CPB Document

No 172
October, 2008

Evaluating CPB's published GDP growth forecasts

A comparison with individual and pooled VAR based forecasts

Adam Elbourne, Henk Kranendonk, Rob Luginbuhl, Bert Smid and Martin Vromans

CPB Netherlands Bureau for Economic Policy Analysis
Van Stolkweg 14
P.O. Box 80510
2508 GM The Hague, the Netherlands

Telephone +31 70 338 33 80
Telefax +31 70 338 33 50
Internet www.cpb.nl

ISBN 978-90-5833-377-3

Abstract in English

We compare the accuracy of our published GDP growth forecasts from our large macro model, SAFFIER, to those produced by VAR based models using both classical and Bayesian estimation techniques. We employ a data driven methodology for selecting variables to include in our VAR models and we find that a randomly selected classical VAR model performs worse in most cases than the Bayesian equivalent, which performs worse than our published forecasts in most cases. However, when we pool forecasts across many VARs we can produce more accurate forecasts than we published. A review of the literature suggests that forecast accuracy is likely irrelevant for the non-forecasting activities the model is used for at CPB because they are fundamentally different activities.

Key words: SEMs, VAR models, Forecast combination, Bayesian methods, Real time

JEL code: C52, C53, E37

Abstract in Dutch

De nauwkeurigheid van de voorspellingen voor de groei van het BBP met het macro-model SAFFIER, gepubliceerd in CEP en MEV, worden vergeleken met de voorspellingen van VAR en BVAR-modellen. De variabelen in de VAR-modellen zijn geselecteerd op basis van hun correlaties met het BBP en een aselect VAR-model geeft in de meeste gevallen slechtere resultaten dan een Bayesiaanse versie. In het algemeen zijn de resultaten van VAR's en BVAR's slechter dan onze gepubliceerde voorspellingen gebaseerd op het SAFFIER-model. 'Pooling' van de voorspellingen van een groot aantal VAR's zorgt voor een vermindering van de voorspelfout en geeft vaak betere resultaten dan voor CEP en MEV. De literatuur suggereert dat het niet aannemelijk is dat de nauwkeurigheid van de voorspellingen relevant is voor de beleidsmatige modelanalyses met SAFFIER omdat het twee fundamenteel verschillende activiteiten zijn.

Sleutelwoorden: Structurele modellen, VAR-modellen, Voorspellingen combineren, Bayesiaanse methoden

JEL code: C52, C53, E37

Een uitgebreide Nederlandse samenvatting is beschikbaar via www.cpb.nl.

Contents

Preface 7

Summary 9

1 Introduction 11

2 Lessons from previous forecasting competitions 15
2.1 Traditional forecasting theory and early forecasting competitions 15
2.2 Findings of recent research 16
2.3 Implications for CPB forecasts 19

3 The competing models 21
3.1 The CEP/MEV process 21
3.2 VAR models 23
3.3 VECM models 24
3.4 Bayesian variants 25

4 Research approach 27
4.1 Model selection 27
4.2 Measuring performance 28
4.3 Testing the relevance of the conclusions of previous studies for the Netherlands 29
4.4 Data 29

5 Results 31
5.1 Comparison with published forecasts 31
5.1.1 Real time forecasts made in March 31
5.1.2 Real time forecasts made in September 33
5.1.3 Conclusion on real time forecasts 35
5.2 Testing the four hypotheses 35
5.2.1 Do simple models do best? 35
5.2.2 Does the accuracy measure matter? 36
5.2.3 Does pooling help? 37
5.2.4 Does the evaluation horizon matter? 39
5.3 Fit versus accuracy 40

6 Conclusion 43

A Data sources 45

B Choice of benchmark 47

C The influence of individual variables on forecast accuracy 49

References 51

Preface

Four times a year CPB publishes macroeconomic forecasts for a two year horizon. These forecasts are made using a large macro model SAFFIER and evaluated regularly by comparing them to the realisations. There are alternative ways to construct macroeconomic forecasts. A popular method is the use of vector-autoregressive (VAR) models. They use little or no theory and instead focus on empirical relationships in the data. The research question is whether we could have made more accurate forecasts by using VAR-models, conditional on information available at the time. The analysis is restricted to forecasts of GDP growth.

This document was written by Adam Elbourne, Henk Kranendonk, Rob Luginbuhl, Bert Smid and Martin Vromans. Adam Elbourne acted as project leader. Paul de Jongh constructed the databases that were used in the project. George Gelauff, Albert van der Horst, Free Huizinga, Debby Lanser, Rocus van Opstal and Johan Verbruggen provided useful comments to earlier versions of this document. Special thanks go to Jan Jacobs of the University of Groningen for his detailed comments.

Coen Teulings
Director CPB

Summary

In this paper we compare the forecast accuracy of VAR based models with that of the macro model SAFFIER. SAFFIER is the model used at CPB for short and medium term analyses including, among other things, producing forecasts for a wide range of macroeconomic variables. This paper focusses on the real time forecast accuracy of our published forecasts of GDP growth for the current and next year are compared with those produced by various classes of VAR models estimated using both classical and Bayesian techniques over the period 1993-2006 (yearly VARs) and 2001-2006 (quarterly VARs). We made this comparison for forecasts made in March (CEP= “Centraal Economisch Plan”) and in September (MEV= “Macro Economische Verkenning”). For this purpose nine variables, chosen on their leading correlations with GDP measures, were selected. We looked at all possible combinations of VAR systems up to five variables as large VAR models are constrained by the number of degrees of freedom. The models are compared regarding the mean error, mean absolute error and root mean square error of the forecasts. Furthermore, we look at the correlation between in-sample fit and forecast accuracy and between forecast accuracy in one period and accuracy in subsequent periods.

The results are discussed in view of four key results regarding forecasting listed by Hendry and Clements (2003): simple, robust forecasting models perform best; pooling forecasts improves accuracy; different measures of accuracy lead to different conclusions; and different methods perform best at different forecast horizons. Recent theory also argues convincingly that forecasting should be seen as distinct from policy analysis - that a model produces more accurate forecasts does not make it more suitable for policy analysis.

The average accuracy of individual VAR based models proves to be worse in most cases than for the published forecasts in the sample period. The main exception is the quarterly forecasts in the current year in the MEV competition and the Bayesian forecasts in the next year in the CEP competition. Pooling the forecasts from the individual models improves accuracy, especially for some classically estimated models. For the CEP forecasts the pooled quarterly VECM forecast beats the published forecast for the current year. For the next year, no pooled VAR based forecast is less accurate than the published forecast except for yearly VAR models in levels, and only when accuracy is measured with squared errors as opposed to absolute errors. For September, the quarterly results of all classes of VAR based models in the current year outperform the published forecast. These results also show that the evaluation horizon matters, at least to some extent.

The finding that simple robust models perform best does not entirely tally with our results. More variables and lags mostly do improve the forecasting accuracy of VECMs, but this does not hold true for all model classes since dVARs do not become more accurate. This is somewhat surprising since the literature argues that VECMs should perform poorly because of their sensitivity to structural breaks. When account is taken of models that display similar accuracy, there is little difference in the ranking of the models whether they are evaluated using the mean absolute error or the root mean square error.

We find no useful correlation between various measures of in-sample fit and forecast accuracy and there is no correlation between forecast accuracy in one period and that in subsequent periods. This means it is not possible to select ‘the best’ model to improve our current forecasting practice. Pooling forecasts get us close to the best performing model anyway, so it is the most relevant finding for improving our GDP growth forecasts in future.

1 Introduction

"*Those who have knowledge, don't predict. Those who predict, don't have knowledge.*"
Lao Tzu

At CPB one of our tasks is to produce forecasts for a wide range of macroeconomic variables for a two year horizon, one of which is GDP growth. These forecasts are made using SAFFIER, a large macro model. The modelling process behind SAFFIER places great emphasis on economic theory: SAFFIER has approximately 2600 equations of which 50 equations represent so-called behavioural equations based on economic theory. The emphasis on economic theory allows forecasts to be made that highlight a broad picture of potential developments in the economy, whilst ensuring that bookkeeping identities are not violated. The theoretically consistent story embodied in a forecast produced with a large macro model is a key demand of many forecast users. CPB has undertaken numerous studies evaluating the accuracy of our forecasts (see Kranendonk and Verbruggen (2006), or Lanser and Kranendonk (2008)) and this study is another attempt to evaluate our published forecasts in light of current practise among academics and other model users.

Since the 1970s, forecasting competitions have shown that atheoretic times series models can often produce more accurate forecasts than large macro models (see Wallis (1989), or Edge et al. (2006), for example).[1] Traditionally this finding would have led to the conclusion that the large macro model was a poor description of the macroeconomy and needed to be respecified: as Clements and Hendry (1998) show, the true model should have the lowest mean squared forecasting error under the assumption that the relationships between variables in the economy are unchanged between the estimation period and the forecast period.[2]

However, recent findings (see Hendry and Clements (2003)) suggest that this is not always the correct conclusion to draw: it might be that the standard assumptions underlying forecasting theory are invalid. In some cases, poor forecasting performance may be due to any number of factors that cause the structure of the economy to change over time. In this case a clear distinction needs to be made between policy analysis and forecasting: a large macro model may be dominated in terms of forecast errors by an atheoretic model that is robust to the types of structural change observed in the period under study, but it may still be the best model for the

[1] As a result of the early competitions, much greater emphasis was placed on the time series properties of large macro models in an attempt to incorporate the forecast accuracy of simple time series models into large macro models. The modelling history at CPB is no exception in this regard.

[2] Consequently, observing that a large macro model produced less accurate forecasts than an atheoretic rival would be seen as evidence that the large model is not an accurate representation of the true structure of the economy. That is, the very theory that allowed a detailed picture of future developments to be put forward was seen as the cause of the inferior forecasting accuracy. Therefore, the greater accuracy available from atheoretic time series forecasts has had to be offset against the lack of story that accompanies them. Often it is this consideration of different factors and the story that accompanies them that produces the clearest picture of the likely prospects for the macroeconomy, from the point of view of the end user of the forecast (see Burns (1986), or Smith (1998), for an example of this argument).

analysis of a given policy issue. To further illustrate this point, Hendry and Clements also report that many authors find that models with good in-sample fit statistics produce no more accurate forecasts than less well fitting models. Moreover, atheoretic models cannot perform the sort of policy analyses that SAFFIER does. The conclusion we should draw from observing more accurate forecasts from atheoretic models is that it may be possible for us to improve the accuracy of the forecasts we publish.

In this paper we compare the published real time forecasts of CPB with those produced by various classes of VAR models using both classical and Bayesian estimation techniques. VAR models were chosen as the competitors because they are atheoretic reduced forms and are commonly used as the benchmark model for producing quick and easy forecasts. VAR models have their roots in the critique of Sims (1980), which mirrors the traditional conclusion sketched above in some respects. Sims argued that many of the restrictions used in large macro models were not valid, in fact he referred to the reliance of large macro models on uncertain theory as 'incredible restrictions'. He proposed that simple VAR models be used in their place since these follow a largely data-driven modelling process and are not as susceptible to the incredible restrictions critique. Furthermore, the relatively small size of VAR models allows them to be estimated as a system. Since VAR models are largely data driven and are relatively simple models to handle, they have been widely used for forecasting. It is therefore of interest to see how a data driven VAR performs relative to SAFFIER for forecasting. It is also straightforward to incorporate leading indicator variables, such as business confidence surveys, alongside traditional economic variables in forecasting VAR models, which may also help in producing accurate forecasts.

