Pooled VAR based forecasts: A comparison with a large macro model

Adam Elbourne,* Henk Kranendonk, Rob Luginbuhl, Bert Smid

Martin Vromans†

September, 2008

Abstract

We compare the accuracy of real-time VAR based forecasts to those from our large macro model, SAFFIER. VAR models using both classical and Bayesian estimation techniques are studied. We employ a data driven methodology for selecting variables to include in our models, which selects a mix of 'traditional' economic variables and leading indicator variables, such as business surveys. 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. We find no difference in accuracy between survey data and 'traditional' data.

Keywords: SEMs, VAR models, Forecast combination, Bayesian methods, real time

JEL codes: C52, C53, E37

1 Introduction

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). 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

*Contact author. Email address: a.c.elbourne@cpb.nl

†All authors: CPB Netherlands Bureau for Economic Policy Analysis, PO Box 80510, 2508GM The Hague, The Netherlands. The auhtors wish to thank the following people. 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.

the relationships between variables in the economy are unchanged between the estimation period and the forecast period. 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.[1] 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. 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. The models are estimated using both 'traditional' economic variables and leading indicator variables, such as business confidence surveys, which may also help in producing accurate forecasts.

Hendry and Clements [2003] list four key findings from the recent literature on forecasting:

- simple, robust forecasting models perform best
- pooling forecasts improves accuracy
- 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. Would VAR models 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.

[1] 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 analysis of a given policy issue.

2 Findings of recent research

Recent research 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 the assumption that future is the same as the past is not met. Since it is difficult to beat simple time series methods, Hendry and Clements [2003] propose two 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. 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. 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.

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. 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. An example of the application of such a model to forecasting Dutch GDP is to be found in den Reijer [2005]. 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.

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. In March ("Centraal Economisch Plan") and September ("Macro Economische Verkenning") detailed forecasts are published for the current and the next year.

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]).

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.

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.[2]

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%.[3]

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 (1).

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

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

[2]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.

[3]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.

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. 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]). 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 (2).

$$\hat{Y}_{t+1} = A_1 Y_t + ... + A_p Y_{t-p+1} \tag{2}$$

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 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), 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 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 (4)

$$\Delta Y_t = \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + ... + \Gamma_{p-1} \Delta Y_{t-p+1} + u_t \tag{4}$$

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 (4).

We estimate our VECMs using Johansen's technique [Johansen, 1995]. 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.[4]

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

In all other respects the models are left unrestricted. We purposefully ignore issues related to the (weak) exogeneity of series within VECMs,[5] 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]. In our BVARs we use 4 lags to make the models more comparable to the classical VARs and VECMs.[6] The BVARs are estimated using the mixed estimation method of Theil and Goldberger [1961].[7] 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

[5]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.

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

[7]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]).

the BVARs.

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 make 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.[8] The nine variables chosen are listed below (all variables are from Statistics Netherlands except world trade, which is from CPB, short-term interest rates from the Dutch Central Bank and the German business climate from IFO).

- 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.[9] 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

[8] 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.

[9] 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.

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 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.[10] 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.[11]. 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

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

[11]The methodologically consistent realisations can be found in Kranendonk and Verbruggen [2006]

differences between the competing methods we report the results per class of models.[12]

4.3 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,[13] 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.

5.2 Real time forecasts made in March

Table 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

[12] All results are available on request.

[13] 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.

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 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.

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, however, SAFFIER is hard to beat for unbiasedness.

Table 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

5.3 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 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.

5.4 Conclusion on real time forecasts

Summarising 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.

When we looked at the performance of individual variables by calculating the accuracy of models containing each specific variable, there was no distinction between survey variables and 'traditional' economic variables. The only variable that performed

Table 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

marginally better than the others was the number of bankruptices for the current year.[14]

5.5 Testing the four hypotheses

5.5.1 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 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.

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 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.

5.5.2 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 functions, especially asymmetric loss functions. Given that we do not directly observe the loss functions of the users of our forecasts we must make

[14] Further details available on request.

Table 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

Table 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

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.

5.5.3 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. 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. 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 4 we can indeed see that the variation in forecasts of the 4 lag models is important.[15] 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 4.

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

5.6 Fit versus accuracy

Table 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.[16] All of the correlations 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

[15] 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.

[16] 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.

are shown in equations 5 and 6 below.

$$AIC = 2k - 2ln(L) \tag{5}$$

$$SBC = kln(n) - 2ln(L) \tag{6}$$

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. 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.[17] Even though individual models with better fit were slightly more accurate, this advantage disappears after pooling. We have seen similar results before, especially in tables 3 and 4, were the average accuracy of the individual models was poor, but after pooling they were relatively accurate.

We also found no correlation between the Quandt-Andrews structural break statistics and forecast accuracy[18], 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 1.

6 Conclusion

We find that the average accuracy of individual VAR based models is not better than CPB's 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. We also found no difference in accuracy between survey data and 'traditional' economic data.

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

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

[18] Details available on request

Table 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

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.

References

Gianna Boero. Comparing ex-ante forecasts from a SEM and VAR model: An application to the Italian economy. *Journal of Forecasting*, 9(1):13–24, 1990.

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

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

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

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

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

Oyvind Eitrheim, Tore Anders Husebo, and Ragnar Nymoen. Equilibrium-correction vs. differencing in macroeconometric forecasting. *Economic Modelling*, 16(4):515–544, December 1999.

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

R. Fildes and S. Makridakis. The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting. *International Statistical Review/Revue Internationale de Statistique*, 63(3):289–308, 1995.

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

David F. Hendry and Michael P. Clements. Economic forecasting: some lessons from recent research. *Economic Modelling*, 20(2):301–329, March 2003.

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

Jan P. A. M. Jacobs and Kenneth F. Wallis. 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, 2007.

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

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

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

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

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

R. B. Litterman. Forecasting with Bayesian Vector Autoregressions: Five Years of Experience. *Journal of Business and Economic Statistics*, 4(1):25–38, 1986.

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

S. Makridakis and M. Hibon. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476, 2000.

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

S. Makridakis, C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord, and LF Simmons. The M2-Competition: A Real-Time Judgmentally Based Forecasting Study. *International Journal of Forecasting*, 9(1):5–22, 1993.

John C. Robertson and Ellis W. Tallman. Vector autoregressions: forecasting and reality. *Economic Review*, (Q1):4–18, 1999.

Christopher A Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, January 1980.

Christopher A. Sims, James H. Stock, and Mark W. Watson. Inference in linear time series models with some unit roots. *Econometrica*, 58(1):113–144, jan 1990.

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

H. Theil and A. S. Goldberger. On Pure and Mixed Statistical Estimation in Economics. *International Economic Review*, 2(1):65–78, 1961.

Kenneth F. Wallis. Macroeconomic forecasting: A survey. *Economic Journal*, 99(394): 28–61, March 1989.