Model selection for forecast combination

Philip Hans Franses

Econometric Institute, Erasmus University Rotterdam

Econometric Institute Report 2008-11

Abstract

In this paper it is advocated to select a model only if it significantly contributes to the accuracy of a combined forecast. Using hold-out-data forecasts of individual models and of the combined forecast, a useful test for equal forecast accuracy can be designed. An illustration for real-time forecasts for GDP in the Netherlands shows its ease of use.

Key words: Forecast combination; Model selection

JEL code: C53

This version: June 01 2008

Author's notes

Many thanks are due to Dick van Dijk and Henk Kranendonk for their helpful comments and Clive Granger for useful discussions. I am particularly grateful to Philip Hubbard of Consensus Economics Inc for allowing me to use their GDP forecasts for the Netherlands. The address for correspondence is Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, franses@few.eur.nl

1. Introduction

Forecast combination is nowadays seen as a useful tool in practical forecasting; see Bates and Granger (1969) for the initial idea and see Clemen (1989) and Timmermann (2006) for surveys. What is usually done in practice is to line up a range of possibly suitable models, see if these models perform well enough in an evaluation sample, and use the forecasts from these models in a hold-out sample to see which combination of the forecasts is best for the yet unseen forecast sample.

One may wonder however if the in-sample evaluation is that much important, knowing that one will combine the forecasts anyway. Indeed, one may expect that irrelevant or inadequate models may turn up to have little or no weight in the combination, and that their inadequacy becomes apparent. The question is now whether looking at those weights, which are typically obtained through auxiliary least-squares based regressions[1], is informative enough. For example, it can happen that some weights are negative when forecasts are all on the same side of the true data points. The hold-out sample may also not be large enough to find significant weights. And, indeed, it may well be that even a small weight in the forecast combination can still be enough to establish better forecast performance.

In this paper I therefore recommend another model selection strategy, which is related to the notion of encompassing. That is, a model is selected in the final combination if the combination with that model yields more forecast accuracy than a combination without that model. For the sake of convenience I choose to look at mean squared prediction errors, but other criteria can be used as well. It is shown that, depending on the empirical setting, the test can be non-standard, and one then needs to follow the methodology outlined in Clark and McCracken (2001). An illustration to real-time forecasts for quarterly growth in Gross Domestic Product (GDP) in the Netherlands shows that useful conclusions can be drawn. A concluding section highlights a few issues for further research.

[1] Examples of regression-based selection methods to examine which models should be included in the combined forecast are discussed in Harvey and Newbold (2000) and Swanson and Zeng (2001).

2. The main idea

In this section, I outline the main idea of the proposed model selection criterion, and I discuss a proper test statistic.

Preliminaries

Consider a time-series variable *y* and assume for the moment that there are two linear regression models to explain the variation in *y*, that is,

(1) $\quad M_1: \quad y = X_1\beta_1 + \varepsilon_1$

and

(2) $\quad M_2: \quad y = X_2\beta_2 + \varepsilon_2$

where there are k_1 regressors in M_1 and k_2 regressors in M_2, each containing an intercept, and where parts of these regressors can overlap. For the moment it is assumed that the models are not nested, and below I will examine the consequence of relaxing this assumption. Further, the viewpoint is that an analyst starts from either M_1 or M_2 and that he or she wonders whether it is worthwhile to include the other model in a subsequent forecast combination.

Assume that the analyst has data for $N = R + P$ observations, of which the first R are used to estimate the parameters of M_1 and M_2, and the second set P is used to evaluate the quality of the one-step-ahead forecasts made from the two models. Eventually, a one-step-ahead forecast for $N + 1$ has to be made.