The published CPB forecasts are not purely based from SAFFIER because the preliminary model outcomes are regularly adjusted by expert opinion. From a CPB point of view the relevant comparison is not between the unadjusted forecasts of SAFFIER and VAR models because expert opinion and add factors make up an integral part of our forecasting process and we would never consider using the pure model-based forecasts. So, the real question is can VAR models improve our forecast accuracy. That is, are VAR model forecasts more accurate than our published forecasts? Moreover, in Franses et al. (2007) the effect of adjustment for forecast accuracy turned out to be small with the exception of some price variables - for the volume of GDP the forecast accuracy of both the model-based forecast and the published forecast are virtually identical.[3]

Hendry and Clements (2003) list four key findings from the recent literature on forecasting:

- Simple, robust forecasting models perform best
- Pooling forecasts improves accuracy

[3] The mean error of the published forecast is slightly higher, but this is partly caused by preliminary quarterly GDP figures which are revised upwards afterwards.

- Different measures of accuracy lead to different conclusions
- Different methods perform best at different forecast horizons

Given the recent findings in the literature, our key research questions are:

1. Could we have made more accurate forecasts than we did, conditional on information available at the time?
2. Are these four key findings applicable for forecasting Dutch GDP growth?
3. Can we use in-sample measures of fit to pick good forecasting models?

The remainder of this paper proceeds as follows. Section 2 briefly reviews the literature. Section 3 details the forecasting process at CPB and introduces VAR models. Section 4 describes our approach and details our attempt to hold a fair contest. Section 5 describes our results and Section 6 concludes.

2 Lessons from previous forecasting competitions

2.1 Traditional forecasting theory and early forecasting competitions

What is the best way to construct a forecast for GDP growth? As described by Hendry and Clements (2003), the standard theory of forecasting relies on the following two assumptions:

1. The model is a good representation of the economy
2. The structure of the economy will remain relatively unchanged

If these two assumptions are met, then a number of results can be proven. For our immediate purposes, the two most important of these are that the best model from the estimation period will produce the most accurate forecasts and that pooling forecasts from different models cannot improve forecast accuracy. The intuition behind the first point is that the best in-sample model has the most accurate representation of the true economy and, therefore, the most accurate representation of the causes and consequences of economic events. In other words, it has the best description of what will happen tomorrow given a particular economic situation today. The fruitlessness of pooling forecasts follows immediately from this point: the average of the best models and any other model is not as good as the best model – otherwise it wouldn't have been the best model to start with. These results (and a number of others) are proven in Clements and Hendry (1998).

These points are relevant here because SAFFIER is intended to be an accurate approximation to the true economy of the Netherlands for the period it was estimated on. Hence, if SAFFIER is a good representation and the structure of the economy has remained relatively unchanged, then SAFFIER should produce the best forecasts possible.[4]

There is a large literature detailing forecasting competitions between different models. Originally, these competitions were intended to test the first assumption. Upon finding that a given macro model did not produce the best forecasts, it was often concluded that the macro model in question was therefore not the best representation of the economy. One of the first studies to find that large macro models produced poor forecasts was Wallis (1989). Since the large macro models were often beaten by simple univariate time series models (which could be employed by people with no knowledge of economic theory) it was concluded that economic theory in the large models was being outperformed by models which concentrated on the time

[4] SAFFIER describes the structure of the Dutch economy, taking certain foreign and some domestic variables as exogenously given. Hence, the forecasts published by CPB do not come solely from SAFFIER; if these exogenous variables are inaccurate or poorly forecast, then the published forecasts will also be detrimentally affected. The same story applies to adjustments made on expert opinion and the use of add factors (see, Franses et al. (2007)). However, for ease of notation we will use SAFFIER as short-hand for the entire modelling process, including the construction of these exogenous series and adjustment made on the basis of expert opinion. See Section 3.1 for a more detailed discussion of the process of making a forecast with SAFFIER.

series properties of the data. For further examples of these type of competitions see Theil (1966); Mincer and Zarnowitz (1969); Dhrymes et al. (1972); Cooper and Nelson (1975).

Later generations of large macro models were adapted to better take into account the time series properties of the data (along with other developments). When models for the UK were compared by Wallis (1989) for forecast performance in the 1980s, he found that the published forecasts were more accurate than those from simple time series methods. Further results for the UK can be found in Holden (1997). Holden concludes that, whilst large macro models produced the most accurate forecasts, Vector Autoregressions estimated using Bayesian methods can improve forecasts when included in the average. He also found that averaging across published forecasts could improve forecast accuracy, implying that none of the individual models under consideration were the best, under the traditional interpretation.

2.2 Findings of recent research

Recent research also suggests that large macro models will not produce the most accurate forecasts in all situations. For example, both Eitrheim et al. (1999) and Edge et al. (2006) report that simple reduced form time series methods can produce more accurate forecasts, at least for some variables some of the time. As such, there is still no definitive answer to the question of how to construct the best forecast. Recent research has tried to summarise the findings of numerous forecasting competitions, though. For example, Hendry and Clements (2003) draw the following conclusions based on many forecasting competitions, including the so-called M competitions (see Makridakis et al. (1982, 1993) and Makridakis and Hibon (2000)):

1. Simple methods do best
2. The accuracy measure matters
3. Pooling helps
4. The evaluation horizon matters

The M competitions were forecasting competitions involving many different time-series methods, each of which was applied by a recognised expert in using that model. The methods employed varied from statistically driven procedures through commercial forecasting software to expert opinion. Many of these methods require expert knowledge to use effectively. One class of model which does not require extensive expert knowledge is the class of Vector Autoregression (VAR); many institutions use the VAR as the workhorse model for short-term forecasting (Elliott and Timmermann, 2007). Linear univariate autoregressions and VAR models have also performed well in various comparisons. For example Stock and Watson (1998) find that linear autoregressions perform better than nonlinear models for a wide range of US macroeconomic series. For VAR models, Boero (1990) finds that VAR models outperform structural equations models for Italy. In a forecasting comparison for Norway, Eitrheim et al. (1999) found that a first

difference VAR could produce more accurate forecasts in some cases than the large macro model used by the central bank of Norway. Another recent comparison of VAR based forecasts and published forecasts based on large macro models is reported in Edge et al. (2006). They find that, for certain macro variables, VAR based forecasts outperform the published forecasts from the Federal Reserve.

In light of the results from these forecasting competitions, Hendry and Clements have argued that the main problem with forecasts from the large models lies not in whether they are a good representation of the economy in the period for which they were estimated, rather assumption 2 is not met – the future is not always the same as the past. Since it is difficult to beat simple time series methods, Hendry and Clements (2003) propose two alternative assumptions upon which forecasting models should be built:

1. Models are simplified representations which are incorrect in many ways
2. Economies both evolve and suddenly shift

Hendry and Clements argue that the second point is the main reason why economic forecasts perform badly in given periods. They argue that sudden shifts in the deterministic components of models that lead to poor forecasting performance and that these are relatively common. This is why users of large macro models often find it useful to adjust the intercept terms of their models when making forecasts.

One potential reason why simple methods are hard to beat is that macroeconomics is limited by relatively short sample periods. Hence, as Robertson and Tallman (1999) note, there is a trade-off between the precision with which one can estimate parameters and the complexity of a model. A univariate AR(1) model for GDP is without doubt misspecified in many ways, which implies that the forecasts from such a model will be biased; the advantage of such a simple model is that one can get a relatively precise estimate of the autoregressive parameter, however. Given the typical sample size available in macroeconomics, the same is not typically true of larger models with more parameters that need to be estimated. In larger models, especially VAR models, the extra parameters may be estimated imprecisely. This can lead to poor forecasts. In other words there is potentially a precision-bias trade-off. Moreover, in a changing world, the more complex a model is the more possible sources of structural change are present in the model. In comparison, certain types of simple model are robust to certain types of structural break, for example, Eitrheim et al. (1999) detail how VARs in first differences are robust to level shifts. This type of robustness is another potential explanation for the performance observed from simple models.

The discussion in Robertson and Tallman already suggests that parsimony may be more important than model fit. In fact, evidence in the literature suggests that in-sample fit may be almost entirely uninformative when it comes to forecast performance. Fildes and Makridakis

(1995) conclude that there is little, if any, correlation between measures of in-sample fit and out-of-sample forecast accuracy. Typical figures for correlations are 0.2 (which implies that only 4% of forecast accuracy is explained by in-sample fit) for very short-run forecasts of up to 3 periods. The correlation drops rapidly to zero thereafter. It is possible that there is no relationship with univariate models because extra parameters improve recorded fit statistics by overfitting. That is, the extra parameters allow the model to 'account' for movements in the data that were really caused by omitted variables. In a multivariate setting it would be, in principle, possible to model the extra variables and deliver a relationship between in-sample fit and forecast accuracy.

One further potential advantage of simple time series methods is that it is relatively easy to incorporate expectations of individuals into the forecasts through the use of leading indicators. Leading indicators, such as surveys of firms' expectations of future sales or consumer confidence surveys, offer the possibility of directly incorporating the expectations of economic agents into forecasting models. Furthermore, leading indicator variables are often available with very little delay compared to official statistics and are not subject to revision, hence they may contain more up-to-date and relevant information for making forecasts. They don't typically enter large macro models but are easy to incorporate into simple models. One potential drawback of leading indicator variables, however, is that it is unclear if the relationship between these series are more or less susceptible to structural shifts than standard economic series. Hence, their worth in forecasting is an open question.

Whilst VAR models are the forecasting workhorse due to their easy estimation, there are other data driven approaches available. One such alternative is the use of dynamic factor models. Dynamic factor models attempt to summarise the key factors that are driving a large number of time series. Since macroeconomic time series tend to move together, it makes sense to try to model the factors that are driving these common movements. An example of the application of such a model to forecasting Dutch GDP is to be found in Den Reijer (2005). He uses 370 time series to predict GDP up to 8 quarters ahead. He concludes that the mean square error of the dynamic factor model is 70% of that of a univariate AR model for one quarter ahead forecasts. For 2-4 quarters ahead this rises to just over 80% and for 5-8 quarters ahead the relative accuracy is 86% to 98%. Hendry and Clements also argue that pooling improves accuracy, in part because it is a simple way of utilising information from many sources. Given that a dynamic factor model takes a number of series and attempts to extract the information content of the different series it will be informative to compare the accuracy of our forecasts, where we pool over a large number of VAR based models, to those from the dynamic factor model.

