Discussion Paper No. 06-087

Accuracy and Properties of German Business Cycle Forecasts

Steffen Osterloh

ZEW

Zentrum für Europäische
Wirtschaftsforschung GmbH

Centre for European
Economic Research

Discussion Paper No. 06-087

Accuracy and Properties of German Business Cycle Forecasts

Steffen Osterloh

Non-technical Summary

The high number of institutions publishing business cycle forecasts has always aroused public as well as scientific interest regarding the evaluation of their forecast accuracy. The emphasis of this paper is on German forecasts of real GDP growth. Related research does not find systematic differences between the forecasters regarding their relative accuracy, besides the date of the publication of the forecast. Therefore data published in the Consensus Forecasts survey is used, which guarantees a high degree of comparability between the individual forecasters as their date of publication is standardized. The sample used in this paper comprises the time span of 1995 until 2005 and includes mainly different types of banks, as well as public and private research institutes.

First, several standard descriptive measures are applied and the individual forecasters are ranked according to their accuracy. The results for forecasts with a horizon of more than 12 months are surprising as a simple naïve forecast shows the by far highest accuracy. The modified Diebold Mariano-test confirms this result of the bad performance of the forecasters compared to the naïve forecast for the longer forecast horizons. However, the Diebold-Mariano-test shows for no forecaster a systematically higher or lower forecast error compared to the arithmetic average of the participants for all horizons.

Two conditions for good forecasts are tested empirically to explain the differences in forecast accuracy. It is shown that the relatively bad accuracy can mainly be explained by a large overestimation found for all forecasters at the longer forecast horizons which decreases slowly towards the end of the target year. Another source of inaccuracy may be due to weak information efficiency which is tested via the assumption of unpredictability of forecast revisions. The results suggest that negative news have been incorporated too slowly, which impaired the adjustment of the forecasts from earlier optimistic views to new information. A further test looks at the imitation behaviour of forecasters. An imitation of the view of other forecasters can be confirmed empirically for the majority of the forecasters, though large differences in magnitude apply.

Finally, a nonparametric test is used to answer the question whether all forecasters were equal or if some forecasted better than the rest. This rank-sum test looks at the positions of the forecasters in a ranking which is calculated on the basis of the forecast errors for every forecasted year. It can be concluded that not all forecasters are equal. Some of them performed significantly better than a random distribution of the ranks would have suggested and showed a better forecast accuracy than other forecasters.

Accuracy and Properties of German Business Cycle Forecasts

Steffen Osterloh[*]
Centre for European Economic Research (ZEW)[†]

December 2006

Abstract: In this paper the accuracy of a wide range of German business cycle forecasters is assessed for the past 10 years. For this purpose, a data set is used comprising forecasts published on a monthly basis by Consensus Economics. The application of several descriptive as well as statistical measures reveals that the accuracy of the 2-years forecasts is low relative to a simple naïve forecast. This observation can mainly be explained by a systematic overestimation of the growth rates by the forecasters. Moreover, the lack of accuracy can also be explained partly by insufficient information efficiency as well as imitation behaviour. Finally, it is shown that notwithstanding the common errors which affected the accuracy of all forecasters mainly because of their systematic overestimation, they differ significantly in their forecast accuracy.

Keywords: business cycle forecasting, forecast evaluation, Consensus forecasts

JEL-Classification: C52, E32, E37

[*] I would like to thank the participants of the workshop "Makroökonomik und Konjunktur" in Dresden as well as Bernhard Boockmann, Friedrich Heinemann, Marcus Kappler and Waldemar Rotfuß for very helpful comments. All remaining errors are mine.

[†] Centre for European Economic Research (ZEW), P.O. Box 103443, D-68034 Mannheim, Germany, Phone: +49/621/1235-165, Fax: +49/621/1235-223; E-mail: osterloh@zew.de

Contents

1. Introduction 1
2. The data 2
    2.1. Consensus Forecasts 2
    2.2. Participants 3
3. Forecast accuracy 4
    3.1. Descriptive measures 4
    3.2. Application 6
4. Test for difference of forecast errors 8
5. Conditions for good forecasts 12
    5.1. Unbiasedness 12
    5.2. Information efficiency 15
6. Imitation behaviour 19
    6.1. Methodology 20
    6.2. Results 20
7. Test for differences in forecast accuracy 22
    7.1. Methodology 22
    7.2. Results 24
8. Conclusions 24

References 26

Appendix 28

1. Introduction

Forecasting the business cycle is one of the activities of economists which are most critically observed by the public. This forecasting comprises variables like GDP growth and its components, prices, unemployment or interest rates and is conducted by a variety of institutions: by public authorities for budgeting, by central banks for monetary policy, by research institutes for policy consultancy and banks for planning investment strategies. The high number of institutions publishing business cycle forecasts has always aroused public as well as scientific interest regarding the evaluation of their forecast accuracy. For the purpose of this evaluation, the science has developed a variety of statistical measures and empirical tests.

In this paper, the emphasis will be on German short-term forecasts (with a forecast horizon up to 24 months) of real GDP growth. This has recently been done extensively in the paper by Döpke and Fritsche (2006), who analyse the forecasting performance of 14 German institutions over a long period of more than 30 years, comprising the public research institutes, the Council of Economic Advisors (Sachverständigenrat), international organizations (OECD, IMF, European Community), some private institutes and the federal government. The authors do not find systematic differences between the forecasters regarding the relative accuracy, besides the date of the publication of the forecast. Thus, in working with intermittently published forecasts an important consideration arises: As they differ in their date of publication, differences in accuracy may not exist because of different economic models or abilities, but forecasters who publish later may be favoured because they possess more information. This problem is addressed in this paper by using the data published in the Consensus Forecasts survey, which has the special feature of a standardized date of publication.

The paper is organised as follows. First, in chapter 2 the data is introduced in detail. In chapter 3, several standard descriptive measures are introduced and applied to the data. In chapter 4, the differences in forecast accuracy are tested for empirical significance. In the subsequent chapter 5, two conditions for good forecasts which help to explain the differences in forecast accuracy are tested empirically. In addition, the special features of the data allow us to conduct a specific empirical test which looks at the imitation behaviour of forecasters (chapter 6). This is followed by a test which looks at systematic differences in the forecast accuracy of the individual forecasters in chapter 7. The final chapter 8 concludes.

2. The data

2.1. Consensus Forecasts

The data used in this analysis is available from Consensus Forecasts, a monthly survey conducted by the London-based company Consensus Economics. This survey, which is conducted since 1989, publishes forecasts of a variety of forecasters for key macroeconomic variables (in addition to GDP these include its components, prices, industrial production, unemployment and interest rates) for currently 70 countries.

Every month, each forecaster submits two point-forecasts for these variables, one for the current and one for the following year. These questionnaires also contain a precise definition of the predicted variables to ensure comparability. Consensus Forecasts is published in the second week of each month, based on the survey of the panellist forecasts in the two weeks before, with a common deadline (Harvey et al. 2001). This standardization of the date of publication and the definitions for all forecasters guarantees a high degree of comparability of the forecasts.

In addition to the individual participants´ forecasts, the arithmetic average of all forecasters is published for every predicted variable, which is broadly known as the ‘Consensus’. This pooled value is often used by forecasters in combination with their own forecast to communicate to the public their own view relative to the generally expected value.

The relatively high frequency of publications of Consensus Forecasts, which is much higher than the usual publication cycles of the well-known forecasters like IMF, OECD or the German institutes, who often only produce 2-4 forecasts a year, allows the application of a number of specialized empirical procedures. This special feature of the data set is called “fixed-event” forecast in the literature, as one fixed event (the GDP growth rate for the year T) is predicted at a high number of horizons. This contrasts to the usually analyzed “fixed-horizon” forecasts, where point estimates for many years are conducted at only one fixed horizon[1].