It is assumed that the analyst has the intention to use a combined forecast for $N +$ 1 and starting from $M_{1,}$ he or she wonders whether M_2 should be considered in that combination or the other way around. So, each model delivers a forecast for that observation at *N*+1, that is, there is a $\hat{y}_{1,N+1}$ from M$_1$ and a $\hat{y}_{2,N+1}$ from M$_2$, but in the end the analyst intends to consider

$$(3) \qquad \hat{y}_{c,N+1} = \alpha_1 \hat{y}_{1,N+1} + \alpha_2 \hat{y}_{2,N+1}$$

also because it is well known that combined forecasts usually perform better, see Clemen (1989) and Timmermann (2006).

The weights α_1 and α_2 are determined using the *P* hold-out observations. For that sample, the analyst has 2 times the *P* one-step-ahead forecasts, that is,

$$(4) \qquad \hat{y}_{1,R+1},\ldots,\hat{y}_{1,R+P} \qquad \text{and} \qquad \hat{y}_{2,R+1},\ldots,\hat{y}_{2,R+P} \quad ,$$

from models M$_1$ and M$_2$, respectively, with forecast errors

$$(5) \qquad \hat{\varepsilon}_{1,R+1},\ldots,\hat{\varepsilon}_{1,R+P} \qquad \text{and} \qquad \hat{\varepsilon}_{2,R+1},\ldots,\hat{\varepsilon}_{2,R+P}$$

The analyst uses the forecasts in (4) to estimate the weights α_1 and α_2 for

$$(6) \qquad \hat{y}_{c,t} = \hat{\alpha}_1 \hat{y}_{1,t} + \hat{\alpha}_2 \hat{y}_{2,t}$$

where the auxiliary regression to get these weights has *t* running from *R* + 1 to *R* + *P*. Note that in practice *P* may be quite small. This combined forecast has one-step-ahead forecast errors $\hat{\varepsilon}_{c,R+1},\ldots,\hat{\varepsilon}_{c,R+P}$. Timmermann (2006) gives a summary of useful methods to estimate these weights. When the variances of the one-step-ahead forecast errors are approximately equal, it often is found that the optimal weights are 0.5 and 0.5, or in general 1/K where K would be the number of models like in (1) and (2). For the moment, it suffices to assume that the analyst either fixes the weights a priori or (somehow) estimates them using least squares.

Finally, it is assumed that the analyst repeats this way of combining forecasts each time a new time-series observation becomes available. When this happens, the analyst can decide to keep the size of the hold-out sample *P* fixed, and increase the model estimation sample *R* to *R*+1. Upon doing so, the ratio

(7) $$\pi = \frac{P}{R}$$

approaches 0, see Clark and McCracken (2001) for the terminology, which will become more relevant later. Below I shall discuss what happens to the test below when this π does not approach 0.

Testing the relevance of a model

The main idea is that I would argue that a model should be selected and be included in the combined forecast if the final combined forecast is better off in terms of forecast accuracy with that particular model than without it.

To examine this, one could now consider the two encompassing regression models

(8) $$y_t = \beta_0 + \beta_1 \hat{y}_{c,t} + \beta_2 \hat{y}_{i,t} + \varepsilon_t$$

where *i* is either 1 or 2 to indicate one of the models, and where *t* runs from *R*+1 to *R*+*P*, and to see if β_2 =0. For example, when i = 1, and β_2 =0, then model M_2 significantly adds to the combined forecast's accuracy. However, simulations in Clark and McCracken (2001), among others, have shown that such encompassing tests do not have much power.

Therefore, I propose to evaluate the actual forecasts using the test proposed in Ericsson (1992), which has proved its usefulness, see again the simulations in Clark and McCracken (2001), that is, the *t*-ratio of α_i in

(9) $$\hat{\varepsilon}_{i,t} = \alpha_i (\hat{\varepsilon}_{i,t} - \hat{\varepsilon}_{c,t}) + \eta_t$$

When $\pi = 0$, Clark and McCracken (2001, pp. 91-92) show that this *t*-test has a standard normal distribution under the null hypothesis of equal forecast accuracy. As the alternative hypothesis is a one-sided hypothesis, that is, as the alternative is that the combined forecast is better, the 95% critical value of this test is 1.645.

