrates how the SPF forecasters' performance was compared to the random walk model. Here the results are not much better than before, emphasizing the good performance of the random walk model. Based on percentages, the numbers are almost exactly the same as one could see when the whole dataset was included. However, when the forecast comparison regression is conducted the results are better than before for the two shortest horizons. Now over 67 percent of forecasters in the one quarter horizon, and 21 percent of forecasters in the two quarter horizon have something to add to the forecast by the random walk model. For the three quarter horizon, on the other hand, the situation is similar to the previous findings, while the four quarter horizon actually has fewer individuals who can beat the random walk.

In total it seems as forecasters have improved over time as tests show that there are fewer biased individuals and more individuals manage to beat the random walk model. Still, the majority fail to give forecasts that contain more information than the forecast from the random walk model for the three longest horizons.

Table 18: Overall performance for two different sub-samples

	Panel A: All observations				Panel B: Observations from 1990q2			
	1q	2q	3q	4q	1q	2q	3q	4q
Overestimation	46.8 %	46.5 %	45.5 %	47.3 %	53.3 %	54.2 %	54.8 %	55.8 %
Underestimation	53.2 %	53.5 %	54.5 %	52.7 %	46.7 %	45.8 %	45.2 %	44.2 %
Biased (α=0)*	33.8 %	35.1 %	35.1 %	36.1 %	25.1 %	28.1 %	30.5 %	30.3 %
Worse than RW	59.9 %	74.5 %	81.2 %	85.9 %	59.8 %	74.1 %	80.7 %	84.4 %
Better than RW	40.1 %	25.5 %	18.8 %	14.1 %	40.2 %	25.9 %	19.3 %	15.6 %
β=0*, p<0.05	49.4 %	12.5 %	9.1 %	16.6 %	67.1 %	21.2 %	9.4 %	8.3 %

Note: Top three rows show the proportion of responses which over- or underestimate real inflation, and the proportion of individuals who are biased. Bottom three rows show the proportion of responses which predict better or worse than the RW model, and the proportion of individuals where the null ofβ equal to zero is rejected on a 5% level
*One individual dropped due to insufficient data for the 4q horizon.

5.7 Comments

Given the information in section 4.2 (p. 27) concerning the timing of the survey, it seems reasonable that I have chosen a benchmark model that use information available in the previous quarter. However, due to the implications with revision in the GDP variable, as mentioned

before, it could have given the benchmark model a too big advantage which in turn biased my results. I have therefore repeated the analysis in the lower part of table 18 with the random walk model lagged one more quarter to see if this affected the results. According to table 19, it did not remarkably change the results. The majority of the forecasts are still worse than the random walk model on all horizons. The results from the forecast comparison regression test are somewhat better for the two shortest horizons, but equal in the two longer horizons.

I also could have included different time series models as comparison, but after seeing the results from the simple random walk model I chose not to. According to Ang *et al.* (2007) an ARMA(1,1) model is one of the best performers among time series models when predicting CPI inflation. This model, which is not very sophisticated either, performs slightly better than the random walk model for my data as well, and therefore I see no point in repeating all the analysis when the results are known in advance. Based on the data and techniques I have used it is apparent that time series models are capable of predicting GDP price index inflation quite well.

Table 19: Performance relative to benchmark after lagging

	1q	2q	3q	4q
Worse than RW	55.6 %	65.5 %	72.9 %	77.8 %
Better than RW	44.4 %	34.5 %	27.1 %	22.2 %
β=0*, p<0.05	54.5 %	21.6 %	9.7 %	16.7 %

Note: Performance relative to RW model after lagging values one more qurter. See table 18.

6. Conclusion

The purpose of this thesis has been two-folded. First, as a part of a research project I cooperated with a co-student in an attempt to give an overview of potential and current problems with the SPF database. It was our intention to document these problems and come up with possible solutions based on our own suggestions or based on earlier findings in the literature concerning the database. Second, I sought to given an assessment of the forecast accuracy of individual forecasters based on data from the SPF. Most previous and up to date studies concerning survey data have used pooled data in form of a mean or median to give an assessment. In that respect, this thesis is a supplement to the existing literature.