In this case of a fixed-event forecast, the monthly forecasting cycle begins in January of the previous year and ends in December of the forecasted year. Forecasts are therefore provided monthly by each participant, moving from horizons of 24 months up to 1 month ahead, producing altogether 24 forecasts. In the following, forecasts for a given year which were made in the same year (at a horizon of 12 months or less) are considered as “current year”

[1] An example for this is the Council of Economic Advisers (Sachverständigenrat), whose forecast for the next year is always published in November of the year before.

forecasts, forecasts produced in the year before (with a horizon of 13-24 months) are denoted “next year” forecasts.[2]

2.2. Participants

The sample used in this paper is restricted to the GDP forecasts for Germany and comprises the time span of 1995 until 2005.[3] This means that the current year forecasts for 1995–2005 and the next year forecasts for 1996-2005 are available. During this period, on average, 30 forecasters participated in the survey each month. These various German participants can be grouped according to their background:

1. Public research institutes: ifo, DIW, RWI, HWWA, IfW
2. Banks:
    a. ‘Großbanken’ (Deutsche Bank, Commerzbank, Dresdner Bank, HVB)
    b. ‘Landesbanken’ (e.g. Helaba, Bayerische LB, West LB)
    c. Co-operative central banks (DZ Bank, WGZ Bank)
    d. Private banks (e.g. Bank Julius Bär, Sal. Oppenheim)
    e. Affiliates of foreign banks (e.g. HSBC Trinkaus & Burkardt, SEB Germany)
    f. Foreign investment banks (e.g. JP Morgan, UBS Warburg)
3. Others:
    a. Private institutes: FAZ Institut, IW, Economist Intelligence Unit
    b. Industry: Hoechst AG

A problem which appears when working with this data set concerns mergers and acquisitions. The composition of the Consensus Economics panel changed several times because the research departments of banks changed their owner and their name. In order to arrive at a continuous time series for the respective banks, the consistency of the forecasts was used as criterion. In the case of the merger of DG Bank and GZ Bank to DZ Bank, the first forecasts published by the new DZ Bank almost completely matched the last forecasts of DG Bank, such that a continuity of the work of the research department was assumed. In cases where it was not possible to get a clear cut result from the figures alone, the criteria of the consistency of the staff was used. In the case of the acquisition of BfG Bank by SEB,

[2] For example, next year forecasts for the GDP growth in 2005 were produced in every month between January and December 2004, current year forecasts between January and December 2005.
[3] Due to the unavailability of 5 issues of Consensus Forecasts, few values had to be interpolated without influencing the explanatory power of the results.

the later economists of SEB Germany were identical to the economists of the former BfG bank research department.

In some cases the continuation of the time series was not possible, as in the case of the merger of HYPO Bank and Bayerische Vereinsbank to HypoVereinsbank. Both banks had participated in the survey before their merger and it was not possible to determine whose researchers dominated the later HVB forecasts.

Another problem was caused by missing values because of the discontinuous participation of the respective forecasters in the survey. This was mostly the case when a forecaster simply forgot to report his forecasts to Consensus Economics. If this happens, the previous forecast is not inserted by Consensus Economics, but no value is published. Yet, for some forecasters missing values appeared more generally. This mainly concerned the German research institutes, whose forecasts are available for almost all of the predicted years, but in most cases they show many missing values at single forecast horizons. This reflects the fact that these institutes usually do not revise their forecasts as often as banks do.

Another group of forecasters published regularly for all horizons, but either stopped participating after some years (very often due to an acquisition) or did not start in 1995 but later.[4] In both cases the implementation of several empirical tests for the respective forecasters becomes impossible. That is why for many of the following empirical tests, different participants had to be dropped, in some cases because the number of forecasted years was too low, in other cases because they did not participate at enough horizons.

3. Forecast accuracy

3.1. Descriptive measures

A first evaluation of the forecast accuracy of the individual forecasters and their pooled forecast, the Consensus, is made with some standard measures, which are briefly introduced (for an overview see Döpke 2003).

A central measure, the forecast error, is generally defined as $e_t = F_t - R_t$, with F_t representing the predicted value for the year t, and R_t the realisation of the value in t. One aspect discussed controversially in the literature considers the question which value should be used as the realisation. The first preliminary figures for German GDP growth are published in January of the following year, but these figures can also be regarded as

[4] Table A1 in the Annex shows the participation of the forecasters and mergers or acquisitions in detail.

estimates as they are usually revised in the course of the following months. It is therefore argued that later publications are more accurate than the first preliminary figures. But in the course of time GDP figures are also often revised due to changes in methodology, which makes them difficult to compare with the initially forecasted values. Therefore Batchelor (2000) proposes to use the actual values published in the middle of the following year as the values used for forecast error computation. In this paper, the values which have been published as realisation in the Consensus Forecasts issue in June of the following year are taken as actual values.

For the descriptive analysis of the accuracy of the forecasters, the following measures are considered:

1. Mean Error: $ME = \frac{1}{T}\sum_{t=1}^{T} e_t$

   The mean error is simply the average of the forecast errors. But it can not be regarded as an appropriate measure for the evaluation of the accuracy, as the simple averaging of the forecast errors leads to a cancelling out of positive and negative errors. This means that a forecaster may have very high positive and negative errors, but show a ME of zero. The ME can instead be regarded as a measure of bias, a positive ME indicating a systematic overestimation of the GDP, a negative value a systematic underestimation. These properties will be discussed in chapter 5.1 in greater detail.

2. Mean absolute error: $MAE = \frac{1}{T}\sum_{t=1}^{T} |e_t|$

   The MAE averages the absolute errors over all periods, giving positive and negative deviations of the same size the same weight. Using this measure (as well as the following ones), one has to assume a symmetric loss function. This means that negative errors amount to the same loss as positive errors of the same magnitude, which is usually assumed in analysing business cycle forecasts.

3. Mean squared error: $MSE = \frac{1}{T}\sum_{t=1}^{T} e_t^2$ and

4. Root mean squared error: $RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_t^2}$

   The idea behind squaring the forecast errors as in these two measures is that large errors should be weighted more than small errors. While an error of 2% is twice as severe as an error of 1% using the MAE, it is four times as severe using the MSE. As it is a widely accepted fact that the main target of GDP forecasting should be to avoid large errors, the RMSE has become the main instrument for measuring forecast accuracy.

4. Theil´s U: $U = \frac{RMSE(Model)}{RMSE(Alternative\_Model)}$

Theil´s U (inequality coefficient) has been introduced to provide comparability between forecasts of various variables which have different variances. This is usually the case if one of the variables is more difficult to predict than the other. For this purpose, the RMSE of the forecast of concern is usually divided by the RMSE of a 'naïve' forecast which is used as alternative model, e.g. a random walk model. A value of lower than 1 shows that the model performs better than the alternative model, whereas a value higher than 1 shows that the alternative model is better.

3.2. Application

The ranking of the forecasters according to their accuracy can take place by simply calculating the RMSE of every forecaster over all periods and all forecast horizons. This has been done by Blix et al. (2001) for several OECD countries. But this is problematic in the present case, as the data set contains many missing values as discussed in chapter 2.2. This is critical because the difficulty to forecast differs both between horizons and target years. Figure 3.1 shows that the average RMSE (across all forecasters and target years) diminishes highly while getting closer to the end of the predicted year. This is in line with the expected trend: Getting closer to the end of the predicted year, the forecasters have available more information, and the forecasting becomes easier and more accurate.

Figure 3.1 Average RMSE 1995-2005

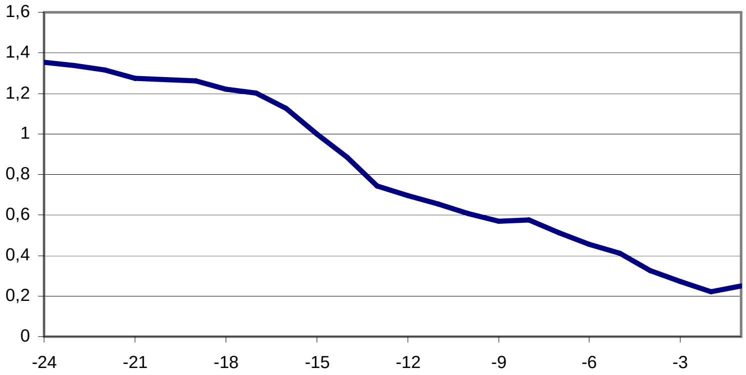
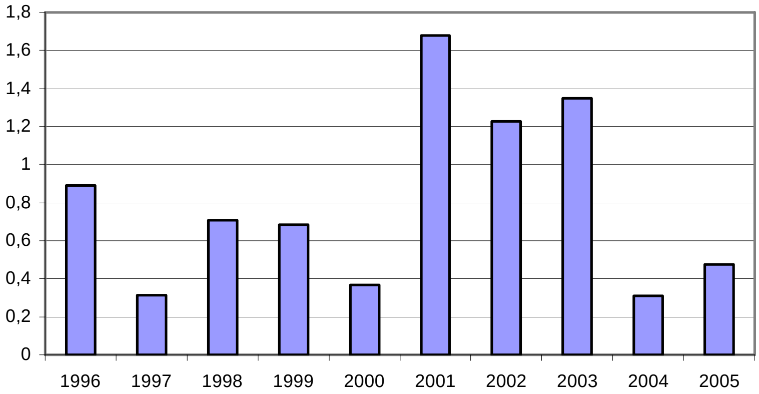


Figure 3.2 shows that the average RMSE (over all periods and forecasters) for some target years (2001-2003) have been more than five times as high as for other years (1997 and 2000). This confirms the assumption that some years are much more difficult to predict than others.

Figure 3.2 Average RMSE for target years



If some forecasters mainly published in the periods which were easy to predict, this would lead to a low RMSE, but would not say anything about the individual forecasting ability. To allow for this, a variant of Theil´s U has been developed. As alternative model not a naïve forecast has been used, but the Consensus. Its RMSE was calculated only for the periods where the individual forecaster also participated. In other words, periods where values of an individual forecaster were missing have been dropped out of the RMSE of the Consensus before calculating Theil´s U. Therefore, Theil´s U value has to be interpreted as the relative accuracy compared to the Consensus for all of the periods where the individual forecaster participated. A value of lower than 1 shows that the individual forecaster was more accurate than the Consensus, a value higher than 1 shows that the Consensus performed better.