2.3 Implications for CPB forecasts

As described in more detail in Section 3, CPB forecasts are made using a large macro model. The findings described above suggest that we may be able to produce more accurate forecasts with the help of simple, robust forecasting models. It is important to stress that this does not mean that the model is 'bad' or 'wrong', especially for answering policy questions. As mentioned at the start of this section, traditional forecasting theory tells us that the best model of an economy should produce the best forecasts. However, there are many reasons, such as structural shifts, why an otherwise good model may produce poor forecasts. The effects of policy interventions may be invariant to structural shifts because the structural shift only changes the value of the intercept term, not the co-variance between the variables. Hence, Hendry and Clements argue:

1. Being the 'best' forecasting model does not justify its policy use
2. Forecast failure is insufficient to reject a policy model.

They justify the first conclusion by noting that "... the class of models that 'wins' forecasting competitions is usually badly mis-specified in econometric terms, and rarely has any implications for economic-policy analysis, lacking both target variables and policy instruments." The second is evident from the observation that "... intercept corrections could improve forecast performance without changing policy advice". There is no reason to doubt that endemic structural change and shifts, as observed in the UK and US, are also relevant for the Netherlands. There are many potential breaks for the Netherlands in the last decade alone: EMU; the dotcom boom; 11 September or the housing boom, to name but a few. It follows that inferior forecast accuracy from CPB models need have no relevance for the policy analysis we perform with these models.

Since VAR and BVAR models are part of the standard forecasting toolkit and have shown the potential to outperform large macro models in terms of forecast accuracy, it seems natural to choose these as the alternative benchmark to test our forecasts against. In light of the findings discussed above we want to see if these are also applicable to the Netherlands. Our key research questions are

1. Could we have made more accurate forecasts than we did, conditional on knowledge we had at the time?
2. Do the 4 conclusions listed above hold for forecasting Dutch GDP growth?
3. Is there any relationship between in-sample measures of fit and out of sample forecast accuracy in our multivariate models?

3 The competing models

3.1 The CEP/MEV process

CPB has a long tradition of using large macroeconomic models to make forecasts and analyses for the Dutch economy. Short-term forecasts are made four times a year. In March (“Centraal Economisch Plan”) and September (“Macro Economische Verkenning”) detailed forecasts are published for the current and the next year. In June and December less detailed forecasts are published. These are effectively an update of the previous forecasts applying recent information on economic developments and government policy.

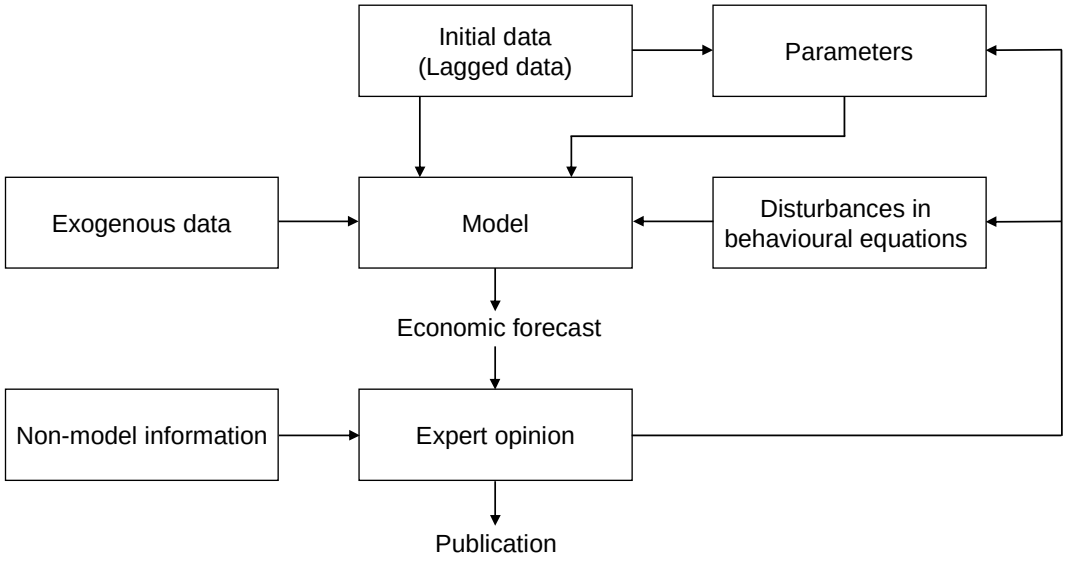
Since 2004 CPB has used the macroeconomic model SAFFIER (see Kranendonk and Verbruggen (2007) for details) for short-term and medium-term macroeconomic analyses. Prior to this the predecessors of SAFFIER were used; SAFE was used between 2002 and 2004 (see CPB (2003)) and FKSEC was used prior to SAFE (see CPB (1992)). The core of SAFFIER are three blocks for the market for goods and services, for the labour market and for the public sector. The first block contains behavioural equations for the final demand components: private consumption, business investments and exports. Part of the demand for goods and services originates from abroad, the remainder is produced in the Netherlands. This production is described on the basis of a constant elasticity of substitution (CES) production function, with labour and capital as the production factors. From this production function the equations for investments and the employment are derived. In the labour market block labour supply is largely exogenous, while the explanation of wages is based on a right-to-manage model. The public sector block consists of a detailed description of all kinds of institutional arrangements in the social security, health-care and tax systems. In addition to this ‘economic’ block SAFFIER has a large book-keeping system, in line with the system of the National Accounts, to guarantee the consistency of all the economic relations.

In numbers, SAFFIER has approximately 2600 equations of which 50 equations represent so-called behavioural equations. The behavioural equations contain about 300 parameters. The remaining equations are rules of thumb or identities. In total, SAFFIER contains over 3000 variables categorised into 2600 endogenous variables and 450 exogenous variables, 200 of which are autonomous terms used for adjusting the forecasts in light of expert opinion.

Figure 3.1 outlines the various components of the forecasting process. Its main component is the macroeconomic model describing the relationship between the endogenous and exogenous variables. The exogenous variables are used as inputs upon which the forecasts are conditioned. In addition, they are also important in defining add factors (adjustments to the constant terms) to the behavioural equations based upon non-model information from expert opinion or leading indicators, for example. Besides this input data, the model requires lagged endogenous data to initialise the forecasting process. This data consists of realised historical values of the various

macroeconomic variables. Furthermore, each behavioural model equation contains several parameters. Incorporation of the above components closes the model, so that a first macroeconomic forecast can be extracted. This first forecast is assessed by several experts within CPB. These experts can suggest adjustments to the results by bringing in non-model information. The experts often rely on their own models which are likely to be better equipped in predicting specific macroeconomic variables, such as social-security or pension-related variables. The non-model information is fed back into the model via the disturbance terms and sometimes via parameter adjustments. Several forecast rounds follow (usually three) resulting in the final forecast publication.

Figure 3.1 The process of producing a forecast with SAFFIER



CPB regularly analyses the forecast accuracy of their short-term forecasts (see Kranendonk and Verbruggen (2006), for example). On average the forecast error is close to zero and tests do not reject that the forecasts are unbiased and efficient. However this result is the balance of significant positive and negative forecast errors in separate years. The size of these errors is declining when more information becomes available, although not very much. Over the sample period studied in this paper, the mean absolute error (MAE) for the forecast in March for the current year is 0.98%, while the MAE for the next year is slightly higher at 1.19%.[5]

The main sources of uncertainty and forecast errors in our published forecasts is given in table 3.1 (from Lanser and Kranendonk (2008)). The scheme in figure 3.1 clearly illustrates the relevant elements that influence the forecasts. Seventy-five percent of the forecast errors can be attributed to the exogenous variables. Assumptions about the international business cycle, like

[5] These figures vary somewhat depending on the specific sample period chosen for the analysis. When the analysis is done over the period 1990-2006 the figures are 0.9% and 1.1% respectively.

Table 3.1 Sources of uncertainty for short-term GDP forecasts made in March (% of total uncertainty)[a]

	Next year	Current year
Exogenous information	78	73
Lagged variables	11	15
Model (parameters)	8	7
Residuals equations	4	6

[a] Numbers do not add up to 100 due to rounding

world trade, competitive prices and interest rates, are crucial for the forecast performance. The second important source of errors is the accuracy of information about the past. Statistics Netherlands revises the yearly data in the National Accounts twice before the figures are 'definitive'.[6] A third source of errors is connected to the macromodel we use: uncertainty about the estimated parameters in the model and the residuals for the behavioural equations.

3.2 VAR models

VAR models became popular econometric tools after Sims (1980) suggested that they could be used as alternatives to large simultaneous equations models. A reduced form pth order VAR is shown in (3.1).

$$Y_t = A_1 Y_{t-1} + \ldots + A_p Y_{t-p} + u_t \tag{3.1}$$

where Y_t is a vector of endogenous variables at time t, A_i are square matrices of parameters and u_t are the reduced form errors. We also include a constant and a trend. In order to facilitate testing the link between in-sample fit and forecast accuracy, we do not use any procedures for selecting the lag length. Rather we estimate all models with four different lag structures: that is, with orders 1 through 4. A VAR is typically estimated by ordinary least squares (OLS) as a reduced form. A number of results justify the use of OLS. Most importantly, the OLS estimates of the autoregressive parameters are consistent and asymptotically normally distributed (see Lütkepohl (1991)), even if the VAR contains integrated variables (see Sims et al. (1990)). There is no distinction made between endogenous and exogenous variables. In order to produce a forecast the VAR model is simply simulated one period ahead to produce the forecast for the next period $\hat{Y}_{t+1}$, as shown in (3.2).

$$\hat{Y}_{t+1} = A_1 Y_t + \ldots + A_p Y_{t-p+1} \tag{3.2}$$

For succesive forecast horizons the procedure is simply repeated. We also include VAR models specified in first differences. This is because using first differences, whilst removing the

[6] This does not include the revisions in the preliminary quarterly figures published during the year on the economic growth.

information regarding the long-run behaviour of the level of the series, helps to make the models robust to level shifts to some degree, as discussed in Section 2. In short-term forecasting the latter robustness may be more important than the lost information from the levels. We call these models dVARs in our notation. The dVAR(p) model is shown in (3.3), where Δ indicates the first difference operator, $Y_t - Y_{t-1}$. A constant is included, which is the equivalent treatment of trends as for the models estimated in levels since a constant in a first difference model implies a trend in the levels specification.

$$\Delta Y_t = A_1 \Delta Y_{t-1} + ... + A_p \Delta Y_{t-p} + u_t \tag{3.3}$$

3.3 VECM models

If cointegrating relations are present in a system of variables, estimating a VECM may be more appropriate. Considering specific parameterisations that support the analysis of the cointegration structure is then useful. The VECM is obtained from the levels VAR form in the previous paragraph by substracting Y_{t-1} from both sides and rearranging terms. This results in the VECM representation shown in (3.4)

$$\Delta Y_t = \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + ... + \Gamma_{p-1} \Delta Y_{t-p+1} + u_t \tag{3.4}$$

where Π and Γ_i are the square matrices of parameters. On the right hand side, the first term represents the long run and the other coefficients are short-run parameters. As with the dVARs, a constant is also included, which implies that there is a trend in the level of the series. This is equivalent to the treatment in the VAR models. again we estimate the models to be as comparable to the basic VARs as possible; so we estimate with the same 4 lag structures as above. For the VECM case, this means that there are zero to three lagged difference terms on the right hand side of (3.4).Whilst the simple act of subtracting Y_{t-1} from both sides of the VAR representation shows that the two models are equivalent, there is a key difference with regards to estimation. The presence of cointegration implies that Π is not of full rank. This has implications for the long-run properties of the model. Namely, the rank of Π, which is the same as the number of cointegrating relationships among the variables, determines how many structural shocks have permanent effects, and conversely, how many only have transitory effects (see King et al. (1991), for details).