We found five problems with the database which are worth mentioning. First, it has been noted in previous literature working with individual data that forecasters respond infrequently to the survey. A well adopted solution is to drop all individuals with fewer than twelve responses. Second, there is a problem with missing values in the one-year-ahead forecast. Due to incomplete survey questionnaires by some respondents there are five quarters which have no value for this forecast horizon. A solution is to start with data from 1972q3, as all quarters with missing values are located in the beginning of the survey. A second solution is to fill in values, e.g. based on a linear projection. A third problem is reallocation of used ID's. Due to lack of control in the early periods of the survey there may be a problem with several individuals forecasting under the same ID number. This will make impossible to separate individuals from one another. It is guaranteed, however, that this is only a potential problem with date from before the Federal Reserve Bank of Philadelphia took over the survey in 1990q2. Fourth, it should be noted that several variables have multiple changes in base year. This will only pose a problem if level data are being used, but can be ignored if changes in these variables are being used. Fifth, there seem to be a problem concerning the consistency in the forecasted values. The one-year-ahead inflation often deviates from the forecast of the annual growth next year, for instance. A solution adopted in the literature is to chop of those observations which seem to be inconsistent. A new problem arises, however, due to the fact that the annual average forecast was introduced in 1980q1 and we have no way of checking how inconsistent forecasters were before this date.

With regards to the forecast accuracy of individual forecasts there are several findings. First, some individuals are able to accurately predict inflation over time and it also seems as they become more accurate over time. This finding is based on analysis of the longest individual forecasting series in the database and forecast accuracy across decades. None of the individuals seem to systematically outperform the other forecasters in the database. Why forecasters have improved over time is hard to test based on this data, so I can only conjecture. It may be that forecasters have improved their skills by learning from previous mistakes and enhancing their knowledge. However, it could also be a result of better technology and a less volatile underlying variable. Second, forecast accuracy during recessions have been worse than the average accuracy during the respective decades suggesting a deteriorating performance during difficult times. Nevertheless, some individuals performed well during these difficult times and outperformed the random walk model. It also seems as the accuracy have improved from the earliest recessions to the newer ones. This underlines the above-mentioned finding that forecasts are getting more accurate over time. No newer papers have documented this, but Zarnowitz and Braun (1993) found no improvement in their dataset up until 1990. Third, I find no difference between industries in terms of forecasting accuracy for any horizon. I did, however, find evidence suggesting biasedness in the two longest horizons which means that they are systematically overestimating inflation. This is, as far as I know, a new finding and thus is a valuable contribution to the existing literature. Fourth, I also find evidence for bias among some individuals. The majority, about two-thirds, seems to be unbiased. This result is similar to Zarnowitz (1985), and somewhat contradicting to Keane and Runkle (1990) who find no evidence of bias. Fifth, most of the individuals fail to add information to the forecast given by the same change random walk model. This applies especially for the three longest horizons. This is somewhat contradicting to previous findings. Ang *et al.* (2007), for example, find that surveys outperform most time series models in forecasting CPI inflation. It should be noted, however, that researchers have found that a random walk model can perform well in forecasting inflation (Ang *et al.* 2007; Atkeson and Ohanian 2001).

Given the importance for most economic actors of having correct inflation forecasts it is imperative to continue the research on this phenomenon and questions related to the issue. Based on the results in this paper, it would be interesting to see a study on individual forecasts of the GDP price index when vintage data are used (in contrast to revised data, which I used).

This might explain some of the difference in forecasting accuracy between individuals and the random walk model. Further studies could also address the importance of having correct expectations. If it turns out forecasters are biased, how will this affect policy decisions? Another important question in this regard is how expectations from surveys affect the overall inflation expectation in the society. If people use surveys as a source of information, which consequences would it have if survey expectations is not satisfactory?