When the null hypothesis is rejected for model M_i, then the accuracy of the combined forecast is higher than that of M_i alone and hence M_i can be improved by adding information from the other model. If the null hypothesis is not rejected, M_i's forecast is equally good as the combined forecast and the other forecast (and model) does not matter, and there is then also no need to include it in the combined forecast.

Variations

There are two directions in which the above approach may need to be expanded to meet relevant practical situations. The first is that $\pi \neq 0$ and the second is that there are $K > 2$ models instead of just 2.

When $\pi \neq 0$ this means that if new data become available, that then *P* and *R* both increase such that π approximately approaches a fixed constant. In that case it also becomes important to recognize that the combined forecast nests the individual forecasts, see Clark and McCracken (2001). In Clark and McCracken (2001) it is shown that the *t*-ratio for α_i in (9) then has no standard normal distribution anymore in the case of nested forecast schemes. Critical values for the cases where π is 0.1, 0.2, 0.4, 1.0, 2.0, 3.0 and 5.0 are given in the first panel of their Table 1 (page 92).

The combined forecast can cover K models. Assuming that one looks at the situation where an earlier combination covering K-1 models can be improved with one additional model, the degree of nesting is 1.

3. An illustration

To illustrate this proposal for model selection, and also to illustrate how results can be interpreted, I consider the real-time forecasts in Table 1, and the realizations as they are presently known (May 26, 2008). I also report the flash values that were published six weeks after the relevant quarter, that is, the first-release data.

3.1 Currently available GDP data

The data concern annual growth rates of Netherlands' GDP, when predicted and observed for the quarters 2004Q4 to and including 2007Q4. This is not very large sample, and this is due to the fact that there are only two real-time forecasts available for the Netherlands, that is, the forecasts in the last two columns were published in the very same quarters as they concern.

Insert Table 1 about here

The second column of Table 1 contains the currently known GDP growth rates (computed as log(y)-log(y_{-4}) where y is GDP in billions of euros), as they are published by Statistics Netherlands (CBS). The fourth column contains the real-time forecasts made by Consensus Economics Inc, a commercial London UK-based company that publishes real-time forecasts for many countries around the world, including the Netherlands. These forecasts are based on weighted expert opinions, and details of the procedure are given on the website www.consensuseconomics.com. The final column contains the real-time forecasts created using the methodology in De Groot and Franses (2005). This so-called Econometric Institute Current Indicator of the Economy (EICIE) is four-weekly published in the (Dutch language) *Economische Statistische Berichten*. The graphs in Figure 1 show that the two forecasts seem to follow the actual GDP values reasonably well. Some observations are predicted rather well, but sometimes the fit is poor.

Insert Figure 1 about here

Before continuing one needs to think for a moment about the possible value of π. The real-time forecasts in the EICIE are based on a regression of GDP (after suitable transformation) on current and lagged growth rates of employees in the temporary staffing sector and on its own past This autoregressive distributed lag model, when estimated in error correction format to incorporate a cointegration relation, is re-estimated each time a new observation becomes available. Hence, the size of P is 1, and R increases each time with 1, which makes π to approach 0. It is not evident whether the value of π is also approximately equal to 0 for the Consensus forecasts, but it seems a reasonable assumption too.

Insert Table 2 about here

In Table 2 I give the one-step-ahead (or better: current) forecast errors for the two forecasts in columns 4 (Consensus) and 5 (EICIE) of Table 1. The mean error, median error and mean squared prediction error of Consensus are smaller than those of the EICIE, and the EICIE has a smaller median squared prediction error. Based on these two individual track records, one would be inclined to favor the Consensus forecasts.

As said, it is assumed that one aims at considering a combined forecast, so now it matters if either the Consensus forecasts or the EICIE forecasts can be improved by including the other. Let us first look at a combination based on equal (0.5) weights. The forecast errors of this new forecast are displayed in the fourth column of Table 2. As expected, the mean and median errors of this equal weight forecast combination are in between those of the two forecasts. The mean squared prediction error (1.01) is closer but larger than that (0.93) of the Consensus forecast, while the median squared forecast error (0.25) is smaller than each of its components. Hence, equal weights do give some improvement, but not that much.