We estimate our VECMs using Johansen's technique (Johansen, 1995), which is effectively a two-step procedure. The first step involves estimating the number of long-run relationships present between the series in question, the second involves estimating the parameters of the model conditional on the outcome of the first step. Rather than estimating the number of cointegrating relationships for each model we simply set this equal to one for all models, then

estimate the cointegrating relationships by maximum likelihood.[7]

In all other respects the models are left unrestricted. We purposefully ignore issues related to the (weak) exogeneity of series within VECMs,[8] due to the significant effect that such restrictions can have on the properties of the model (see Jacobs and Wallis (2007) for a discussion of these issues). Exogeneity tests are standard zero restriction tests - the null hypothesis is that the parameter in question is zero and is rejected if the estimate is less likely under the null hypothesis than a pre-selected critical value. However, it is not valid to reverse this process - if a null hypothesis is not rejected it does not imply that it is true, just that it is not rejected. A data driven method for imposing exogeneity would necessarily be based on this reverse of the standard hypothesis test. Minimising the chance of imposing the null hypothesis incorrectly would require that the maximum likelihood estimate of a particular parameter be close to zero, which would limit the effect of the restriction anyway. So we leave our models unrestricted.

By placing greater emphasis on producing a good estimate of Π, a VECM model is placing more emphasis on the long-run properties of the model. Whether this improves short-run forecasts is an open question. As discussed in Section 2 there is some debate in the literature as to the benefits of forecasting with VECMs. The mechanics of forecasting in a VECM are the same as forecasting with a VAR.

3.4 Bayesian variants

It is also possible to estimate the VAR model presented above by Bayesian methods rather than OLS. This proceeds by specifying a prior distribution for each of the A_i matrices, which is incorporated into the estimation using Bayesian methods. One widely employed prior distribution for VAR models is the so-called Minnnesota prior of Litterman (1980, 1986). This prior specifies that the mean of A_1 is the identity matrix and the mean of the remaining A_i matrices is the null matrix. That is, the prior mean is that each series follows a random walk unrelated to the other series in the VAR model. Cross-correlation is allowed if there is sufficient evidence for it in the data to outweigh the effects of the prior. The variance of the prior distributions becomes smaller at greater lags, which implies that more recent events are more important for forecasting future events. In this sense, lag length choice is of much less significance for BVARs than for VARs because the prior weights higher lags heavily towards zero. In our BVARs we use 4 lags to make the models more comparable to the classical VARs

[7] We also estimated the models with more cointegrating relationships but these provided slightly less accurate forecasts.

[8] Exogeneity of a given series implies that this variable does not respond to any of the other variables in the model. Weak exogeneity implies that the series does not respond to deviations from the long-run relationship, but it does to the remaining lagged difference terms.

and VECMs.[9] The BVARs are estimated using the mixed estimation method of Theil and Goldberger (1961).[10] Again, once the model is estimated the forecast is produced in an identical way to VAR or VECM forecasts. For further discussion of BVARs and various prior distributions see Robertson and Tallman (1999).

We also estimate VECM models using Bayesian methods. The cointegrating relationships are estimated using Johansen's maximum likelihood technique with one cointegrating relationship, whilst the remaining parameters are estimated using Theil-Goldberger mixed estimation with an equivalent prior to the Minnesota prior used for the BVARs.

[9] Increasing the number of lags made the yearly forecasts slightly worse and the quarterly forecasts slightly better.

[10] We also estimated the Bayesian models using Gibbs sampling, which gave similar results to the Theil-Goldberger method. In some cases the Gibbs sampling forecasts were slightly worse than the Theil-Goldberger forecasts. Gibbs sampling BVARs and Theil-Goldberger BVARs should produce similar results if the data satisfy the Gauss-Markov assumptions: zero mean, serially uncorrelated and homoskedastic (see LeSage (1999)).

4 Research approach

4.1 Model selection

Since the literature suggests that simple models should produce forecasts that are hard to beat, our point of departure for the VAR models is a simple univariate AR(1) in the yearly growth rate. This is the simplest VAR model for the growth rate of GDP. We then compare such a simple model with the published CEP/MEV forecasts. Then we made the models progressively more complicated by adding lags and variables. In total we selected nine additional variables to include alongside GDP in our models. These nine series were selected based upon their leading correlations with GDP growth in the period ending 1992.[11] The nine variables chosen are listed below (see appendix for a detailed data description).

- Consumption
- Total worker compensation
- Consumer price index
- World trade
- Short term interest rates
- Business climate survey
- Consumer confidence
- Bankruptcies
- German business confidence (the Ifo survey)

All levels series enter the models in logarithms. By stopping in 1992, we ensure that we do not give our VAR based models an unfair advantage compared to our published forecasts.[12] Since VAR models are limited in terms of the number of degrees of freedom available and because theory tells us that there is likely to be a precision-bias trade-off, we estimate all possible combinations of lower dimension models rather than a 10 variable model. We vary the lag length of our models from 1 to 4. In addition to the 4 univariate models we have estimated 1020 versions of each classically estimated model class (there are 9 bivariate combinations, 36 trivariate, 84 combinations of 4 variables and 126 combinations of 5 variables; each is estimated with four different lag structures), except for the yearly models where degrees of freedom limitations restricted 4 variable models to a maximum of 3 lags and 5 variable models to a

[11] We chose nine as the number of series for a number of reasons, the two most important being that nine was computationally feasible and that this allowed us to cover a wide variety of types of variables.

[12] One potentially important distinction between the VAR based models and SAFFIER is that the VAR based models make no use of information we already know with a reasonably high degree of certainty for the forecast period. One of the most important of these is that, at the time a forecast is made with SAFFIER, current and future wage growth in many industries is already known, due to the existence of multiyear wage bargaining. Ceteris paribus this entails an advantage for SAFFIER.

maximum of 2 lags. In total, therefore, there are 520 combinations in each yearly model class. There are also 256 versions of each Bayesian model class (this is the 255 combinations of variables plus the univariate case but without multiplying for different lag structures since the same lag structure of four lags is used for all BVARs).

4.2 Measuring performance

Our analysis of the comparative forecasting performance is based solely on the measure of forecast performance with real time data.[13] Using the latest available data the forecasting models are estimated with their estimation period ending at the end of 1992. Forecasts for 1993 and 1994 are then made. This is similar to what would have been done for the CEP publication in 1993, since provisional data for the whole of 1992 would have been published prior to the CEP forecasts being published in March. Then the process is repeated but with the end of the estimation period shifted one year later. That is, the estimation ends in 1993 and forecasts are made for 1994 and 1995. This is repeated until the last forecasts are made for 2006 and 2007. During this process the start period for the estimation is held constant, so subsequent forecasts use more information. We also make forecasts for comparison to MEV, which is published every September.

For yearly data we have real time data sets from 1993 up to 2006, but for quarterly we only have real time data from 2001 to 2006. The forecasts are evaluated against a series of realisations appropriate for the data set in question, not the latest figures. This is because methodological changes have taken place and some elements of the series are measured differently to what they were in the past – we decided that our analysis should proceed by using realisations and forecasts that were methodologically consistent.[14] (See Appendix B for a discussion of the appropriate benchmark). The relatively short span of real time data available for quarterly models sometimes necessitates the use of the recursive approach alongside the real time approach. When this is the case for the reported statistics it is made clear in the text.

Since we are not the final users of our forecasts, it is not clear which loss function should be applied to judge forecast accuracy. We pick commonly used measures: we compare the mean error, the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) for each of the methods. We base our answer to our first research question on the performance on these measures. Furthermore, these measures have been used in previous accuracy studies that CPB has undertaken and their use here facilitates ease of comparison with previous results. Since we are also interested in distilling differences between the competing methods we report the results per class of models.[15]

[13] In this study we only forecast GDP growth; in CEP and MEV a much broader range of macroeconomic variables are forecasted.

[14] The methodologically consistent realisations can be found in Kranendonk and Verbruggen (2006)

[15] All results are available on request.

4.3 Testing the relevance of the conclusions of previous studies for the Netherlands

In order to test the four conclusions listed in Section 2.2 we test the implications of the conclusions as they apply to our modelling set-up. Firstly, the conclusion that simple/robust methods do best implies that our simplest model, and AR(1) of the growth rate of GDP should perform well. As more lags and more variables are added the models are becoming more complicated and have more relationships that may be subject to structural shifts. Therefore, adding lags and variables should not produce more accurate forecasts. Furthermore, models in levels should not perform better than models specified in growth rates, since the latter are robust against structural shifts. This last implication should be especially relevant for VECM models due to the extra emphasis placed on estimating the long-run relationship between the levels of variables in VECM models. Secondly, if the accuracy measure matters, we should observe differences in our ranking accross our three measures of forecast accuracy: mean error, mean absolute error, and root mean squared error. Thirdly, if pooling helps, we should observe that pooled forecasts are more accurate than the average accuracy of the underlying models. Moreover, the improved accuracy should bring the pooled forecasts into the group of most accurate forecasts. Fourth, we should observe different models, or different classes of models, performing best at different horizons. In our study we effectively have four different horizons: two for CEP and two for MEV.

To answer our last research question regarding the relationship between in-sample fit and forecast accuracy we look at the correlation between various measures of in-sample fit and our forecast accuracy statistics on a model-by-model basis. The fit statistics we choose to look at are the log-likelihood, the Akaike Information Criteria, the Schwartz Bayesian Criteria, r^2 and adjusted r^2. We also look for a relationship between the Quandt-Andrews (see Andrews (1993)) measures of within sample structural break and forecast accuracy.

4.4 Data

The yearly data are taken from the appendices published in ‘Centraal Economisch Plan’, the spring-forecast of CPB. The table ‘Main Economic Indicators’ is available in electronic format since 1993. This table contains the assumptions of the economic international environment and the forecasts for the Dutch economy. The time series start in 1970. The 2007 version is used for the recursive estimations and forecasts. The real-time analysis is based on all available versions since 1993.

The quarterly time series databases from Statistics Netherlands (CBS) are available, for the series we have selected, for a first forecast year of 2001. These databases are limited to Dutch GDP and its main components and do not contain quarterly information on international data or the Dutch labour market. These databases start in the first quarter of 1977.