Bibliography

Akhtar, M., Los, C. & Stoddar, R. (1983) Surveys of Inflation Expectations: Forward or Backward Looking? *Quarterly Review of Federal Reserve Bank of New York,* 8**,** 63-66.

Ang, A., Bekaert, G. & Wei, M. (2007) Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better? *Journal of Monetary Economics,* 54**,** 1163-1212.

Atkeson, A. & Ohanian, L. E. (2001) Are Phillips Curves Useful for Forecasting Inflation? *Federal Reserve Bank of Minneapolis Quarterly Review,* 25**,** 2-11.

Batchelor, R. A. (1990) All Forecasters Are Equal. *Journal of Business & Economic Statistics,* 8**,** 143-144.

Batchelor, R. A. (2000) The IMF and OECD Versus Consensus Forecasts. *City University Business School, London*.

Batchelor, R. A. & Dua, P. (1995) Forecaster Diversity and the Benefits of Combining Forecasts. *Management Science,* 41**,** 68-75.

Bates, J. M. & Granger, C. W. J. (1969) The Combination of Forecasts. *Operational Research Quarterly,* 20**,** 451-468.

Chow, G. C. (2011) Usefulness of Adaptive and Rational Expectations in Economics. *CEPS Working Paper, Princeton University,* No. 221.

Clemen, R. T. (1989) Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting,* 5**,** 559-583.

Clements, M. P. (2006a) Evaluating the Survey of Professional Forecasters Probability Distributions of Expected Inflation Based on Derived Event Probability Forecasts. *Empirical Economics,* 31**,** 49-64.

Clements, M. P. (2006b) Internal Consistency of Survey Respondents` Forecasts: Evidence Based on the Survey of Professional Forecasters. *The Warwick Economics Research Paper Series***,** No. 772.

Croushore, D. (1993) Introducing: The Survey of Professional Forecasters. *Business Review,* 6.

Croushore, D. (1996) Inflation Forecasts: How Good Are They? *Business Review***,** 15-25.

Croushore, D. (2006) An Evaluation of Inflation Forecasts from Surveys Using Real-Time Data. *Federal Reserce Bank of Philadelphia Working Paper* No. 06-19.

Croushore, D. (2009) Philadelphia Fed Forecasting Surveys: Their Value for Research.

Croushore, D. & Stark, T. (2001) A Real-Time Data Set for Macroeconomists. *Journal of Econometrics,* 105**,** 111-130.

D'Agostino, A., McQuinn, K. & Whelan, K. T. (2010) Are Some Forecasters Really Better Than Others? *Research Technical Papers*.

Diebold, F. X., Tay, A. S. & Wallis, K. F. (1997) Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters. *NBER Working Paper,* No. 6228.

Elliott, G. & Timmermann, A. (2008) Economic Forecasting. *Journal of Economic Literature,* 46**,** 3-56.

Federal Reserve Bank of Philadelphia. (2011). *Survey of Professional Forecasters Documentation* [Online]. Available: http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/spf-documentation.pdf [Accessed 25.01 2012].

Galati, G., Poelhekke, S. & Zhou, C. (2011) Did the Crisis Affect Inflation Expectations? *International Journal of Central Banking,* 7**,** 167-207.

Gerberding, C. (2006) Household Versus Expert Forecasts of Inflation: New Evidence from European Survey Data. Deutsche Bundesbank.

Giordani, P. & Söderlind, P. (2003) Inflation Forecast Uncertainty. *European Economic Review,* 47**,** 1037-1059.

Granger, C. W. J. & Newbold, P. (1974) Spurious Regressions in Econometrics. *Journal of Econometrics,* 2**,** 111-120.

Grant, A. P. & Thomas, L. B. (1999) Inflationary Expectations and Rationality Revisited. *Economics Letters,* 62**,** 331-338.