This result is emphasized by the outcomes of the test regression in (9). The first regression assumes that the Consensus forecasts are the starting point and it looks at whether it can be improved by incorporating the EICIE forecasts in an equal-weight combination. It reads as

$$(10a) \qquad \hat{\varepsilon}_{consensus,t} = \mu + \alpha(\hat{\varepsilon}_{consensus,t} - \hat{\varepsilon}_{combined-equalweights,t}) + \eta_t$$

and this regression gives a t-ratio for α of 1.139, while for the test regression

$$(10b) \qquad \hat{\varepsilon}_{eicie,t} = \mu + \alpha(\hat{\varepsilon}_{eicie,t} - \hat{\varepsilon}_{combined-equalweights,t}) + \eta_t$$

one gets a t-ratio of 2.183. From (10b) it can thus be learned that the EICIE can be improved by including the Consensus forecast in the combination, while (10a) tells us that this equal-weights combination is not significantly better than the Consensus forecast already is.

One possible reason for the above finding is that a combined forecast based on other than equal weights is perhaps better. When I regress the CBS data on an intercept, the Consensus and EICIE forecasts I get as a combined forecast

1.257 + 0.429 Consensus + 0.215 EICIE

with an R^2 value of 0.541. The t-ratios of these parameters do not indicate significance, which is of course due to the small sample size of just 13 observations. This emphasizes the problems of interpreting the t-ratios of parameters in combining regressions, and it reiterates the very reason to look at test statistics like those in (10a) and (10b). Note that the estimated intercept takes a large positive value, and this reflects the commonly found phenomenon that later vintages of data typically move upwards, relative to the first release (flash) values. When the intercept is not included, the combined forecast is

0.851 Consensus + 0.285 EICIE

where the parameter for the Consensus forecast is significant at the 5% level, while that of the EICIE is not.

The penultimate column of Table 2 gives the forecast errors of this least-squares-weights-combined forecast with an intercept included, while the last column gives it for the case without an intercept. Clearly, this combined forecast outperforms its constituents by far, and especially the mean squared prediction error is reduced substantially.

To verify if each of the two models is contributing significantly or whether one of the component forecasts is better on its own, I run the regression for the case of the intercept included and get

$$\text{(11a)} \qquad \hat{\varepsilon}_{consensus,t} = \mu + \alpha(\hat{\varepsilon}_{consensus,t} - \hat{\varepsilon}_{combined-leasstsquaresweights,t}) + \eta_t$$

with a t-ratio for α equal to 2.162, while the t-ratio for α in

$$\text{(11b)} \qquad \hat{\varepsilon}_{eicie,t} = \mu + \alpha(\hat{\varepsilon}_{eicie,t} - \hat{\varepsilon}_{combined-leasstsquaresweights,t}) + \eta_t$$

is equal to 3.016. Hence, now both models are relevant for the combined forecast, albeit that the EICIE needs less weight than the Consensus forecast in this final combined forecast. The conclusion here is that both individual real time forecasts for Netherlands GDP can be improved by combining them with the other in a linear combination which has no equal weights. In case the least-squares-weights combined forecast does not include an intercept, the t-ratios for α parameters in (11a) and (11b) are -0.107 and 1.618, respectively, which shows that the intercept was needed indeed.

Overall, these results show that for the final value of real GDP growth a combination of the Consensus and EICIE forecasts can be beneficial, where the contribution of the EICIE is relatively smaller than that of the Consensus. Both forecasts seem to underestimate the final value, and hence the combination requires a non-zero intercept to accommodate for this.