Tables A.2 and A.3 in the annex show the descriptive statistics for the individual forecasters, limited to those forecasters who participated in at least 6 out of the 11 years. In addition, the Consensus and a naïve forecast are published as benchmarks.

The choice of a naïve forecast is more or less arbitrary. Often simply the last available actual value is used to account for short term trends (“no change” forecast). Others use a long-term growth average of sometimes 20 or more years to avoid a domination of short-time cyclical effects. Here, the naïve forecast has been calculated as the average of the published GDP growth rate of the respective past three years (rolling average). This is done to give the most recent growth rates a high weight, but avoid a domination of the cyclical component of the growth rate of a no change forecast.

The forecasters are ranked according to their Theil´s U. In addition, the RMSE and MAE are provided, as well as a measure which shows the percentage of the periods where the forecaster's value was closer to the realisation than the Consensus. It can easily be seen that in both rankings the different measures do not result in the exact same order, but still a ranking based on any other measure would be similar to the Theil´s U criteria.

The position of the Consensus is in both rankings above the average, which is consistent with other authors´ results. These results are also confirmed by theory, as McNees (1992) discusses. Extreme forecasts (which are very often wrong) cancel out, which lets McNees say that "many heads are better than one".

Table A.2 shows the results of the current year forecasts only (comprising the horizons of 12 until 1 months). Here, unsurprisingly, the naïve forecast reaches the last position. This is what one would expect, as this naïve forecast only contains the information of the past three years, but does not account at all for the development in the current year.

A very different picture emerges from the values of the next year forecasts (table A.3), which comprise the horizons of 24 until 13 months. Very surprisingly, the naïve forecast performs by far the best. This striking result, which to the author's knowledge can nowhere be found in the literature, has to be explained in the following chapters.

Regarding the relative accuracy of the individual forecasters, it can be observed that in both rankings mainly less renowned forecasters rank at the first positions, confirming a similar result by Blix et al. (2001). The three forecasters who constantly show the highest accuracy are HSBC Trinkaus & Burkhardt, BfG Bank (later SEB Germany) and MM Warburg, while more prestigious forecasters like one research institute and one 'Großbank' can be found at the last ranks. But this result has to be regarded cautiously. As will be shown in the next chapters, this period seems to be quite special in its predictability. In addition, the results of the study conducted by Blix et al. for an earlier period show very different results. There, Trinkaus & Burkardt was among the worst, while some underperformers of this study are among the best. Therefore, a generalization of the results should not be made.

4. Test for difference of forecast errors

In addition to the descriptive measures presented in the previous chapter, a statistical test introduced by Diebold and Mariano (1995) is used to test the differences of forecast errors for statistical significance. This test is an asymptotic t-test, testing for the null hypothesis that the difference of the mean squared errors of two forecast models A and B is zero for a given forecast horizon, i.e. $\overline{d} = MSE_A - MSE_B = 0$.[5] The test statistic of the Diebold-Mariano test is

$$DM = \frac{\overline{d}}{\sqrt{\hat{V}(\overline{d})}},$$

which follows a t-distribution with (N-1) degrees of freedom, N = number of forecasts.

In this test, the variance $\hat{V}$ is estimated robustly as in the work of Newey and West (1987). This allows to account for the autocorrelation problem due to overlapping periods. As for a 2-step forecast an unpredicted event does not only affect the forecast error for the current year (which was predicted in the year before), but also for the next year (which has been predicted earlier in the year of the event), a positive autocorrelation between the two forecast errors can be expected.

In this paper, a modified version of the Diebold-Mariano test is used. This has been proposed for small samples by Harvey, Leybourne and Newbold (1997) (see e.g. Schröder 2002). The test statistic of this modified version is $mDM = C \cdot DM$, with the correction factor

$$C = \left(\frac{N + 1 - 2h + N^{-1}h(h-1)}{N} \right)^{\frac{1}{2}},$$

with N = number of forecasts and h = steps of forecast. The calculated correction factors are 0.849 for the next year forecasts and 0.953 for the current year forecasts.

Table 4.1 shows the results of the bilateral comparisons of the individual forecasters with the Consensus. The forecast horizons of 23, 18, 11 and 6 months have been used. Only forecasters have been considered who published forecasts in every year for the selected horizons.

Comparing individual forecasts with the Consensus may be problematic, because all participants' forecasts are part of the Consensus and therefore correlated. But as each individual forecast represents only a very small share of the Consensus (around 1/30), this problem is disregarded in the following.

It can be seen that both negative values (forecaster has lower MSE than Consensus) and positive values (Consensus has lower MSE) can be observed, but in most cases the null hypothesis that the forecast errors are equal can not be rejected. No forecaster has positive or negative results significantly different from zero for more than two horizons. Therefore,

[5] Diebold and Mariano (1995) consider a general loss function. For the evaluation of business cycle forecasts, it is common to use a MSE loss differential.

it can be said that this test does not allow us to identify any forecaster who performed significantly better or worse than the Consensus for all horizons.

Table 4.1 Diebold Mariano test, accuracy compared to Consensus

	Horizon			
	-23	-18	-11	-6
Institute 2		-0.12* (0.058)		-0.04 (0.061)
Institute 4		-0.01 (0.133)	0.23** (0.098)	0.30** (0.137)
Institute 5			0.37** (0.146)	
Others 1	-0.30 (0.380)	0.22* (0.110)	0.11 (0.088)	0.12** (0.048)
Großbank 1	0.18 (0.403)	0.53 (0.345)	-0.21* (0.137)	0.05 (0.044)
Großbank 2	0.21 (0.355)	0.31** (0.114)	0.00 (0.046)	0.06*** (0.018)
Großbank 4	0.06 (0.135)	0.28 (0.208)	0.31** (0.153)	0.01 (0.087)
Co-operative 1	0.02 (0.109)	0.11* (0.059)	0.06* (0.037)	0.03 (0.025)
Co-operative 2	0.42 (0.294)	0.34 (0.238)	0.08 (0.173)	-0.02 (0.051)
Landesbank 1	-0.55* (0.293)	-0.14 (0.089)	-0.05 (0.053)	0.05 (0.038)
Landesbank 2	0.30 (0.319)	0.18 (0.301)	0.14** (0.047)	0.05 (0.043)
Landesbank 3	0.50 (0.511)	0.31 (0.269)	0.04 (0.077)	0.03 (0.021)
Landesbank 4	0.58** (0.229)	0.48* (0.235)	0.08 (0.147)	0.04 (0.039)
Landesbank 5	-0.32* (0.191)	-0.17* (0.098)	0.19 (0.150)	0.03 (0.036)
Affiliate 1	-0.20 (0.185)	-0.12 (0.173)	-0.07 (0.054)	-0.04 (0.044)
Affiliate 2	-0.66 (0.461)	-0.39 (0.340)	-0.02 (0.085)	-0.02 (0.028)
Affiliate 3	0.08 (0.428)	0.67** (0.301)	-0.04 (0.182)	-0.01 (0.078)
Private 1	-0.07 (0.110)	0.16 (0.163)	-0.27** (0.130)	-0.05* (0.030)
Private 2	-0.23 (0.235)	0.01 (0.144)	-0.14** (0.063)	0.01 (0.019)

The symbol ***,**, and * denotes rejection of the null hypothesis at the 1%, 5%, and 10% level respectively. Standard errors in parentheses.

Table 4.2 applies the modified DM-test to a bilateral comparison with the naïve forecast introduced in chapter 3.2. This test confirms the results of the descriptive statistics. For the two longer horizons (next year forecasts), the null hypothesis of equal mean squared errors can be rejected. The naïve forecast performs better than any individual forecast (negative values), being always at least significantly different from zero at the 10% level, in most cases also at the 5% level.