5 Results

5.1 Comparison with published forecasts

Since our competition contains relatively few comparison points,[16] it would be highly unlikely that at least a few of the models we have run did not produce more accurate forecasts just out of luck. Since there is no way of adequetely distinguishing luck from some underlying reason, it would be foolish to simply pick the best performing model and claim it would remain so. As a result, we focus our discussion on averages since this gives us an idea of how well we could have done if we did not know which model would do best beforehand. In other words, how well could we have expected to have done if we had randomly picked a VAR model to use instead of SAFFIER back in 1993. It turns out that we find very little correlation between the relative rankings of the models over time, so it may be that a search for the best performing VAR model is even more pointless than the discussion here would suggest. See section 5.3 for more information.

5.1.1 Real time forecasts made in March

Table 5.1 shows a comparison between the average accuracy, the accuracy of pooled forecasts and the accuracy of our published forecasts. Those model classes that were more accurate than SAFFIER for MAE or RMSE are shown in italics. For the current year, the average accuracy of yearly VARs and VECMs compares unfavourably with the accuracy of forecasts for SAFFIER for both MAE and RMSE. For quarterly models, both classical and Bayesian dVARs and VECMs have lower MAE but higher RMSE. In fact, none of the VAR based model classes is more accurate than SAFFIER on average when using RMSE. For forecasts for the following year, the comparison is less favourable to SAFFIER. An average yearly Bayesian VAR or dVAR has lower MAE and RMSE than SAFFIER, whilst all quarterly models have lower RMSE than SAFFIER. All in all, however, the performance of our published forecasts is relatively good - even at those forecast horizons where VAR based models are more accurate, the margin is not large.

If we compare the different classes of VAR based models we can see that an average Bayesain model is, in general, slightly more accurate than its classical counterpart, especially for the yearly models. This is evidence that the use of prior information can alleviate the degrees of freedom problem associated with the yearly models to some extent. BVARs are widely used in

[16] For yearly models there are 14 current year comparison points and 13 for the next year. For quarterly data there are 6 and 5 respectively. The small number of comparison points for quaterly models makes inference difficult. However, we also used a single data set rather than the real time data set refered to here – this allowed us to compare both quarterly and yearly models over a 14 year period. The key results were still visible in the non-real-time data. Further details available on request.

forecasting because of this very reason (see, for example, Iacoviello (2001), for a comparison of a VAR and a BVAR for forecasting Italian GDP). For quarterly models, VECMs and BVECMs are the most accurate; in contrast, the VAR models are least accurate.

Table 5.1 Accuracy of real time forecasts made in March

	Current year			Next year		
	Mean error	MAE	RMSE	Mean error	MAE	RMSE
1993-2006						
SAFFIER (CEP)	− 0.13	0.98	1.17	0.01	1.19	1.48
Average from individual models						
Yearly VAR	0.05	1.29	1.59	0.23	1.65	2.11
Yearly dVAR	0.05	1.23	1.53	0.07	1.42	1.83
Yearly BVAR	− 0.08	1.01	1.20	− 0.03	*1.13*	*1.36*
Yearly BdVAR	− 0.05	1.13	1.31	0.13	*1.12*	*1.40*
Pooled across models						
Yearly VAR	0.05	1.04	1.24	0.23	*1.15*	1.56
Yearly dVAR	0.05	1.04	1.22	0.07	*1.09*	*1.41*
Yearly BVAR	− 0.08	0.99	*1.13*	− 0.03	*1.04*	*1.28*
Yearly BdVAR	− 0.05	1.11	1.28	0.13	*1.05*	*1.34*
All yearly models	0.03	1.00	1.19	0.13	*1.06*	*1.41*
2001-2006						
SAFFIER (CEP)	0.47	0.97	1.14	0.86	1.34	1.81
Average from individual models						
Quarterly VAR	0.89	1.09	1.32	1.10	1.37	*1.68*
Quarterly dVAR	0.72	*0.96*	1.24	0.98	*1.29*	*1.68*
Quarterly VECM	0.67	*0.94*	1.20	0.98	1.37	*1.71*
Quarterly BVAR	0.82	1.03	1.27	1.12	1.38	*1.70*
Quarterly BdVAR	0.72	*0.95*	1.21	0.88	*1.19*	*1.59*
Quarterly BVECM	0.61	*0.90*	1.17	0.95	*1.33*	*1.64*
Pooled across models						
Quarterly VAR	0.89	*0.97*	1.19	1.10	*1.18*	*1.55*
Quarterly dVAR	0.72	*0.86*	1.17	0.98	*1.24*	*1.64*
Quarterly VECM	0.67	*0.85*	*1.13*	0.98	*1.33*	*1.65*
Quarterly BVAR	0.82	*0.93*	1.16	*1.12*	*1.18*	*1.59*
Quarterly BdVAR	0.72	*0.85*	1.15	0.88	*1.19*	*1.59*
Quarterly BVECM	0.61	*0.81*	*1.12*	0.95	*1.30*	*1.60*
All quarterly models	0.75	*0.89*	1.15	1.02	*1.15*	*1.55*
All yearly models	0.48	1.40	1.55	0.91	1.39	1.86

Pooling the forecasts within a model class improves accuracy across the board. In particular, pooling improves the accuracy of classically estimated models more than it improves the accuracy of Bayesian estimated models. For yearly models, pooled dVAR forecasts are more

accurate than the pooled BdVAR forecasts for the current year. The pooled BVAR is the most accurate for the following year. In fact, for next year forecasts, only the VAR models do not produce lower RMSE forecasts than SAFFIER when pooled, although they do have lower MAE. For the current year, the pooled VAR and dVAR models are now approaching the accuracy of SAFFIER. The pooled BdVAR forecasts improve only slightly and remain less accurate than SAFFIER.

For the current year, none of the pooled quarterly models is less accurate than SAFFIER. The most accurate for the current year is the pooled BVECM, although the VECM is not far behind. For the following year the pooled VAR is the most accurate, although there is very little difference between the pooled VAR, BVAR or BdVAR forecasts. Comparing the pooled quarterly forecasts to the pooled forecast from all yearly models over the same period we use for evaluating the quarterly models, the quarterly models are more accurate. This suggests that there is extra information content in the quarterly series that can be used for forecasting. Furthermore, pooling all quarterly models is close to the most accurate for both the current and next year.

An alternative measure of accuracy is the mean error. This can show if forecasts are systematically biased. The mean error is much lower over the 1993-2006 period than over 2001-2006. In the period longer period the average growth rate of GDP was 2.5%; whereas in the latter period the growth rate was only 1.5%. The higher mean error for the latter period shows the effects of the downturn in the business cycle during these years. For the yearly models over the period 1993-2006, there is very little difference between the yearly VARs and SAFFIER. Pooling both VARs and dVARs produces mean errors comparable to our published forecasts. For the period 2001-2006, SAFFIER is hard to beat, although quarterly BdVARs are comparable. One further point of note is that the average of all yearly models produces lower mean errors than the quarterly models, even when evaluated over the later period. All in all, for unbiasedness, SAFFIER is hard to beat.

5.1.2 Real time forecasts made in September

As described in Section 3, CPB produces forecasts in March and September. We also made forecasts for all models using the September data to compare to the forecasts published in various MEVs. As with the forecasts made in March, the forecasts are for the GDP growth rate in the current year and the following year. Since the MEV forecasts are published in September preliminary data are available for the first two quarters of the current year when the forecasts are made. The yearly models do not use this extra quarterly information, but a newer revision of the yearly data is available, which they do use.

As expected, the accuracy of the quarterly models, and of SAFFIER, is better for September forecasts than for March forecasts. The yearly forecasts, however, generally become *less* accurate. The reason for this deteriation is unkown. Having said this, it is worth noting that the pooled yearly VAR forecasts for the next year are still of comparable accuracy to the published

Table 5.2 Accuracy of real time forecasts made in September

	Current year			Next year		
	Mean error	MAE	RMSE	Mean error	MAE	RMSE
1993-2006						
SAFFIER (MEV)	− 0.21	0.69	0.77	0.09	1.13	1.37
Average from individual models						
Yearly VAR	0.19	1.30	1.61	0.37	1.67	2.15
Yearly dVAR	0.16	1.33	1.62	0.10	1.49	1.88
Yearly BVAR	0.03	1.12	1.33	− 0.07	1.16	1.38
Yearly BdVAR	0.07	1.15	1.34	0.09	1.20	1.46
Pooled across models						
Yearly VAR	0.19	0.97	1.20	0.37	1.21	1.60
Yearly dVAR	0.16	1.14	1.28	0.10	1.17	1.49
Yearly BVAR	0.03	1.11	1.26	− 0.07	*1.09*	*1.31*
Yearly BdVAR	0.07	1.12	1.31	0.09	*1.11*	1.41
All yearly models	0.15	1.05	1.21	0.19	*1.12*	1.47
2001-2006						
SAFFIER (MEV)	− 0.03	0.53	0.62	0.90	1.27	1.62
Average from individual models						
Quarterly VAR	− 0.19	*0.38*	*0.47*	0.81	1.43	1.76
Quarterly dVAR	− 0.28	*0.42*	*0.49*	0.61	1.35	1.71
Quarterly VECM	− 0.28	*0.42*	*0.52*	0.53	1.40	1.69
Quarterly BVAR	− 0.20	*0.39*	*0.48*	0.72	1.43	1.79
Quarterly BdVAR	− 0.26	*0.40*	*0.47*	0.60	1.37	1.73
Quarterly BVECM	− 0.30	*0.43*	*0.53*	0.35	1.39	1.71
Pooled across models						
Quarterly VAR	− 0.19	*0.31*	*0.40*	0.81	1.28	*1.55*
Quarterly dVAR	− 0.28	*0.38*	*0.44*	0.61	1.32	*1.57*
Quarterly VECM	− 0.28	*0.38*	*0.47*	0.53	1.36	*1.56*
Quarterly BVAR	− 0.20	*0.31*	*0.42*	0.72	1.30	*1.61*
Quarterly BdVAR	− 0.26	*0.37*	*0.43*	0.60	1.32	*1.59*
Quarterly BVECM	− 0.30	*0.37*	*0.48*	0.35	1.38	*1.59*
All quarterly models	− 0.25	*0.35*	*0.44*	0.63	1.30	*1.55*
All yearly models	0.86	1.53	1.60	1.23	1.61	2.05

forecasts despite ignoring the extra two quarters of information available. As with the forecasts made in March, the pooled BVARs are the most accurate yearly model. For the current year, the published forecasts are less accurate than both the average model in each class and the pooled forecast from each class of quarterly model. For the following year, the published forecast is more accurate than the average quarterly VAR based forecast. With regards to the pooled quarterly forecasts the conclusion depends on whether MAE or RMSE is the accuracy measure - SAFFIER does best with MAE whereas the VAR based models do best on RMSE. In contrast

with the March forecasts, there is little difference in accuracy between Bayesian and classical models.

With regards the mean error, the picture for forecasts made in September is similar to that for forecasts made in March. Once again, the forecasts produced using SAFFIER are hard to beat except for next year forecasts over the period 2001-2006. Whereas the mean error for each class of VAR based models falls when we compare the forecasts made in September to those made in March; for SAFFIER, the mean error rises.