3.2 First-release GDP data

To see if the results for the final data carry trough to the first-release data, I consider the flash values of GDP. The third column of Table 1 gives the first-release growth rates, as they were published about six weeks after the end of the relevant quarter. Comparing the numbers in columns 2 and 3 of Table 1, one can see that there can be substantial differences between first-release and the currently seen as "final" values. This can also be observed from the graphs in Figures 2 and 3.

Insert Figures 2 and 3 about here

Insert Table 3 about here

Columns 2 and 3 of Table 3 show that both the Consensus and EICIE forecasts do much better on the flash data, where again the Consensus forecasts outperform.

To see if either the Consensus forecasts or the EICIE forecasts can be improved by including the other, let us again look at a combination based on equal (0.5) weights. The forecast errors of this new forecast are displayed in the column "Equal weights" of Table 3. As expected, the mean and median errors of this equal weight forecast combination are in between those of the two forecasts. The mean squared prediction error (0.66) is closer but larger than that (0.56) of the Consensus forecast, while the median squared forecast error (0.06) is substantially smaller than each of its components. Hence, equal weights certainly do give some improvement.

The first test regression assumes that the Consensus forecasts are the starting point and it looks at whether it can be improved by incorporating the EICIE forecasts. It reads as (10a) and it gives a t-ratio for α of 0.496, while for the test regression (10b) one gets a t-ratio of 3.011. From (10b) it can thus be leaned that the EICIE can be improved by including the Consensus forecast in the combination, while (10a) tells us that this equal-weights combination is not significantly better than the Consensus forecast.

When I regress the CBS data on an intercept, the Consensus and EICIE forecasts I get as a combined forecast

0.359 + 0.779 Consensus + 0.090 EICIE

with an R^2 value of 0.649, where now the intercept is considerably smaller than before, and it is now found not statistically significant. When the intercept is not included, the combined forecast is

0.891 Consensus + 0.110 EICIE

where the parameter for the Consensus forecast is significant at the 5% level, while that of the EICIE is not. Note that the sum of the parameters is about equal to 1.

The penultimate column of Table 3 gives the forecast errors of this least-squares-weights-combined forecast with an intercept included, while the last column gives it for the case without an intercept. Clearly, this combined forecast outperforms its constituents by far, and especially the median squared prediction error is reduced substantially.

Running regression (11a) for the case of the intercept included I get a t-ratio for α equal to 0.837, while the t-ratio for α in (11b) is equal to 3.143. In case the least-squares-weights combined forecast does not include an intercept, the t-ratios for α parameters in (11a) and (11b) are 0.488 and 3.009, respectively.

These results show that for the flash value of real GDP growth a combination of the Consensus and EICIE forecasts is not beneficial, as the contribution of the EICIE is not significant relative to that of the Consensus. Hence there only the Consensus forecasts will do. As we saw, for the revised ("final") GDP figures, the EICIE does seem to be relevant for the final combination.

4. Consequences

This paper has put forward a simple methodology to see if forecasts from models can be significantly improved by combining them with forecasts from other models. It has a single model or perhaps already a combination of K-1 forecasts as the starting points, and it can be used in case one wonders whether a further combination with yet a new model can yield even better forecasts.

Looking at the use of model forecasts this way, in-sample model diagnostics have become less relevant. What is needed is a set of consistent forecasts from 2 or K models. Moreover, forecasts in the past do not need to be accurate. Individual track records are no guarantee that the models will successfully contribute to the combined forecasts, so if the intention is to consider combined forecasts anyway, only studying in-sample performance becomes obsolete.

Future work on the issue of this paper could include model selection for the combination of multi-step-ahead forecasts. Also, many more examples would be needed to illustrate the merits.