Table 4.2 Diebold Mariano test, accuracy compared to naïve model

	Horizon			
	-23	-18	-11	-6
Institute 2		0.68** (0.305)		-0.87* (0.491)
Institute 4		0.79** (0.309)	-0.23 (0.329)	-0.53 (0.401)
Institute 5			-0.08 (0.345)	
Others 1	0.95** (0.418)	1.01** (0.389)	-0.35 (0.340)	-0.71* (0.443)
Großbank 1	1.42* (0.718)	1.33* (0.683)	-0.66** (0.332)	-0.78** (0.401)
Großbank 2	1.45** (0.475)	1.11** (0.385)	-0.46* (0.298)	-0.77* (0.443)
Großbank 4	1.31** (0.455)	1.08** (0.467)	-0.14 (0.278)	-0.82* (0.442)
Co-operative 1	1.26** (0.453)	0.91** (0.391)	-0.39 (0.310)	-0.80* (0.451)
Co-operative 2	1.67** (0.721)	1.13* (0.559)	-0.38 (0.420)	-0.85** (0.418)
Landesbank 1	0.70** (0.261)	0.65* (0.363)	-0.50* (0.311)	-0.78* (0.436)
Landesbank 2	1.55** (0.471)	0.98* (0.526)	-0.32 (0.297)	-0.78* (0.467)
Landesbank 3	1.75* (0.940)	1.10* (0.609)	-0.42 (0.302)	-0.80* (0.434)
Landesbank 4	1.82** (0.679)	1.28** (0.491)	-0.38* (0.235)	-0.79* (0.428)
Landesbank 5	0.93* (0.534)	0.63* (0.314)	-0.26 (0.295)	-0.80* (0.447)
Affiliate 1	1.05** (0.446)	0.68* (0.331)	-0.52* (0.337)	-0.87* (0.470)
Affiliate 2	0.59** (0.231)	0.40* (0.217)	-0.47* (0.292)	-0.85** (0.422)
Affiliate 3	1.33** (0.504)	1.47** (0.607)	-0.49 (0.424)	-0.84* (0.503)
Private 1	1.17** (0.442)	0.95* (0.452)	-0.72** (0.338)	-0.88** (0.435)
Private 2	1.02* (0.479)	0.81* (0.388)	-0.59* (0.332)	-0.82* (0.435)
Consensus	1.25** (0.499)	0.80** (0.359)	-0.46* (0.300)	-0.83* (0.437)

The symbol ***,**, and * denotes rejection of the null hypothesis at the 1%, 5%, and 10% level respectively. Standard errors in parentheses.

Analogous to the results of the descriptive statistics, this changes with the forecast horizons getting shorter. For all forecasters the sign becomes negative for the two horizons of the current year forecasts (significantly different from zero in most cases at the 10% level), indicating that their errors are lower than the errors of the naïve forecast. In most cases this coincides with a constant shift of the difference of errors with declining horizons, which shows that the forecasts fare better compared to the naïve forecast the closer you get to the end of the predicted year.

Summing up, the results of the modified DM-test confirm the finding of the descriptive statistics of a surprisingly bad performance of the individual forecasters as well as the Consensus compared to a simple naïve forecast for the next year forecasts. However, this test is not able to answer the question if any of the individual forecasters performed better or worse than the other participants. The respective test can not show whether there are any forecasters who show systematically higher or lower forecast errors than the Consensus.

5. Conditions for good forecasts

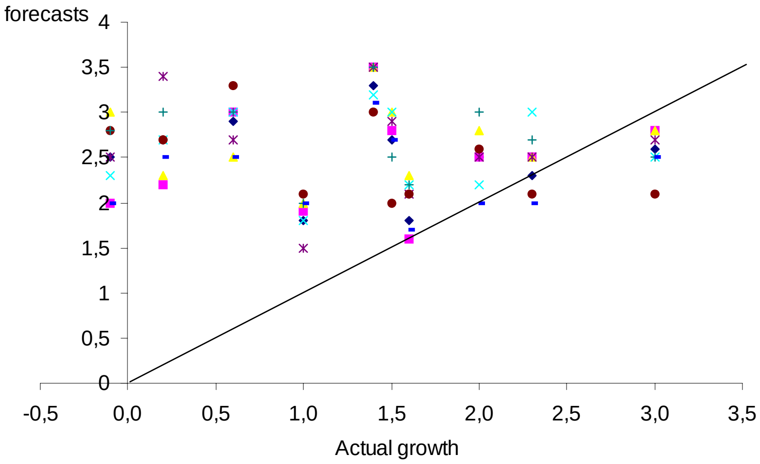
The following sections are intended to explain the findings of the surprisingly bad accuracy of the next year forecasts found in the previous chapters. For this purpose two main conditions for good forecasts are introduced and tested empirically, unbiasedness and information efficiency.

5.1. Unbiasedness

A first condition for good forecasts which is analyzed is unbiasedness. A forecast is considered to be biased if it is systematically too high or too low. If this is the case, a forecast is suboptimal because it can easily be improved on the basis of the bias known from past forecasts. In the case of an upward bias, this improvement could easily be made by subtracting the average overestimation from the forecast.

A first impression of a possible bias can be drawn from plotting the actual values against the respective forecasts. Figure 5.1 shows this for the longest forecast horizon (24 months) for a number of forecasters which have been chosen arbitrarily out of the sample. A perfect forecast where it exactly matches the realisation would lie on the 45° line, unbiased forecasts should cluster around the line. It can easily be seen in the graph that most observations are above the line. This represents cases where the forecasts were higher than the actual values, which means that the growth rates were overestimated in the respective years.

Figure 5.1 Actual and forecasted (24 months ahead) GDP growth rate



To verify this observation empirically, a simple t-test is used, regressing the forecast errors for a given forecast horizon on a constant: $e_t = \alpha + u_t$ (Fildes and Stekler 2002). The null hypothesis, that the forecasts are unbiased, would hold if α =0. A negative value would show an underestimation, a positive value an overestimation. For this test a normal distribution of the forecast errors has to be assumed, which can not be rejected on the basis of the Jarque-Bera statistics. Similarly to the Diebold-Mariano-test, robust standard errors (Newey/West) have to be used because of possible autocorrelation.

Table 5.1 shows the result for the 4 horizons (23, 18, 11 and 6 months) for all forecasters who participated in all of the years at the given horizons. It can be observed that all forecasters showed for the two longest horizons highly positive biases which turn out to be highly significantly different from zero in all cases. The null hypothesis of unbiasedness can be rejected there. For the longest horizon, they show values of 0.84-1.37, which suggests that the GDP growth rate was overestimated on average by more than 1 percentage points by most forecasters, and by 1.16 points by the Consensus. Getting closer to the end of the predicted year, this bias gets smaller in magnitude and loses significance in some cases. But even 6 months before the end of the target year, in many cases a significant positive bias can be observed, in the case of the Consensus it reaches 0.25 percentage points.

Table 5.1 Test for biasedness

	Horizon			
	-23	-18	-11	-6
Institute 2		0.95** (0.302)		0.16* (0.080)
Institute 4		1.02*** (0.297)	0.54** (0.188)	0.32* (0.161)
Institute 5			0.74*** (0.136)	
Others 1	1.13*** (0.268)	1.14*** (0.296)	0.55*** (0.150)	0.25* (0.133)
Großbank 1	1.20** (0.382)	1.13** (0.380)	0.35* (0.180)	0.31** (0.126)
Großbank 2	1.32*** (0.292)	1.22*** (0.287)	0.50*** (0.149)	0.35*** (0.083)
Großbank 4	1.24*** (0.318)	1.13*** (0.271)	0.69*** (0.161)	0.35*** (0.100)
Co-operative 1	1.20*** (0.309)	1.02*** (0.312)	0.55*** (0.150)	0.32*** (0.093)
Co-operative 2	1.28*** (0.371)	1.13*** (0.322)	0.58*** (0.149)	0.35*** (0.098)
Landesbank 1	0.95** (0.319)	0.85** (0.343)	0.37** (0.165)	0.12 (0.172)
Landesbank 2	1.13** (0.436)	1.05** (0.360)	0.52** (0.174)	0.07 (0.142)
Landesbank 3	1.12** (0.479)	0.91** (0.396)	0.42* (0.195)	0.19 (0.139)
Landesbank 4	1.37*** (0.359)	1.14*** (0.332)	0.48** (0.163)	0.31*** (0.080)
Landesbank 5	0.92** (0.399)	0.89** (0.323)	0.56** (0.193)	0.23* (0.118)
Affiliate 1	1.05** (0.345)	0.96*** (0.294)	0.34* (0.157)	0.19* (0.100)
Affiliate 2	0.84** (0.285)	0.75** (0.278)	0.45** (0.174)	0.18* (0.099)
Affiliate 3	1.21*** (0.268)	1.08** (0.332)	0.49*** (0.106)	0.28*** (0.047)
Private 1	1.16*** (0.311)	1.05** (0.328)	0.32* (0.149)	0.23** (0.093)
Private 2	1.07*** (0.325)	0.95** (0.334)	0.28 (0.173)	0.24* (0.121)
Consensus	1.16*** (0.341)	1.00*** (0.303)	0.50*** (0.149)	0.25** (0.102)
Naïve	0.26 (0.392)	0.26 (0.392)	0.09 (0.322)	0.09 (0.322)

The symbol ***,**, and * denotes rejection of the null hypothesis at the 1%, 5%, and 10% level respectively. Standard errors in parentheses.

This test has also been conducted for the naïve forecast introduced in chapter 3.2. This naïve forecast only shows a very small positive bias, which is not significantly different from zero. The null hypothesis of unbiasedness can not be rejected here. Comparing this result with the results of the individual forecasters, a first explanation can be drawn for the bad accuracy of the forecasters for the next year forecasts. It seems that the high positive bias in the early forecasts accounts for the much better accuracy of the naïve forecast for the two longer horizons as shown by the Diebold-Mariano test in chapter 4.