5.1.3 Conclusion on real time forecasts

In summary over both March and September forecasts, pooling all quarterly models is a reasonable strategy. Whilst this does not always produce the most accurate forecasts, it is never beaten convincingly by our published forecasts on both MAE and RMSE. The only case where a class of VAR based model does not convincingly beat the published forecasts is for next year forecasts published in September. However, since pooled yearly forecasts ignore the extra information that is available in September and still produce a similarly accurate forecast, it still suggests that the accuracy of our published forecasts could be improved by considering pooled forecasts from VAR based models. Pooled quarterly models also perform comparably to SAFFIER - they are more accurate on RMSE but less accurate on MAE.

5.2 Testing the four hypotheses

5.2.1 Do simple models do best?

One of the key implications of the literature is that VECM models should perform poorly because of their sensitivity to structural breaks. However, we have found that VECMs were the best performers in our competition for current year forecasts made in March. There are two possible reasons for our disagreement with the literature: either simple models do not always perform best or there were no structural breaks between 1979 and 2006. Furthermore, whilst VARs perform worse individually in most cases, when pooled they improve the most and even become the most accurate for next year forecasts in March and September.

Furthermore, adding more variables to the model (see table 5.3) improves the forecast accuracy of the average of individual quarterly VECM models, the root mean square error for the average 5 variable model is 10-15% lower than for the bivariate model. For the average of individual yearly dVARs the picture is less clear, the picture detoriates for the 3 variable model and then improves slightly by adding 1 or 2 variables more. Adding variables is favourable for the results of pooled forecasts: whilst univariate yearly models have the lowest RMSEs of the yearly models when considered individually and 3 variable models the highest, this is reversed after pooling. More included variables means more estimated models and more potential sources of information, so this is not entirely surprising.

Table 5.3 The effect of increasing the number of variables on forecast accuracy in March

	Current year			Next year		
	Mean error	MAE	RMSE	Mean error	MAE	RMSE
Average from individual models						
Univariate yearly dVAR	− 0.20	1.25	1.36	− 0.19	1.19	1.44
Bivariate yearly dVAR	− 0.07	1.19	1.38	− 0.08	1.25	1.52
3 variable yearly dVAR	0.03	1.26	1.56	0.09	1.47	1.96
4 variable yearly dVAR	0.06	1.24	1.55	0.06	1.43	1.84
5 variable yearly dVAR	0.09	1.21	1.53	0.10	1.41	1.80
Bivariate quarterly VECM	0.69	1.12	1.39	1.07	1.56	1.91
3 variable quarterly VECM	0.66	1.00	1.28	1.01	1.43	1.77
4 variable quarterly VECM	0.66	0.93	1.20	0.98	1.38	1.71
5 variable quarterly VECM	0.67	0.91	1.17	0.96	1.34	1.67
Pooled across models						
Univariate yearly dVAR	− 0.20	1.24	1.35	− 0.19	1.18	1.44
Bivariate yearly dVAR	− 0.07	1.09	1.22	− 0.08	1.08	1.38
3 variable yearly dVAR	0.03	1.04	1.21	0.09	1.00	1.33
4 variable yearly dVAR	0.06	1.05	1.22	0.06	1.09	1.42
5 variable yearly dVAR	0.09	1.07	1.26	0.10	1.15	1.49
Bivariate quarterly VECM	0.69	0.98	1.30	1.07	1.42	1.83
3 variable quarterly VECM	0.66	0.87	1.20	1.01	1.36	1.72
4 variable quarterly VECM	0.66	0.84	1.13	0.98	1.33	1.65
5 variable quarterly VECM	0.67	0.85	1.11	0.96	1.31	1.62

Generally, increasing the lag length only improves the forecast accuracy for the average of individual and pooled forecasts of VECM models in both the current and the next year (see table 5.4). For yearly dVARs adding lags is bad for the individual models. When pooled, however, the next year forecasts become more accurate with extra lags. Again, this could be evidence that the extra information available with extra lags is being usefully extracted through the pooling process. That quarterly models benefit more from extra lags is intuitive since the quarterly models have approximately 4 times the number of observations for estimation than the yearly models have available.

The conclusion of Hendry and Clements that simple robust models perform best is not entirely met by the above results. VECMs with more variables and lags mostly do improve the forecasting accuracy. For yearly dVARs the results, particularly for individual models, look more in line with the Hendry and Clements rule.

5.2.2 Does the accuracy measure matter?

Whilst we find that the ranking of our models differ occasionally when they are evaluated using the mean absolute error or the root mean square error, this only occurs when the models are of similar accuracy on both measures. However, this does not rule out differences for other loss

Table 5.4 The effect of increasing the lag length on forecast accuracy in March

	Current year			Next year		
	Mean error	MAE	RMSE	Mean error	MAE	RMSE
Average from individual models						
Yearly dVAR(1)	0.28	1.16	1.47	0.28	1.31	1.68
Yearly dVAR(2)	− 0.12	1.21	1.49	− 0.16	1.41	1.75
Yearly dVAR(3)	− 0.04	1.33	1.62	0.04	1.25	1.51
Yearly dVAR(4)	− 0.01	1.49	1.91	0.29	1.83	2.73
Quarterly VECM(0)	0.72	1.00	1.27	1.01	1.42	1.76
Quarterly VECM(1)	0.65	0.93	1.19	0.94	1.39	1.71
Quarterly VECM(2)	0.65	0.93	1.19	0.99	1.38	1.72
Quarterly VECM(3)	0.61	0.91	1.17	0.93	1.31	1.64
Pooled across models						
Yearly dVAR(1)	0.28	0.97	1.27	0.28	1.11	1.54
Yearly dVAR(2)	− 0.12	1.09	1.23	− 0.16	1.22	1.45
Yearly dVAR(3)	− 0.04	1.11	1.28	0.04	1.10	1.42
Yearly dVAR(4)	− 0.01	1.19	1.34	0.29	0.99	1.33
Quarterly VECM(0)	0.72	0.93	1.20	1.01	1.34	1.68
Quarterly VECM(1)	0.65	0.83	1.13	0.94	1.35	1.66
Quarterly VECM(2)	0.65	0.81	1.12	0.99	1.34	1.68
Quarterly VECM(3)	0.61	0.81	1.09	0.93	1.26	1.58

functions, especially asymmetric loss functions. Given that we do not directly observe the loss functions of the users of our forecasts we must make some assumptions in order to produce the most relevant forecast for our customers; the relative accuracy of models is robust to these two commonly used measures of accuracy when one allows for some uncertainty around the reported accuracy figures. A further interesting observation is that SAFFIER and the yearly models have lower mean errors than the quarterly models when we confine ourselves to the March forecasts, when the yearly models use the same vintage of data as the quarterly models and SAFFIER. This does not seem to confer any accuracy advantage on the other two measures, especially for the current year.

5.2.3 Does pooling help?

Within each and every class of models we find that pooling helps reduce MAE and RMSE towards the best performing models. For yearly models, BVARs do best and better than pooling everything, especially when evaluated on RMSE. For quarterly models, the question of what is the optimal number of variables or lags to include is made redundant by the observation that the pooled forecast from all models has comparable accuracy as the pooled from the ‘best’ size and lag length. We also found using a single data set[17] over a longer period that quarterly models

[17] That is, using the latest data rather than real time data. Results on request.

Why pooling works

Our results show that the pooled forecast outperforms the average of the individual forecasts in terms of RMSE. So, a better forecast is obtained by combining the forecasts of the underlying models. This is a well-known phenomenon in the literature, see Clemen (1989) and Timmermann (2006) for literature surveys. In some studies it is even found that the pooled forecast outperforms the best underlying model. Empirically, it turns out that a simple average of different forecasts is hard to beat. Why pooled forecasts perform so well is not completely understood and is still the subject of research. A practical example is that of Consensus Economics, who combine the forecasts of several institutions and publishes the simple average and has been quite succesful.

Whilst Hendry and Clements (2004) show that it is impossible to beat the 'optimal' model under certain assumptions, these assumptions are not met in practice so the optimal model does not exist. These departures can be due to mis-specification, mis-estimation or non-stationarities. Therefore departures from 'optimality' are necessary to gain from combining forecasts.

Timmermann (2006) sums up a number of possible explanations why pooling may be successful. First, the individual forecasts may use different information sources so a combination allows more information to bear upon the forecast than from an individual model. Second, individual forecasts may be very differently affected by structural breaks. Some will be adapt quickly and will only be temporarily affected; other models have parameters that will adjust only slowly to new post-break data. Since it is difficult to detect structural breaks in real time, it is plausible that combinations of forecasts from models with different degrees of adaptability will outperform forecasts from individual models. Third, pooled forecasts may be more robust against misspecification biases and measurement errors in the dataset: forecasting models can be seen as local approximations and it is implausible that the same model dominates all others at all points in time. Fourth, the underlying forecasts may be based on different loss functions. If the loss function for a specific forecast entails large losses when the forecast is above the realisation, then the forecast produced using this loss function will below the mean realisation. Combining forecasts using many different loss functions results in an overall loss function centred on the mean realisation in much the same way as the central limit theorem for sample means works.

In our forecasting exercise, the individual VAR-models are based on a limited number of variables. The maximum number of variables is 5, but the dataset we use consists of 10 variables. So the individual models do not use all available information. In this case, combining the forecasts of these incompletely overlapping models might do better than the individual forecasts because they utilise more information without necessarily suffering from degrees of freedom problems that a 10 variable model would. For example, our pooled quarterly forecasts employ 3830 individual forecasts, which allows the information in our data set many opportunities to enter the forecast.

We also pool forecasts of models with variables in levels and in growth rates. These models will react differently to structural breaks. The models with growth rates will react fast to the new dataseries, while the parameters of the models with levels will only change slowly. This does not appear to be the reason our pooled forecasts do well: the pooled forecast across all classes of model is never as good as pooling within one class of models. A related issue is that the combination of models is more robust to misspecification.

Using a simple average of forecasts seems to work well. Since there is little or no relationship between fit statistics and forecast accuracy of the individual models, there appears to be no alternative basis upon which to weight the individual models. Combining forecasts in this way is relatively easy way of producing competitive forecasts, especially when compared to the technical complexity of other methods that attempt to extract information from many different sources, such as factor models.

were, in general, more accurate than the yearly models. One important consequence of this is that it is possible to produce a competitive forecast for GDP growth without a large amount of specialist knowledge. Our procedure, where we first look at simple cross-correlations to pick variables, then estimate all possible combinations and pool the forecasts, produces forecasts that are more accurate than our published forecasts. It is not necessary to choose an individual VAR based model – pooling across all models produces a competitive forecast.