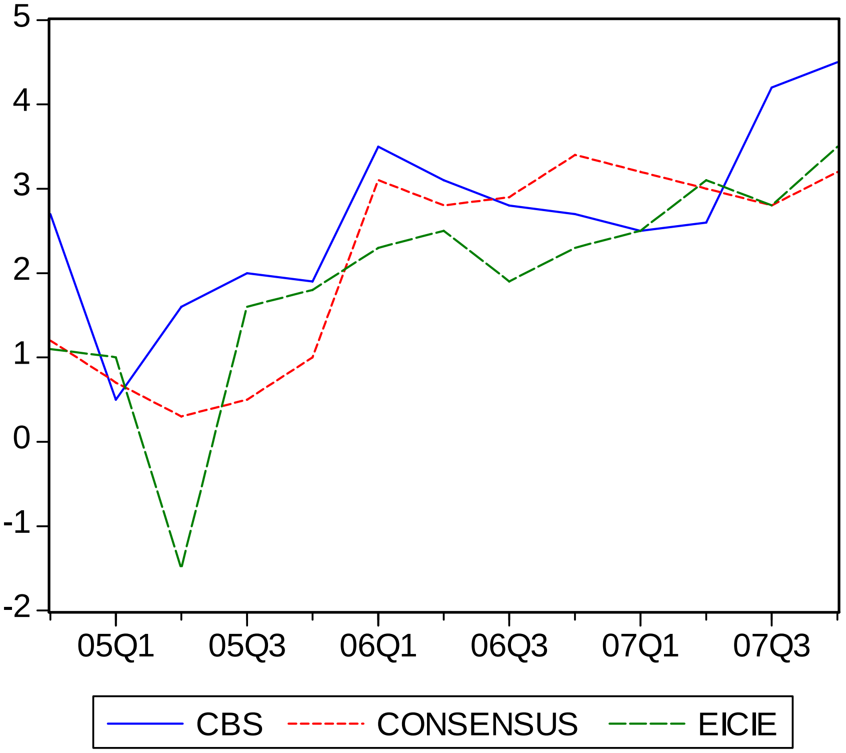

Figure 1: The data from Table 1 (Final data)

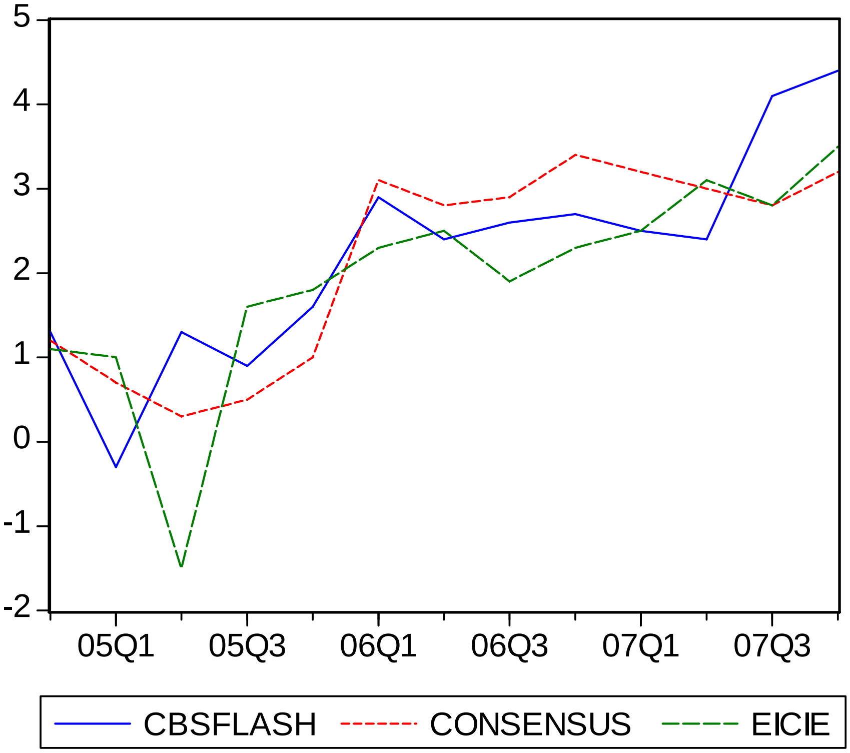

Figure 2: The data from Table 1 (First release data)