5.2. Information efficiency

Another condition of a good forecast is information efficiency. This means that an efficient forecast should efficiently incorporate a certain set of information which is known at the moment of the production of the forecast. If this is not the case, the forecast can easily be improved by incorporating the missing information.

5.2.1 Theory

The strong definition of information efficiency comprises all information which is available at the date of the production of a forecast. This means that a forecast is said to be strongly information efficient if it efficiently incorporates any available information. It is obviously very difficult to test for strong efficiency empirically as it is practically impossible to find all data available. For practical use, tests are usually restricted to some key variables like interest rates, oil prices or business surveys.

A more feasible definition is weak information efficiency. In this case the set of information is restricted to the past forecast errors. Weak information efficiency holds if a forecast efficiently incorporates all information about its past forecast errors. These errors therefore have to be unpredictable. For this purpose, the following regression is commonly used: $e_t = \alpha + \beta e_{t-1} + u_t$. The null hypothesis is that β=0, which means that the current forecast error can not be explained by its past errors. Otherwise it would be possible to improve the forecast on the basis of the knowledge from past errors.

Because of the low number of predicted years, this test seems to be inappropriate for the data set used in this paper. For the special case of a fixed-event forecast, Nordhaus (1987) proposes a specific test which does not aim at the unpredictability of the forecast errors, but at the unpredictability of forecast revisions. In his approach, if weak efficiency holds, the forecast revision process should look like a random walk $_jF_T = {}_kF_T + \sum_{s=k+1}^{j} \varepsilon_s$, j>k, s=(k,...,j).

Each value predicted in later periods $_jF_T$ should be a combination of the initial published value $_kF_T$ and the sum of the revisions ε_s in the periods before. These revisions have an expected value of zero and should be identically independent distributed. Nordhaus (1987) gives the following intuition for this idea: “If I could look at your most recent forecasts and accurately say, ‘Your next forecast will be 2 percent lower than today’s,’ then you can surely improve your forecasts.”

The test for this random walk behaviour therefore looks at the correlation of forecast revisions, which are defined as ${}_t v_T = {}_t F_T - {}_{t-1}F_T$, i.e. the difference between the forecast in the current and the previous period for the same target year. Weak efficiency holds if these are unpredictable (white noise, $\mathrm{cov}({}_{t-1}v_T, {}_t v_T) = 0$). To test for this, the following model is estimated: ${}_t v_T = \alpha({}_{t-1}v_T) + u_t$, under the null hypothesis that the revisions are unpredictable: $H_0 : \alpha = 0$.

Very often, this test is not applied to the predictability of the revisions of subsequent months, but subsequent quarters. This has to be done because many forecasters usually do not revise their forecasts every month, but only every two or three. If, for example, one forecaster revises his forecasts every two month in the same direction, it would be predictable and therefore not information efficient, but a test using monthly revisions would not capture this behaviour. But a positive autocorrelation would be found if the quarterly revisions (the difference between the current forecast and the forecast published 3 months before) are regressed on their lagged values.

5.2.2 Application

The results of the test applied to the Consensus and a number of individual forecasters can be found in the appendix (table A.4). Both monthly and quarterly revisions have been analyzed. Many participants had to be dropped because of too few subsequent periods with a revision of the forecast. The German research institutes constitute the most extreme case: For none of them more than 5 cases of subsequent monthly revisions could be found.

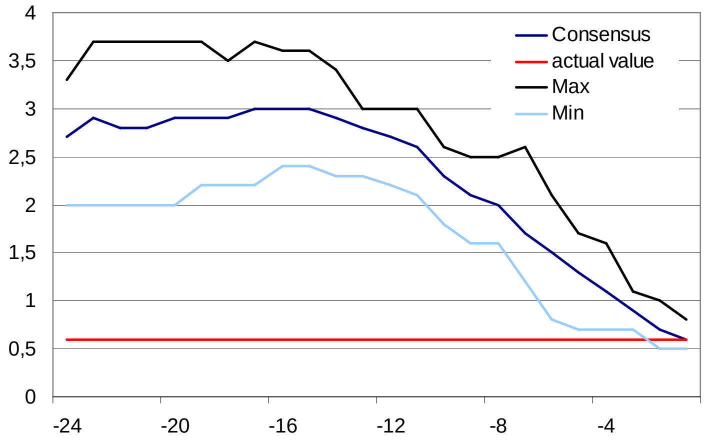
The results of the estimation using ordinary least squares (OLS) show in almost all cases positive values. They are highly significantly different from zero for the Consensus and in some cases for individual forecasters, which lets us reject the null hypothesis of unpredictability. For some of the forecasters they are only significant for the quarterly revisions. This represents forecasters who are seemingly reluctant to revise their forecasts in two subsequent months.

However, these results found using OLS are problematic for two reasons. Firstly, as Harvey et al. (2001) notice, the standard assumption of uncorrelated error terms may be violated. As

However, these results found using OLS are problematic for two reasons. Firstly, as Harvey et al. (2001) notice, the standard assumption of uncorrelated error terms may be violated. As in every period both the next and the current year forecasts may be revised because of the dissemination of news, a positive autocorrelation between these revisions can be suspected, i.e. $E({}_{t-12}\varepsilon_{T+1}, {}_{t}\varepsilon_{T}) \neq 0$. Therefore Harvey et al. propose the use of a consistent OLS or GLS estimator.

Moreover, a second problem casts the general applicability of the Nordhaus test for this data set into doubt. The general assumption of an expected value of zero for the forecast revisions seems to be violated. For all the participants the number of negative revisions is much higher than the number of positive revisions. The Consensus shows 30 upward and 92 downward revisions. Its mean revision has a value of -0.06 percentage points. This also holds for all of the individual forecasters, e.g. the forecasters with the most revisions, Landesbank 3 (34 upward, 75 downward) and Co-operative 2 (23 upward and 67 downward). The reason for this can be seen from figure 5.2. This shows the development of the Consensus and the maximum and the minimum value of the individual forecasters in a year with a very high number of negative revisions. As it has been shown in chapter 4.2, in an average year all forecasters started out with a large overestimation of the growth rate for the 24-months forecast. With decreasing horizons towards the end of the target year, more news became available contradicting their optimistic views. This caused a high necessity to revise the forecasts downwards, which can be seen in the downward trend of the Consensus forecast in the chart. This pattern can be observed for a high number of the years between 1995 and 2005.

Figure 5.2 Forecast revision process (target year: 2001)



A measure introduced by Isiklar et al. (2005) is used to account for the violation of the assumption of an expected value of zero for the forecast revisions. They compare the frequency distribution of the direction of the forecast revisions (upward or downward) with the expected distribution under the hypothesis of independence. Under the null hypothesis, the probability of the direction of a revision should be independent from the direction of the revision one month earlier and therefore be equal to its expected value.

This is applied to the monthly revisions of the Consensus and some selected individual forecasters (those who show the highest number of monthly revisions) in table 5.2. Regarding positive revisions, for all forecasters a very low number of revisions in two subsequent months can be observed. This value is mostly similar to the expected value which is also very low because of the low number of positive revisions in general. But for the negative revisions, a much higher frequency of subsequent revisions can be seen for the Consensus than to be expected under the null hypothesis of independence. That is, given the hypothetical independent distribution, approximately 38 cases of two negative revisions in series would be expected, but actually 52 can be observed. This is also the case for the majority of the individual forecasters.

Table 5.2 Subsequent forecast revisions

	Frequency / Expected frequency	
	${}_t v_T$ >0 and ${}_{t-1} v_T$ >0	${}_t v_T$ <0 and ${}_{t-1} v_T$ <0
Consensus	4 / 4.1	52 / 38.1
Others 1	1 / 1.2	28 / 18.9
Großbank 1	3 / 2.9	20 / 13.8
Landesbank 1	1 / 2.5	15 / 14.5
Landesbank 3	5 / 5.4	28 / 26.3
Landesbank 5	1 / 1.9	17 / 15.9
Co-operative 1	2 / 1.6	23 / 14.2
Co-operative 2	1 / 2.4	26 / 20.6
Affiliate 1	4 / 1.9	20 / 14.7
Affiliate 3	5 / 2.1	17 / 15.3
Foreign 5	2 / 2.0	19 / 20.9

Without going further into an empirical analysis, these results point to a possible lack of information efficiency regarding the negative revisions. In most years, the forecasters started the forecasting cycle with an overestimation at the 24-months-ahead forecast. With the appearance of news, the forecasters revised their forecasts downwards. But the positive correlation and the higher than expected occurrence of subsequent periods with downward revisions show that these revisions were predictable. It seems that the news were not incorporated by conducting large negative revisions, but many small steps of subsequent negative revisions.

5.2.3 Explanations

Nordhaus (1987) gives two possible explanations for this behaviour of the individual forecasters which he calls “forecast smoothing”. First, he argues that professional forecasters show a specific utility function which makes them fearful that “jumpy” forecasts will be treated as inconsistency by their bosses or customers. Therefore a high jump which is indicated by new data or events is distributed over several small jumps in subsequent months. They are especially reluctant to make a positive revision following a negative one or vice versa, even if the data points to this direction.