With regards yearly models, one important conclusion is that pooling works better for classically estimated models than for Bayesian estimated models. There are a number of potential explanations for this all related to the limited degrees of freedom available for yearly models. Firstly, pooling works like a sort of 'poor man's Bayesian' estimation as far as the lag structure is concerned, at least when there are not enough degrees of freedom available to make 4 lag models reliable on their own. Models with 4 lags contain all of the lags from 1 to 4, models with 3 lags only 1 to 3; when these are pooled this implies more weight on lag 1 than on lag 2, and so on. Alternatively it may be because all Bayesian models as being biased towards the Minnesota prior specification. Hence, there is less variation to take advantage of when it comes to pooling. If we look at table 5.3 we can see that 4 lag models benefit considerably more from pooling than the other lag lengths. The longer lag length is more likely to be sensitive to degrees of freedom and overfitting problems, which leads to large variations in the forecasts from the different models. It appears that this can be overcome by pooling. If we look at table 5.4 we can indeed see that the variation in forecasts of the 4 lag models is important.[18] If the 'poor man's Bayesian' story were the more important we would expect to see less variation in the benefits of pooling at a given lag length in table 5.4. However, pooling classically estimated models does not entirely remove the accuracy advantage traditionally associated with Bayesian VARs. For quarterly models there is no discernable difference between Bayesian and classical estimated models in the effects of pooling. The only noticeable result is that VARs and BVARs in levels tend to benefit more from pooling than the other specifications.

5.2.4 Does the evaluation horizon matter?

For forecasts made in March, pooled quarterly VECMs and BVECMs are the most accurate for the current year, whilst pooled quarterly VARs and BVARs are the most accurate for the following year. The most accurate forecasts for the current year made in September were produced by quarterly VARs and BVARs. For the following year it is difficult to beat our published MEV forecasts, although pooled quarterly VARs and BVARs are comparable. This shows that the horizon clearly matters for the choice of forecasting model. It is also interesting to note that the date the forecast is made is important for our published forecasts: when made in

[18] Due to degrees of freedom limitations the 4 variable models were not estimated with 4 lags and the 5 variable models were not estimated with either 3 or 4 lags.

March, the next year forecasts are convincingly beaten by both yearly and quarterly models; when made in September they are among the most accurate.

5.3 Fit versus accuracy

Table 5.5 shows correlations between various measures of in-sample fit for the period up to 1992 and forecast accuracy in the entire subsequent evaluation period for different classes of VAR model.[19] All of the correlation in the tables have been adjusted so that a positive correlation corresponds to better in-sample fit being associated with more accurate forecasts. The vast majority of the correlations are, however, negative. This is in line with other similar studies reported in the literature.

Quarterly dVARs are an exception, though. There is a positive correlation between current year accuracy and the two information criteria: the Akaike Information Criterion (AIC) and the Schwartz Bayesian Criteria (SBC). For next year forecasts there are some positive correlations but these are close to zero. Does this mean that it would be possible to select good models using these information criteria? The AIC and SBC are shown in equations 5.1 and 5.2 below.

$$AIC = 2k - 2ln(L) \tag{5.1}$$

$$SBC = kln(n) - 2ln(L) \tag{5.2}$$

Here, k is the number of estimated parameters in the model, $ln(L)$ is the log-likelihood of the model and n is the sample size. For both criteria a lower number implies a better fit. In samples of size 8 or above, the SBC penalises extra parameters more then the AIC. This goes some way towards explaining the positive correlations for the quarterly dVARs. Whilst there is a negative correlation between the log-likelihood and accuracy, when it is adjusted for the number of parameters estimated it becomes positive and the SBC is higher than the AIC correlation. For quarterly dVARs, increasing the lag length decreased the average accuracy, so these positive correlations are simply picking up the relationship between lag length and average accuracy. Indeed, if the correlations are recalculated separately for models with a given lag length, the positive correlation disappears. Still, it is useful to ask if this information is useful for selecting which models to pool. The quarterly pooled dVARs had an MAE of 0.84 and an RMSE of 1.05 when evaluated against the same data set they were estimated on.[20] When we pooled only those models that had a better than average AIC the MAE was also 0.84 and the RMSE was also 1.05. Doing the same for SBC gave an MAE of 0.83 and an RMSE of 1.04.[21] Even though individual

[19] The results for the quarterly models presented here are based on using a single data set. That is, rather than using a separate real-time data set for each year, equivalent statistics were calculated using a single data set where the estimation period and benchmark were recursively moved through the sample period. This was necessary to overcome the short selection of real-time data sets for quarterly data.

[20] See Appendix B for further discussion of the appropriate benchmark.

[21] Other cut off points were also used without improving the accuracy of the pooled forecasts.

models with better fit were slightly more accurate, this advantage disappears after pooling. We have seen similar results before, especially in tables 5.3 and 5.4, were the average accuracy of the individual models was poor, but after pooling they were relatively accurate.

Table 5.5 The relationship between in-sample fit statistics up to 1992 and forecast accuracy post 1992

			Log-likelihood	AIC	SBC	r^2	Adjusted r^2
Yearly VAR	1yr	MAE	− 0.35	− 0.40	− 0.38	− 0.16	− 0.09
		RMSE	− 0.43	− 0.47	− 0.45	− 0.24	− 0.18
	2yr	MAE	− 0.44	− 0.44	− 0.42	− 0.30	− 0.25
		RMSE	− 0.48	− 0.50	− 0.48	− 0.32	− 0.28
Yearly dVAR	1yr	MAE	− 0.45	− 0.33	− 0.25	− 0.33	− 0.02
		RMSE	− 0.46	− 0.41	− 0.35	− 0.39	− 0.17
	2yr	MAE	− 0.42	− 0.37	− 0.31	− 0.31	− 0.13
		RMSE	− 0.34	− 0.32	− 0.28	− 0.26	− 0.13
Quarterly VAR (recursive)	1yr	MAE	− 0.40	− 0.38	− 0.22	− 0.25	− 0.39
		RMSE	− 0.39	− 0.30	− 0.11	− 0.30	− 0.35
	2yr	MAE	− 0.19	− 0.19	− 0.20	− 0.01	− 0.15
		RMSE	− 0.20	− 0.31	− 0.37	0.08	− 0.20
Quarterly dVAR (recursive)	1yr	MAE	− 0.11	0.24	0.39	− 0.39	− 0.17
		RMSE	− 0.10	0.21	0.35	− 0.36	− 0.16
	2yr	MAE	0.05	− 0.06	− 0.12	0.15	0.09
		RMSE	0.06	− 0.02	− 0.09	0.12	0.09

We also found no correlation between the Quandt-Andrews structural break statistics and forecast accuracy[22], suggesting that the Netherlands has not been subject to significant structural changes, at least as far as forecasting GDP growth is concerned. This is also in line with the good performance of VECM and BVECM models in table 5.1.

[22] Details available on request

6 Conclusion

At CPB one of our tasks is to produce forecasts for a wide range of macroeconomic variables for a two year horizon. These forecasts are made using SAFFIER, a large macro model. In this paper we have compared the real time forecast accuracy of our published GDP growth forecasts with those made with VAR based models over the period 1993-2006. We selected nine variables based on their leading correlations with GDP measures to include in our VAR models. Since large VAR models are constrained by degrees of freedom issues we looked at all possible combinations of smaller VAR systems (up to five variables) rather than a ten-variable VAR.

We find that the average accuracy of individual VAR based models is not better than our published forecasts at most forecast horizons, although some individual VAR models were more accurate (and some less accurate) in our sample period. The main exception is for current year forecasts made in September where quarterly models perform markedly better regardless of the estimation technique. Bayesian models also perform well for next year forecasts made in March. However, when we looked further into whether it would have been possible to pick good models based on available real time information, we found that it would not have been possible. for most classes of model there is no correlation between various measure of in-sample fit and forecast accuracy and there is no correlation between forecast accuracy in one period and forecast accuracy in subsequent periods. For quarterly dVAR models there was a correlation between both the Akaike Information Criterion (AIC) and the Schwartz Bayesian Criteria (SBC). Upon further investigation this was entirely caused by models with fewer lags being more accurate - among models with the same number of lags there was no correlation.

Selecting the ‘best’ model may not be necessary, however, since if we pool the forecasts from many VAR based models, the pooled forecast is more accurate than the average accuracy of the individual models. In our competition we find that pooled VAR based forecasts are either better or as good as our published forecasts for all horizons. Interestingly, pooling allowed classically estimated models that were inaccurate due to degrees of freedom constraints to approach the accuracy of Bayesian estimated models. We suggest that this is because Bayesian methods bias the estimates of all models towards the prior, which results in less variation to take advantage of when pooling. Further research into forecasts for other variables may be of interest too.

If we consider the relative performance of the competing models in historical perspective, we can see that our large macro model still outperforms individual VAR based forecasts on average, as reported by Wallis (1989) for the UK. However, the recent advances in the application of pooled forecasts show that data driven models can still produce more accurate forecasts than our large macro model. Since pooling attempts to utilise information from many sources, it is of interest to compare the accuracy to dynamic factor models, which seek to do the same. Den Reijer (2005) finds that a dynamic factor model has mean square errors that are smaller than an AR model; at one-quarter-ahead they are 70% of those from the AR model, rising to 98% for

eight-quarter-ahead forecasts. We note that the accuracy of the pooled forecasts from the best performing class of models for the current year is 66% and 80% of the MAE and RMSE, respectively, of a univariate model. The MAE and RMSE for the following year are 82% and 84%. These magnitudes compare well with those reported for the dynamic factor model using 370 time series; our pooled VARs use only 10 series.

Appendix A Data sources

Tables A.1 and A.2 below show the sources of the data used in this paper and a description of the difference between the yearly versions and the quarterly versions. Where necessary, yearly levels series were created from the growth rate series listed below. For the yearly versions of short term interest rates, production expectations, consumer confidence, bankruptcies and German business climate the observation for the last quarter of the year is used. For yearly dVAR models the first difference of these series is taken to be the difference between the fourth quarter observation in one year and the fourth quarter observation in the previous year, rather than the third quarter observation of the original year.

Table A.1 Data sources for yearly time series

Variable	Source
GDP, real growth rate	Statistics Netherlands
Private consumption, real growth rate	Statistics Netherlands
Compensation per employee, market sector growth rate	Statistics Netherlands
CPI (inflation)	Statistics Netherlands
Relevant world trade, growth rate	CPB
Short term interest rates (3 months)	DNB
Production expectations, manufacturing indsutry	Statistics Netherlands
Consumer confidence	Statistics Netherlands
Bankruptcies	Statistics Netherlands
German business climate (Ifo)	IFO

Table A.2 Data sources for quarterly time series

Variable	Source
GDP, real level	Statistics Netherlands
Private consumption, real level	Statistics Netherlands
Compensation per employee, market sector level	Statistics Netherlands
CPI	Statistics Netherlands
Relevant world trade, level	CPB
Short term interest rates (3 months)	DNB
Production expectations, manufacturing indsutry	Statistics Netherlands
Consumer confidence	Statistics Netherlands
Bankruptcies	Statistics Netherlands
German business climate (Ifo)	IFO

Appendix B Choice of benchmark

There were two possible choices of benchmark GDP growth realisations available to us: GDP growth rates computed from the latest available GDP series or GDP growth rates using data compiled to match the data set used to produce the forecasts. The difference from these two series arises because the former is subject to methodological revisions including changes to the way that components of GDP are defined and measured. It follows that the latest available GDP figure for previous years will differ from that available at the time because of standard revisions and methodological revisions. For example, suppose that it is decided that the current method for measuring investment does not adequately take into account quality changes in, for example, information and communication technology and, hence, the current method understates the true level of investment. If a new method is used to better account for these quality changes the new values for investment and GDP will be higher than the original figures. GDP growth rates will also be different depending on which method is used. In the present study we have used realisations of the GDP growth rate that use the same method as the real time data we use to estimate the models and make forecasts with as our benchmark. Table B.1 highlights the main differences that would have arisen had we chosen to use the latest figures as our benchmark. The relative ranking of the different classes of models is robust to this choice. One interesting point of note is that all models have lower MAE and RMSE when evaluated against the latest data as opposed to the methodologically consistent data for forecasts made in March (but not for those made in September, which are not shown). The precise cause of this is unclear and more research is needed to determine the exact cause.