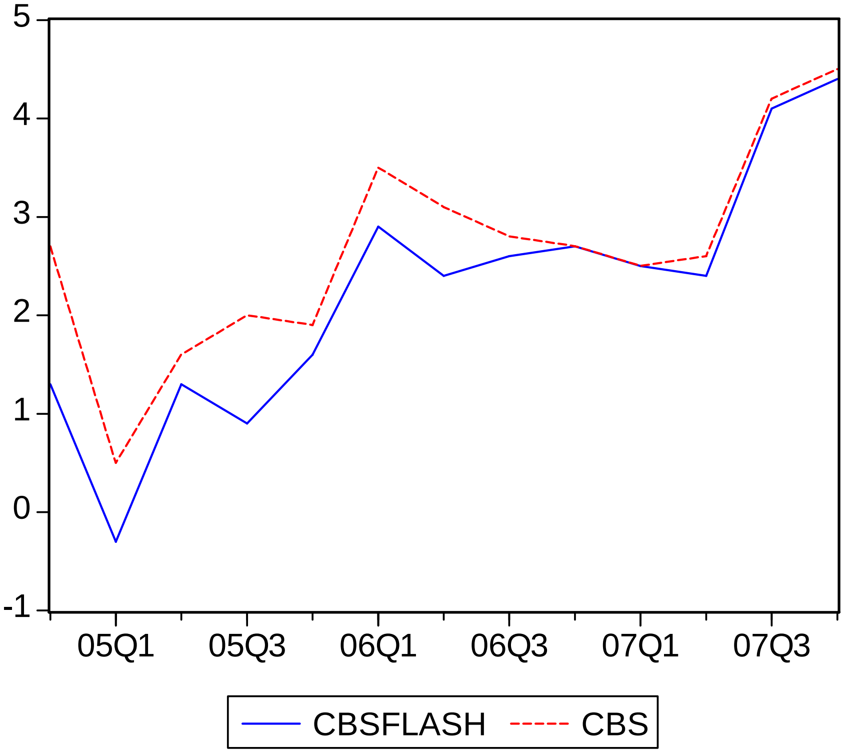

Figure 3: the two CBS series (First release data and final data, as of May 26 2008)

Table 1:
The data used in the illustrations

			Forecasts	
Quarter	CBS[1]	CBS[2]	Consensus[3]	EICIE[4]
2004Q4	2.7	1.3	1.2	1.1
2005Q1	0.5	-0.3	0.7	1.0
2005Q2	1.6	1.3	0.3	-1.5
2005Q3	2.0	0.9	0.5	1.6
2005Q4	1.9	1.6	1.0	1.8
2006Q1	3.5	2.9	3.1	2.3
2006Q2	3.1	2.4	2.8	2.5
2006Q3	2.8	2.6	2.9	1.9
2006Q4	2.7	2.7	3.4	2.3
2007Q1	2.5	2.5	3.2	2.5
2007Q2	2.6	2.4	3.0	3.1
2007Q3	4.2	4.1	2.8	2.8
2007Q4	4.5	4.4	3.2	3.5

[1] This column gives the data on annual growth rates of GDP in the Netherlands, as they are published by the Central Bureau of Statistics, and as they are published on the website www.cbs.nl on April 22 2008.

[2] This column gives the data on annual growth rates of GDP in the Netherlands, as they are published by the Central Bureau of Statistics, and as they are published on the website www.cbs.nl six weeks after the end of the relevant quarter.

[3] This column gives the forecasts as they are created by Consensus Economics Inc., using the method outlined on their website www.consensuseconomics.com

4 This column gives the quotes of the Econometric Institute Current Indicator of the Economy, as they are published in the (Dutch language) *Economische Statistische Berichten*, and on the website www.esbonline.nl

Table 2:

Forecast errors corresponding to individual forecasting models and to forecast combinations (final CBS values, as quoted on May 26 2008).