The other explanation is a psychological one, saying that forecasters tend to hold to their prior views too long, and therefore incorporate data opposing their views too slowly.

A possible explanation for low information efficiency of the Consensus can be found in a herding behaviour. If news is not incorporated by all forecasters at the same time, it is not reflected in a large revision of the Consensus, but spread via smaller revisions over several months. This will be discussed further in the next chapter.

6. Imitation behaviour

One further possible source of the inaccuracy of forecasts is introduced by Gallo, Granger and Jeon (2002), who suspect an imitation behaviour of forecasters. This is the case when views expressed by other forecasters in the previous periods have an influence on an individual’s current forecast. They confirm this empirically by using the lagged mean value (the Consensus) as a proxy for the view of the other forecasters.

This imitation behaviour might reduce the forecast accuracy as it leads to a convergence to “a value which is not the “right” target”. Gallo et al. explain this behaviour with a possible aversion of the forecasters to produce extreme forecasts. If they see that their own forecast is too far away from the other forecasts (or the Consensus), they start wondering that they are possibly wrong in their view and revise towards the Consensus.

6.1. Methodology

Gallo et al. (2002) test the assumption of imitation behaviour empirically by running the following regression using OLS: ${F_{T,t}}^{i} = \alpha + \beta F_{T,t-1}^{i} + \gamma Cons_{T,t-1} + \delta \sigma_{T,t-1} + u_{T,t}$.

The current forecast of an individual forecaster is regressed on a constant, his own forecast published in the previous month ($F_{T,t-1}^{i}$), the value of the Consensus which was published in the previous month, but which is only known in the current period ($Cons_{T,t-1}$), and the standard deviation of all forecasts in the previous month. This is meant to capture the effect related to the forecasts moving closer together as the time-horizon decreases ($\sigma_{T,t-1}$). A high β indicates a low likelihood that a forecaster changes his mind in subsequent periods. The sign of γ shows whether the movement of the individual is in agreement with the movement of the Consensus. A positive value shows that the forecaster tends to revise his forecasts in the same direction as the rest.

6.2. Results

Table 6.1 shows the results of the regression using OLS applied to the German data used in this paper. Similarly to the results of Gallo et al.(2002), a high explanatory capability can be observed (R² ranging from 0.85 to 0.98). Also confirmed can be a high persistence effect, all coefficients are positive and highly significantly different from zero. The interesting factor regarding the hypothesis of imitation behaviour is theγ. It is apparent that for all forecasters this factor is positive, but they differ highly in their magnitude and significance. The three lowest values come from three U.S. investment banks; two of the values are not significantly different from zero. This does not seem to be a coincidence, as it may be expected that these foreign banks in their forecasts put more emphasis on global factors and less on the views of other German forecasters. Furthermore, relatively low values can be observed for the four largest German banks, whose researchers probably have the highest reputation (Deutsche Bank, Dresdner Bank, Commerzbank, HVB). Meanwhile some other forecasters, often from smaller research departments, reach very high values. Some of them are even of the same magnitude as the values of their own lagged forecasts.

Table 6.1 Test for imitation behaviour

	Beta	Gamma	Delta	R²
Institute 1	0.56*** (0.072)	0.42*** (0.071)	-1.47*** (0.475)	0.91
Institute 2	0.75*** (0.078)	0.25*** (0.076)	0.49 (0.367)	0.98
Institute 3	0.47*** (0.107)	0.58*** (0.109)	-0.52 (0.460)	0.97
Institute 4	0.78*** (0.052)	0.23*** (0.052)	-0.25 (0.337)	0.95
Institute 5	0.74*** (0.066)	0.30*** (0.068)	-0.09 (0.375)	0.97
Großbank 1	0.90*** (0.052)	0.11* (0.060)	-0.36 (0.318)	0.95
Großbank 2	0.80*** (0.058)	0.22*** (0.060)	0.02 (0.263)	0.96
Großbank 3	0.78*** (0.095)	0.23** (0.105)	0.06 (0.707)	0.92
Großbank 4	0.81*** (0.056)	0.20*** (0.055)	-0.06 (0.282)	0.95
Co-operative 1	0.60*** (0.062)	0.41*** (0.061)	-0.08 (0.216)	0.98
Co-operative 2	0.62*** (0.058)	0.40*** (0.060)	-0.22 (0.301)	0.94
Landesbank 1	0.59*** (0.057)	0.39*** (0.051)	-0.47* (0.244)	0.95
Landesbank 2	0.63*** (0.052)	0.40*** (0.052)	-0.83*** (0.267)	0.95
Landesbank 3	0.78*** (0.056)	0.24*** (0.062)	-0.23 (0.301)	0.95
Landesbank 4	0.77*** (0.054)	0.24*** (0.053)	-0.08 (0.280)	0.95
Landesbank 5	0.57*** (0.053)	0.44*** (0.053)	-0.61** (0.243)	0.96
Affiliate 1	0.68*** (0.060)	0.35*** (0.061)	-0.49** (0.206)	0.97
Affiliate 2	0.77*** (0.037)	0.26*** (0.039)	-0.76*** (0.256)	0.97
Affiliate 3	0.87*** (0.053)	0.15** (0.067)	-0.03 (0.380)	0.94
Private 1	0.84*** (0.062)	0.19*** (0.065)	-0.19 (0.262)	0.96
Private 2	0.81*** (0.052)	0.20*** (0.054)	-0.79*** (0.291)	0.96
Private 3	0.63*** (0.058)	0.41*** (0.062)	-0.19 (0.272)	0.96
Private 4	0.71*** (0.046)	0.30*** (0.043)	-0.34 (0.302)	0.94
Foreign 1	0.92*** (0.052)	0.08 (0.056)	-0.44 (0.349)	0.96
Foreign 2	0.82*** (0.063)	0.20*** (0.068)	-1.14** (0.441)	0.95
Foreign 3	0.66*** (0.090)	0.32*** (0.086)	-1.35** (0.565)	0.91

	Beta	Gamma	Delta	R²
Foreign 4	0.90*** (0.061)	0.12* (0.067)	-0.41 (0.506)	0.95
Foreign 5	0.96*** (0.057)	0.03 (0.072)	-0.01 (0.439)	0.95
Foreign 6	0.70*** (0.098)	0.20** (0.088)	0.63 (0.620)	0.84
Foreign 7	0.60*** (0.122)	0.42*** (0.143)	-0.67 (0.859)	0.90
Others 1	0.64*** (0.048)	0.40*** (0.048)	-0.44** (0.202)	0.97
Others 2	0.51*** (0.065)	0.51*** (0.061)	-0.10 (0.320)	0.97
Others 3	0.52*** (0.105)	0.53*** (0.123)	-0.43 (0.511)	0.92

The symbol ***,**, and * denotes rejection of the null hypothesis at the 1%, 5%, and 10% level respectively. Standard errors in parentheses.

Summing up, the results indicate that for almost all of the forecasters the hypothesis of imitation behaviour can not be rejected empirically, but this effect differs highly in its magnitude between the individual forecasters.

7. Test for differences in forecast accuracy

Chapter 3 illustrated the difficulties one has to face when the errors between two individual forecasters are tested for statistically significant differences. As the Diebold-Mariano test only allows for bilateral comparisons for a fixed horizon, it is not an appropriate test to answer the question if the differences in accuracy found with descriptive measures in chapter 2 show statistical significance. In this chapter, an alternative nonparametric test proposed by Stekler (1987) is presented to answer empirically the question whether all forecasters were equal or if some forecasted better than the rest.

7.1. Methodology

A first idea to test for individual differences in accuracy would be to conduct an F-test of the forecast errors across the forecasters, but Batchelor (1990) argues that this would not be legitimate. As was shown in figure 3.2, some years were more difficult to predict than others. If the average forecast errors were used for the analysis, years which were hard to forecast (with a higher variance of the errors) would dominate the results. Therefore, Stekler proposes a rank-sum-test, which constitutes a nonparametric test. This test only looks at the positions of the forecasters in a ranking which is calculated on the basis of the forecast errors for every forecasted year.

In a first step, all forecasters (1,…,n) are ranked for every target year according to their forecast error, with the value of 1 given to the best and the last rank (n) to the worst. Stekler (1987) uses the RMSE over all forecast horizons to rank the forecasters for every target year. But this approach does not seem to be appropriate for this data set given its many missing values as discussed in chapter 2. The usage of the RMSE would favour the forecasters who mainly participated at the short horizons. Therefore, in this paper, the forecasters have been ranked according to their Theil´s U compared to the Consensus as introduced in chapter 2.2. Then, for every forecaster i, the rank sum

$$r_i = \sum_{t=1}^{T} r_{it}$$

over all predicted years (1,…,T) was calculated.

For this procedure the sample has been restricted to the 22 forecasters who took part in the survey in all years between 1995 and 2005. Table A.4 in the annex shows their ranks for Theil´s U calculated for both the next and the current year forecasts (comprising all 24 horizons). It appears that there are big differences in forecast accuracy, with some forecasters consistently performing better than the average and others consistently ranking lowest.