Table B.1 The effect of different benchmarks GDP growth series on the accuracy of quarterly real time forecasts made in March

	Method consistent			Latest revision		
	Mean error	MAE	RMSE	Mean error	MAE	RMSE
Average VAR	0.89	1.09	1.32	0.74	0.88	0.99
Average dVAR	0.72	0.96	1.24	0.57	0.74	0.87
Average VECM	0.67	0.94	1.20	0.52	0.72	0.82
Average BVAR	0.82	1.03	1.27	0.67	0.82	0.93
Average BdVAR	0.72	0.95	1.21	0.57	0.72	0.82
Average BVECM	0.52	0.95	1.20	0.37	0.70	0.81
Pooled VAR	0.89	0.97	1.19	0.74	0.74	0.82
Pooled dVAR	0.72	0.86	1.17	0.57	0.62	0.75
Pooled VECM	0.67	0.85	1.13	0.52	0.62	0.72
Pooled BVAR	0.82	0.93	1.16	0.67	0.69	0.78
Pooled BdVAR	0.72	0.87	1.15	0.57	0.63	0.72
Pooled BVECM	0.52	0.84	1.13	0.37	0.58	0.70

Appendix C The influence of individual variables on forecast accuracy

To get an impression of the importance of the different individual variables in explaining forecast accuracy we present the results for the quarterly VECM models grouped by variable in table C.1. We restrict this analysis to one class of models as the general conclusions for VARs and dVARs are comparable to those for VECMs.

The picture for the different variables is very close to the average of individual and pooled results for all VECM models. Only models with the number of bankruptcies as a variable produce better forecasts on all three criteria for both years: mean error, mean absolute error and root mean square error. The improvement is about 10-15%, which is quite a lot since there is a considerable overlap between the different models per variable and thus a bias to the mean. A case could also be made that the leading indicator variables (the last 5) do marginally better on average than the economic variables.

Table C.1 The accuracy of real time forecasts made in March of quarterly VECM models containing each variable

	Current year			Next year		
	Mean error	MAE	RMSE	Mean error	MAE	RMSE
Average from individual models						
All	0.67	0.94	1.20	0.98	1.37	1.71
Consumption	0.61	0.98	1.24	0.89	1.37	1.69
Inflation	0.74	0.98	1.24	1.07	1.35	1.73
Employee compensation	0.74	0.96	1.22	1.06	1.32	1.68
World trade	0.73	0.97	1.24	1.02	1.40	1.72
Short term interest	0.62	0.94	1.19	0.97	1.38	1.73
Business climate survey	0.59	0.87	1.13	0.91	1.36	1.64
Consumer confidence	0.66	0.93	1.20	0.91	1.42	1.72
Bankruptcies	0.61	0.84	1.07	0.86	1.29	1.62
German business climate	0.69	0.88	1.18	1.04	1.38	1.72
Pooled across models						
All	0.67	0.85	1.13	0.98	1.33	1.65
Consumption	0.61	0.88	1.19	0.89	1.34	1.63
Inflation	0.74	0.94	1.19	1.07	1.30	1.69
Employee compensation	0.74	0.92	1.16	1.06	1.26	1.64
World trade	0.73	0.87	1.16	1.02	1.34	1.66
Short term interest	0.62	0.82	1.13	0.97	1.32	1.67
Business climate survey	0.59	0.81	1.08	0.91	1.34	1.60
Consumer confidence	0.66	0.84	1.13	0.91	1.39	1.66
Bankruptcies	0.61	0.77	1.00	0.86	1.26	1.57
German business climate	0.69	0.81	1.14	1.04	1.32	1.67

References

Andrews, D.W.K., 1993, Tests for parameter instability and structural change with unknown change point, *Econometrica*, vol. 61, no. 4, pp. 821–56.

Boero, G., 1990, Comparing ex-ante forecasts from a SEM and VAR model: An application to the Italian economy, *Journal of Forecasting*, vol. 9, no. 1, pp. 13–24.

Burns, S.T., 1986, The interpretation and use of economic predictions, Predictability in science and society, *Proceedings of the Royal Society of London Series A*, vol. 407, no. 1832, pp. 103–125.

Clemen, R.T., 1989, Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting*, vol. 5, pp. 559–583.

Clements, M.P. and D.F. Hendry, 1998, *Forecasting Economic Time Series*, Cambridge University Press.

Cooper, J.P. and C.R. Nelson, 1975, The Ex Ante Prediction Performance of the St. Louis and FRB-MIT-PENN Econometric Models and Some Results on Composite Predictors, *Journal of Money, Credit and Banking*, vol. 7, no. 1, pp. 1–32.

CPB, 1992, *FKSEC: A Macro-econometric Model for the Netherlands*, Stenfert Kroese.

CPB, 2003, SAFE; a quarterly model of the Dutch economy for short-term analyses, CPB Document 42.

Dhrymes, P.J., E.P. Howrey, S.H. Hymans, J. Kmenta, E.E. Leamer, R.E. Quandt, J.B. Ramsey, H.T. Shapiro and V. Zarnowitz, 1972, Criteria for evaluation of econometric models, *Annals of Economic and Social Measurement*, vol. 1, no. 3, pp. 291–324.

Edge, R.M., M.T. Kiley and J.P. Laforte, 2006, A comparison of forecast performance between Federal Reserve staff forecasts, simple reduced-form models, and a DSGE model, mimeo, IMF.

Eitrheim, O., T.A. Husebo and R. Nymoen, 1999, Equilibrium-correction vs. differencing in macroeconometric forecasting, *Economic Modelling*, vol. 16, no. 4, pp. 515–544.

Elliott, G. and A. Timmermann, 2007, Economic forecasting, CEPR Discussion Paper 6158.

Fildes, R. and S. Makridakis, 1995, The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting, *International Statistical Review/Revue Internationale de Statistique*, vol. 63, no. 3, pp. 289–308.

Franses, P.H., H.C. Kranendonk and D. Lanser, 2007, On the optimality of expert-adjusted forecasts, CPB Discussion Paper 92.

Hendry, D.F. and M.P. Clements, 2003, Economic forecasting: some lessons from recent research, *Economic Modelling*, vol. 20, no. 2, pp. 301–329.

Hendry, D.F. and M.P. Clements, 2004, Pooling of forecasts, *Econometrics Journal*, vol. 7, pp. 1–31.

Holden, K., 1997, A comparison of forecasts from UK economic models and some Bayesian vector autoregressive models, *Journal of Economic Studies*, vol. 24, no. 4, pp. 242–256.

Iacoviello, M., 2001, Short-term forecasting: Projecting Italian GDP, one quarter to two years ahead, IMF Working Papers 01/109.

Jacobs, J.P.A.M. and K.F. Wallis, 2007, Cointegration, long-run structural modelling and weak exogeneity: Two models of the uk economy, CAMA Working Papers 2007-12, Australian National University, Centre for Applied Macroeconomic Analysis.

Johansen, S., 1995, *Likelihood-based inference in cointegration in the vector autoregressive model*, Oxford University Press.

King, R.G., C.I. Plosser, J.H. Stock and M.W. Watson, 1991, Stochastic trends and economic fluctuations, *American Economic Review*, vol. 81, no. 4, pp. 819–40.

Kranendonk, H.C. and J.P. Verbruggen, 2006, Trefzekerheid van korte-termijnramingen en middellange-termijnverkenningen, CPB Document 131.

Kranendonk, H.C. and J.P. Verbruggen, 2007, SAFFIER; a multi-purpose model of the Dutch economy for short-term and medium-term analyses, CPB Document 144.

Lanser, D. and H.C. Kranendonk, 2008, Investigating uncertainty in macro-economic forecasts, CPB Discussion Paper 112.

LeSage, J.P., 1999, Applied Econometrics using MATLAB, Department of Economics, University of Toledo.

Litterman, R.B., 1980, A Bayesian Procedure for Forecasting with Vector Autoregression, Department of economics working paper, Massachusetts Institute of Technology.

Litterman, R.B., 1986, Forecasting with Bayesian Vector Autoregressions: Five Years of Experience, *Journal of Business and Economic Statistics*, vol. 4, no. 1, pp. 25–38.

Lütkepohl, H., 1991, *Introduction to multiple time series analysis*, Springer-Verlag, Berlin.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen and R. Winkler, 1982, The Accuracy of Extrapolation (time series) Methods: Results of a Forecasting Competition, *Journal of Forecasting*, vol. 1, pp. 111–153.

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord and L. Simmons, 1993, The M2-Competition: A Real-Time Judgmentally Based Forecasting Study, *International Journal of Forecasting*, vol. 9, no. 1, pp. 5–22.

Makridakis, S. and M. Hibon, 2000, The M3-Competition: results, conclusions and implications, *International Journal of Forecasting*, vol. 16, no. 4, pp. 451–476.

Mincer, J. and V. Zarnowitz, 1969, The Evaluation of Economic Forecasts and Expectations, *Economic Forecasts and Expectations*.

Reijer, A.H.J. den, 2005, Forecasting Dutch GDP using large scale factor models, DNB Working Papers 028, Netherlands Central Bank.

Robertson, J.C. and E.W. Tallman, 1999, Vector autoregressions: forecasting and reality, *Economic Review*, no. Q1, pp. 4–18.

Sims, C.A., 1980, Macroeconomics and reality, *Econometrica*, vol. 48, no. 1, pp. 1–48.

Sims, C.A., J.H. Stock and M.W. Watson, 1990, Inference in linear time series models with some unit roots, *Econometrica*, vol. 58, no. 1, pp. 113–144.

Smith, R., 1998, Emergent policy-making with macroeconometric models, *Economic Modelling*, vol. 15, no. 3, pp. 429–442.

Stock, J.H. and M.W. Watson, 1998, A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, NBER Working Paper 6607.

Theil, H., 1966, *Applied Economic Forecasting*, North-Holland, Amsterdam.

Theil, H. and A.S. Goldberger, 1961, On Pure and Mixed Statistical Estimation in Economics, *International Economic Review*, vol. 2, no. 1, pp. 65–78.

Timmermann, A., 2006, Forecast combinations, in G. Elliott, C.W.J. Granger and A. Timmermann, eds., *Handbook of Economic Forecasting*, pp. 135–196, North Holland, Amsterdam.

Wallis, K.F., 1989, Macroeconomic forecasting: A survey, *Economic Journal*, vol. 99, no. 394, pp. 28–61.