Gregory, A. W., Smith, G. W. & Yetman, J. (2001) Testing for Forecast Consensus. *Journal of Business & Economic Statistics,* 19**,** 34-43.

Hafer, R. W. & Hein, S. E. (1985) On the Accuracy of Time-Series, Interest Rate, and Survey Forecasts of Inflation. *The Journal of Business,* 58**,** 377-398.

Holden, K. & Peel, D. A. (1990) On Testing for Unbiasedness and Efficiency of Forecasts. *The Manchester School,* 58**,** 120-127.

Keane, M. P. & Runkle, D. E. (1990) Testing the Rationality of Price Forecasts: New Evidence from Panel Data. *The American Economic Review,* 80**,** 714-735.

Kershoff, G. & Smit, B. (2002) Conducting Inflation Expectation Surveys in South Africa. *South African Journal of Economics,* 70**,** 205-212.

Mankiw, N. G., Reis, R. & Wolfers, J. (2003) Disagreement About Inflation Expectations. *NBER Macroeconomics Annual,* 18**,** 209-248.

McNees, S. K. (1992) How Large Are Economic Forecast Errors. *New England Economic Review***,** 25-42.

Mehra, Y. P. (2002) Survey Measures of Expected Inflation: Revisiting the Issues of Predictive Content and Rationality. *Economic Quarterly-Federal Reserve Bank of Richmond,* 88**,** 17-36.

Mestre, R. (2007) Are Survey-Based Inflation Expectations in the Euro Area Informative. *ECB Working Paper Series,* No. 721.

Mincer, J. A. & Zarnowitz, V. (1969) The Evaluation of Economic Forecasts. *In:* Mincer, J. A. (ed.) *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance.* NBER, 1-46

National Bureau of Economic Research. (2010). *U.S. Business Cycle Expansions and Contractions* [Online]. Available: http://www.nber.org/cycles.html#announcements [Accessed 23.04 2012].

Newey, W. K. & West, K. D. (1987) A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica,* 55**,** 703-708.

Pearce, D. K. (1979) Comparing Survey and Rational Measures of Expected Inflation: Forecast Performance and Interest Rate Effects. *Journal of Money, Credit and Banking,* 11**,** 447-456.

Roberts, J. M. (1998) Inflation Expectations and the Transmission of Monetary Policy. *Finance and Economics Discussion Series,* No. 1998-43**,** 1-41.

Shiller, R. J. (2000) *Irrational Exuberance.* Princeton: Princeton University Press.

Smith, G. W. & Yetman, J. (2010) Estimating Dynamic Euler Equations with Multivariate Professional Forecasts. *Economic Inquiry*.

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

Stark, T. & Croushore, D. (2002) Forecasting with a Real-Time Data Set for Macroeconomists. *Journal of Macroeconomics,* 24**,** 507-531.

Stock, J. H. & Watson, M. W. (1999) Forecasting Inflation. *Journal of Monetary Economics,* 44**,** 293-335.

Su, V. & Su, J. (1975) An Evaluation of ASA/NBER Business Outlook Survey Forecasts. *Explorations in Economic Research,* 2**,** 588-618.

Thomas, L. B. (1999) Survey Measures of Expected Us Inflation. *The Journal of Economic Perspectives,* 13**,** 125-144.

Zarnowitz, V. (1984) The Accuracy of Individual and Group Forecasts from Business Outlook Surveys. *Journal of Forecasting,* 3**,** 11-26.

Zarnowitz, V. (1985) Rational Expectations and Macroeconomic Forecasts. *Journal of Business and Economic Statistics,* 3**,** 293-311.

Zarnowitz, V. & Braun, P. (1993) Twenty-Two Years of the NBER-ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance. *In:* Stock, J. H. & Watson, M. W. (eds.) *Business Cycles, Indicators and Forecasting.* University of Chicago Press, 11-94