The assumption that forecasters are not equal in their forecast accuracy – i.e. that their positions are not random – was tested empirically. The null hypothesis claims that all forecasters are equal. This means that looking at different periods, there should be no systematic differences in the rank distribution. Thus, for every year the rank of an individual forecaster has to be identical to the average rank, i.e. (n+1)/2. For 22 forecasters this average rank is 11.5. Therefore it is claimed that the null hypothesis is $H_0 : r_i = T(n+1)/2$, such that for every forecaster the rank sum over all target years T should be identical to its expected value.

To test this empirically, Batchelor (1990) developed the following test statistic, which follows a $\chi_{n-1}{}^2$ distribution under the null hypothesis:

$$f = \sum_{i=1}^{n} \frac{\{r_i - T(n+1)/2\}^2}{Tn(n+1)/12}.$$

The nominator shows the squared deviation of each forecaster’s rank sum from its expected value. In the denominator, the variance of an individual rank statistic [n(n+1)/12] for the sum of T individual ranks is used.

7.2. Results

Three test statistics have been calculated: one comprising all 24 forecast horizons, one only for the next year forecasts and one for the current year forecasts. Table 7.1 presents the results of the test statistics. The result of the test statistic for all 24 forecast horizons shows the highest significance: The null hypothesis that the forecast accuracy of all participants is equal is rejected at the 1% level. The values of the current and next year forecasts are significant at the 5% and 10% level respectively.

Table 7.1 Results of the rank sum test

Whole period	Current year	Next year
40.41***	34.85**	29.99*

Critical values for 21 degrees of freedom: 1%: 38.98; 5%: 32.67; 10%; 29.62.
The symbol ***,**, and * denotes rejection of the null hypothesis at the 1%, 5%, and 10% level respectively.

Therefore, it can be concluded that the different positions found in the rankings are not random. Not all forecasters are equal. Some of them performed significantly better than a random distribution of the ranks would have suggested and showed a better forecast accuracy than other forecasters.

8. Conclusions

The empirical analysis of the data from the Consensus Forecasts survey allows us to draw a number of interesting conclusions regarding the performance of German business cycle forecasters in the last 10 years.

The most striking result is the weak forecast accuracy of all forecasters for the next year forecasts, especially relative to a simple naïve forecast. As it has been shown, this can mainly be explained by the large positive bias which all forecasters show for the longer forecast horizons in the analyzed time period. This bias could be confirmed empirically for all forecasters, showing that they started off with a systematic overestimation at the longest forecast horizon which decreased slowly while getting closer to the end of the target year.

Although the comprehensive discussion of the reasons for this bias is beyond the scope of this paper, two aspects seem to be important. The most obvious explanation for the overestimation appears to be the assumption that the decline of the German trend growth rate since the mid-90s was not expected and therefore not incorporated into the forecasts. But this explanation may be dissatisfying because the reasons underlying the decline of potential growth were known by the forecasters for some time before the decline happened. Moreover it should not be neglected that the past 10 years were dominated by a number of unpredictable negative macroeconomic shocks, like the bursting of the new economy bubble or the terrorist attacks of 9/11. A more detailed analysis of these explanatory factors should be undertaken in future research.

Another source of inaccuracy may be associated with weak information efficiency. Although this is much more difficult to confirm empirically than bias, it can be assumed from the analyses that negative news have been incorporated too slowly. This impaired the adjustment of the forecasts from earlier optimistic views to new information. As a final source of inaccuracy an imitation effect could be identified. An imitation of the view of other forecasters can be confirmed empirically for the majority of the forecasters and differs highly in its magnitude.

Notwithstanding these common errors which affected the forecast accuracy of all forecasters, the results of a rank-sum test indicate that they differ significantly in their forecast accuracy. This makes the rankings based on different descriptive measures interesting. Still, it remains an open question whether the best performers, who were mainly less renowned forecasters, reached their relatively better results because of their generally better models and abilities, or if they simply coped better with the challenges of this specific period which was apparently very difficult to predict. This question can only be answered by either extending the data set to earlier years, or keeping an eye on the performance of the forecasters in the next years. Finally, another interesting research would be to analyze whether these results would also hold for other variables published in Consensus Forecasts, e.g. inflation forecasts.

References

Batchelor, Roy A. (1990), All forecasters are equal, Journal of Business and Economic Statistics, 8(1), pp.143-144.

Batchelor, Roy A. (2001), How useful are the forecasts of intergovernmental agencies? The IMF and OECD versus the consensus, Applied Economics, 33(2), pp. 225-235.

Blix, Mårten, Joachim Wadefjord, Ulrika Wienecke and Martin Ådahl (2001), How Good Is the Forecasting Performance of Major Institutions?, Sveriges Riksbank Economic Review, 2001(3), pp. 38-68.

Diebold, Francis X. and Roberto S. Mariano (1995), Comparing Predictive Accuracy, Journal of Business and Economic Statistics, 13(3), pp. 253-264.

Döpke, Jörg (2004), Zur Qualität von Konjunkturprognosen, Wirtschaftswissenschaftliches Studium, 33(1), pp. 8-13.

Döpke, Jörg and Ulrich Fritsche (2005), Growth and Inflation Forecasts for Germany. A Panel-based Assessment of Accuracy and Efficiency, Empirical Economics, 31(3), pp. 777-798.

Fildes, Robert and Herman Stekler (2002), The State of Macroeconomic Forecasting, Journal of Macroeconomics, 24(4), pp. 435-468.

Gallo, Giampiero M., Clive W. J. Granger and Yongil Jeon (2002), Copycats and Common Swings: The impact of the use of forecasts in information sets, IMF Staff Papers, 49(1), pp. 4-21.

Harvey, David I., Stephen J. Leybourne, Paul Newbold (2001), Analysis of a panel of UK macroeconomic forecasts, Econometrics Journal, 4(1), pp. 37-55.

Harvey, David I., Stephen J. Leybourne, Paul Newbold (1997), Testing the Equality of Prediction Mean Squared Errors, International Journal of Forecasting, 13(2), pp. 281 - 291.

Isiklar, Gultekin, Kajal Lahiri and Prakash Loungani (2005), How quickly do forecasters incorporate news? Evidence from cross-country surveys, Journal of Applied Econometrics, forthcoming.

McNees, Stephen K. (1992), The Uses and Abuses of “Consensus” Forecasts, Journal of Forecasting, 11(8), pp.703-710.

Newey, Whitney K. and Kenneth D. West (1987), A simple positive semi-definite heteroskedasticity and autocorrelation-consistent covariance matrix, Econometrica, 55(3), pp. 703-708.

Nordhaus, William D. (1987), Forecasting Efficiency: Concepts and Applications, The Review of Economics and Statistics, 69(4), pp. 667-674.

Schröder, Michael (2002), Erstellung von Prognosemodellen, in: Michael Schröder (ed.), Finanzmarkt-Ökonometrie: Basistechniken, Fortgeschrittene Verfahren, Prognosemodelle, Schäffer-Poeschel, Stuttgart, pp. 397-465.

Stekler, Herman O. (1987), Who forecasts better?, Journal of Business and Economic Statistics, 5(1), pp.155-158.

Appendix

Table A.1 Availability of forecasters

	Group	Availability	Mergers/Acquisitions
DIW	Institute	01/1995 – 12/2005	
HWWA	Institute	02/1996 – 12/2005	
Ifo	Institute	01/1995 – 12/2005	
RWI	Institute	01/1995 – 12/2005	
HWWA	Institute	01/1995 – 12/2005	
FAZ Institut	Others	01/1995 – 11/2005	
Deutsche Bank	Großbank	01/1995 – 12/2005	
Commerzbank	Großbank	01/1995 – 12/2005	
Dresdner Bank	Großbank	01/1995 – 12/2005	
DZ Bank	Co-operative	01/1995 – 12/2005	Until 12/2001: DG Bank
WGZ Bank	Co-operative	01/1995 – 12/2005	
Bayerische Vereinsbank	-	01/1995 – 08/1998	
HYPO Bank	-	01/1995 – 07/1998	
Bankgesellschaft Berlin	Landesbank	01/1995 – 12/2005	
Bayerische Landesbank	Landesbank	01/1995 – 12/2005	
Westdeutsche Landesbank	Landesbank	01/1995 – 12/2005	
DekaBank	Landesbank	01/1995 – 12/2005	Until 12/1998: Deutsche Girozentrale
BfG Bank	Affiliate	01/1995 – 12/2005	Since 04/2001: SEB
BHF-Bank	Affiliate	01/1995 – 12/2005	
Bank Julius Bär	Private	01/1995 – 12/2005	
Delbruck & Co	Private	01/1995 – 04/2003	
Sal Oppenheim	Private	01/1995 – 12/2005	
MM Warburg	Private	01/1995 – 12/2005	
HSBC Trinkaus & Burkhardt	Affiliate	01/1995 – 12/2005	Until 11/1998: Trinkaus & Burkhardt
JP Morgan	Foreign	01/1995 – 12/2005	
Morgan Stanley	Foreign	06/1996 – 12/2005	
Invesco Bank	Foreign	08/1998 – 10/2004	
Merrill Lynch	Foreign	07/1998 – 11/2002	
UBS Warburg	Foreign	05/1998 - 12/2005	Until 04/2000: Warburg Dillon Reed; since 07/2003: UBS
HypoVereinsbank	Großbank	09/1998 – 12/2005	
IW Köln	Others	12/1999 – 12/2005	
Lehman Brothers	Foreign	03/2002 – 12/2005	
UBS Frankfurt	-	01/1995 - 03/1998	
SMH Bank	-	01/1995 - 05/1998	