	Forecasts		Combined forecasts, weights		
Quarter	Consensus	EICIE	Equal	LS, intercept	
				With	Without
2004Q4	1.5	1.6	1.55	0.68	1.37
2005Q1	-0.2	-0.5	-0.35	-1.27	-0.38
2005Q2	1.3	3.1	2.20	0.55	1.77
2005Q3	1.5	0.4	0.95	0.19	1.12
2005Q4	0.9	0.1	0.50	-0.18	0.54
2006Q1	0.4	1.2	0.80	0.35	0.21
2006Q2	0.3	0.6	0.45	0.04	0.00
2006Q3	-0.1	0.9	0.40	-0.18	-0.21
2006Q4	-0.7	0.4	-0.15	-0.59	-0.85
2007Q1	-0.7	0.0	-0.35	-0.74	-0.94
2007Q2	-0.4	-0.5	-0.45	-0.68	-0.84
2007Q3	1.4	1.4	1.40	1.08	1.02
2007Q4	1.3	1.0	1.15	1.04	0.78
Mean error	0.50	0.75	0.62	0.02	0.28
Median error	0.40	0.60	0.50	0.04	0.21
Mean SPE	0.93	1.43	1.01	0.48	0.82
Median SPE	0.49	0.36	0.25	0.38	0.70

Table 3:

Forecast errors corresponding to individual forecasting models and to forecast combinations (flash CBS values).

	Forecasts		Combined forecasts, weights		
Quarter	Consensus	EICIE	Equal	LS, intercept	
				With	Without
2004Q4	0.1	0.2	0.15	-0.09	0.11
2005Q1	-1.0	-1.3	-1.15	-1.29	-1.03
2005Q2	1.0	2.8	1.90	0.84	1.20
2005Q3	0.4	-0.7	-0.15	0.01	0.28
2005Q4	0.6	-0.2	0.20	0.30	0.51
2006Q1	-0.2	0.6	0.20	-0.08	-0.11
2006Q2	-0.4	-0.1	-0.25	-0.37	-0.37
2006Q3	-0.3	0.7	0.20	-0.19	-0.19
2006Q4	-0.7	0.4	-0.15	-0.51	-0.58
2007Q1	-0.7	0.0	-0.35	-0.58	-0.63
2007Q2	-0.6	-0.7	-0.65	-0.58	-0.61
2007Q3	1.3	1.3	1.30	1.31	1.30
2007Q4	1.2	0.9	1.05	1.23	1.16
Mean error	0.05	0.30	0.18	0.00	0.08
Median error	-0.20	0.20	0.15	-0.09	-0.11
Mean SPE	0.56	1.09	0.66	0.52	0.55
Median SPE	0.36	0.49	0.06	0.27	0.34

References

De Groot, Bert and Philip Hans Franses (2005), Real time estimates of GDP growth, Econometric Institute Report 2005-01, Erasmus University Rotterdam.

Bates, J.M. and C.W.J. Granger (1969), The combination of forecasts, *Operations Research Quarterly,* 20, 451-468.

Clark, Todd E. and Michael W. McCracken (2001), Tests of equal forecast accuracy and encompassing for nested models, Journal of Econometrics, 105, 85-110

Clemen, Robert T. (1989), Combining forecasts: A review and annotated bibliography (with discussion), *International Journal of Forecasting 5*, 559-583.

Ericsson, Neil. R.(1992)., Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions and illustration, *Journal of Policy Modeling*, 14, 465-495.

Harvey, David and Paul Newbold (2000), Tests for multiple forecast encompassing, *Journal of Applied Econometrics,* 15, 471-482

Swanson, Norman. R. and Tian Zeng (2001), Choosing among competing econometric forecasts: Regression-based forecast combination using model selection, *Journal of Forecasting*, 20, 425-440

Timmermann, Allan (2006), Forecast combinations, Chapter 4 in Graham Elliott, Clive W.J. Granger and Allan Timmermann (eds.), *Handbook of Economic Forecasting Volume I*, Amsterdam: Elsevier, 135-196.