Bank in Liechtenstein	Foreign	01/1995 – 08/1998	
Hoechst AG	Others	01/1995 – 04/1999	
Economist Intelligence Unit	-	10/2003 - 12/2005	
Goldman Sachs	-	10/2003 - 12/2005	
Bank of America	-	10/2003 – 12/2005	
Global Insight	-	10/2004 – 12/2005	
Citigroup	-	05/2004 – 12/2005	

Table A.2 Descriptive statistics: Current year forecasts

	Forecaster	Theil´s U	RMSE	MAE	Percentage better than Consensus
1	Foreign 1	0.86	0.45	0.34	0.56
2	Private 1	0.89	0.51	0.36	0.55
3	Foreign 2	0.91	0.56	0.37	0.63
4	Affiliate 1	0.92	0.48	0.36	0.52
5	Affiliate 2	0.94	0.50	0.36	0.55
6	Landesbank 1	0.98	0.55	0.40	0.42
7	Affiliate 3	0.98	0.55	0.40	0.46
8	Institute 1	0.99	0.58	0.42	0.45
9	Institute 2	1.00	0.60	0.43	0.47
10	Großbank 1	1.00	0.56	0.40	0.46
11	Consensus	1.00	0.56	0.39	-
12	Institute 3	1.03	0.59	0.40	0.44
13	Großbank 2	1.04	0.58	0.42	0.39
14	Foreign 3	1.04	0.60	0.44	0.35
15	Co-operative 1	1.05	0.60	0.43	0.38
16	Co-operative 2	1.05	0.59	0.41	0.50
17	Private 2	1.06	0.56	0.39	0.56
18	Landesbank 2	1.06	0.59	0.45	0.34
19	Private 3	1.07	0.61	0.43	0.36
20	Landesbank 3	1.07	0.60	0.44	0.45
21	Landesbank 4	1.09	0.61	0.45	0.40
22	Großbank 3	1.09	0.60	0.37	0.51
23	Foreign 4	1.13	0.58	0.42	0.36
24	Landesbank 5	1.14	0.64	0.45	0.28
25	Others 1	1.15	0.65	0.46	0.31
26	Private 4	1.19	0.78	0.58	0.11
27	Others 2	1.21	0.75	0.53	0.18
28	Großbank 4	1.21	0.66	0.47	0.33
29	Institute 4	1.23	0.71	0.54	0.20
30	Foreign 5	1.23	0.62	0.44	0.38
31	Institute 5	1.24	0.69	0.51	0.30
32	naïve	1.90	1.06	0.89	0.22

Table A.3 Descriptive statistics: Next year forecasts

	Forecaster	Theil´s U	RMSE	MAE	Percentage better than Consensus
1	naïve	0.79	1.11	0.89	0.59
2	Foreign 2	0.84	1.30	0.97	0.82
3	Foreign 3	0.88	1.36	1.11	0.68
4	Affiliate 2	0.90	1.26	1.03	0.60
5	Großbank 3	0.93	1.41	1.12	0.53
6	Landesbank 1	0.95	1.31	1.08	0.53
7	Institute 1	0.96	1.38	1.14	0.52
8	Affiliate 1	0.96	1.32	1.05	0.55
9	Private 4	0.97	1.46	1.20	0.52
10	Institute 3	0.98	1.49	1.21	0.59
11	Private 1	0.99	1.40	1.12	0.55
12	Institute 4	1.00	1.39	1.16	0.36
13	Others 1	1.00	1.40	1.13	0.37
14	Consensus	1.00	1.40	1.11	-
15	Landesbank 5	1.00	1.42	1.13	0.46
16	Private 2	1.00	1.37	1.08	0.42
17	Institute 2	1.00	1.43	1.19	0.46
18	Foreign 1	1.01	1.40	1.01	0.59
19	Co-operative 1	1.03	1.42	1.16	0.29
20	Landesbank 2	1.04	1.46	1.16	0.40
21	Co-operative 2	1.06	1.46	1.13	0.47
22	Landesbank 3	1.06	1.49	1.14	0.45
23	Großbank 2	1.06	1.49	1.24	0.29
24	Others 2	1.07	1.80	1.55	0.20
25	Großbank 4	1.07	1.49	1.25	0.23
26	Großbank 1	1.08	1.49	1.11	0.50
27	Affiliate 3	1.09	1.54	1.27	0.22
28	Landesbank 4	1.09	1.53	1.24	0.26
29	Private 3	1.11	1.52	1.23	0.23
30	Institute 5	1.12	1.58	1.32	0.15
31	Foreign 4	1.14	2.04	1.72	0.42
32	Foreign 5	1.17	1.44	1.1	0.30

Table A.4 Test for information efficiency

	Coefficient (-1)	Obs.	Coefficient (-3)
Großbank 1	0.21*** (0.070)	28	0.18 (0.130)
Großbank 2	0.05 (0.074)	9	0.54*** (0.126)
Großbank 4	0.23*** (0.080)	16	0.14 (0.140)
Co-operative 1	0.26*** (0.142)	31	0.56*** (0.083)
Co-operative 2	-0.09 (0.070)	36	0.25** (0.109)
Landesbank 1	0.16** (0.070)	20	0.35*** (0.127)
Landesbank 2	0.18*** (0.067)	26	0.22* (0.119)
Landesbank 3	0.09 (0.072)	52	0.34** (0.132)
Landesbank 4	0.18** (0.069)	14	0.43*** (0.120)
Landesbank 5	0.01 (0.071)	29	0.36*** (0.129)
Affiliate 1	0.35*** (0.073)	32	0.48*** (0.115)
Affiliate 2	0.01 (0.085)	17	0.48*** (0.139)
Affiliate 3	0.23*** (0.073)	25	0.25* (0.131)
Private 1	0.10 (0.083)	11	0.41*** (0.127)
Private 2	0.06 (0.078)	16	0.18 (0.134)
Private 3	0.17** (0.078)	22	0.40*** (0.146)
Private 4	0.26** (0.116)	15	0.29* (0.145)
Foreign 5	0.17** (0.083)	27	0.17 (0.140)
Others 1	0.31*** (0.066)	36	0.43*** (0.138)
Consensus	0.54*** (0.060)		0.63*** (0.110)

The symbol ***,**, and * denotes rejection of the null hypothesis at the 1%, 5%, and 10% level respectively. Standard errors in parentheses.

Table A.5 Rank sum test

	Institute 1	Institute 2	Institute 4	Institute 5	Großbank 1	Großbank 2	Großbank 4	Co-operative 1	Co-operative 2	Landesbank 1	Landesbank 2	Landesbank 3	Landesbank 4	Landesbank 5	Private 1	Private 2	Private 3	Affiliate 1	Affiliate 2	Affiliate 3	Foreign 5	Others 1
1995	1	7	21	17	11	16	5	10	18	2	12	6	13	8	3	15	14	9	4	19	22	20
1996	8	20	14	21	6	17	10	9	16	3	7	5	18	1	2	4	15	12	11	22	19	13
1997	22	4	15	6	2	7	19	11	1	21	8	20	14	16	10	3	17	12	13	9	18	5
1998	6	9	18	20	15	17	21	14	5	1	7	2	16	11	10	13	8	3	4	19	22	12
1999	6	10	9	22	12	17	20	18	16	7	2	5	4	14	11	1	13	3	21	19	8	15
2000	14	10	20	6	2	4	13	19	7	21	18	17	15	16	5	12	1	11	9	22	8	3
2001	2	4	3	15	20	6	19	14	10	5	7	21	11	13	9	16	17	8	1	18	22	12
2002	9	2	16	14	11	5	13	7	22	8	18	20	17	21	12	15	19	4	1	6	3	10
2003	4	13	18	15	17	20	10	14	11	5	21	12	22	9	8	7	16	3	2	1	19	6
2004	14	3	13	15	22	21	18	2	12	1	10	9	16	4	6	11	7	8	20	17	19	5
2005	14	11	5	12	3	18	7	9	4	17	15	2	16	6	10	20	22	8	1	19	21	13
Sum	100	93	152	163	121	148	155	127	122	91	125	119	162	119	86	117	149	81	87	171	181	114
Expected Sum= 126.5																						