Decomposing bias in expert forecasts

Philip Hans Franses

Econometric Institute
Erasmus School of Economics

Econometric Institute Report 2010-26

Abstract

Forecasts in the airline industry are often based in part on statistical models but mostly on expert judgment. It is frequently documented in the forecasting literature that expert forecasts are biased but that their accuracy is higher than model forecasts. If an expert forecast can be approximated by the weighted sum of a part that can be replicated by an analyst and a non-replicable part containing managerial intuition, the question arises which of two causes the bias. This paper advocates a simple regression-based strategy to decompose bias in expert forecasts. An illustration of the method to a unique database on airline revenues shows how it can be used to improve their experts' forecasts.

Key words: Expert forecasts, Forecast bias, Airline revenues

This version: April 2010

Author's notes
Many thanks are due to Pieter Bootsma and Stefan ten Haaf (KLM Royal Dutch Airlines) for providing the data. Address for correspondence is Econometric Institute, Erasmus University Rotterdam, PO Box 1738, NL-3000 DR Rotterdam, The Netherlands, franses@ese.eur.nl

1. Introduction

This paper considers a situation that is very common in forecasting practice (also in the airline industry), as is recently evidenced by a stream of literature on expert adjustment of forecasts; see for example the recent study of Fildes, et al. (2009). This situation concerns a modeler who creates a forecast using a statistical or econometric model. Usually, the input to this model is publicly available, at least, within the forecast setting (the airline has revenues data, and there are data from government agencies). The statistical forecast is however not directly communicated to management or policy makers, as it may happen that finally an expert or manager quotes an alternative forecast. This expert forecast may have had the model forecast as input, but this is rarely known. It may be that the expert makes his own model, based on another set of publicly available variables, but it can also be that she fully dismisses the model forecast, see Boulaksil and Franses (2009). The third person in this situation is the analyst, who needs to evaluate the quality of the model forecast and that of the expert forecast, if available. This analyst also has access to publicly available data, but he has no access whatsoever to the managerial intuition which makes the expert forecast to differ from the model forecast.

In this paper the following variant of the above situation is studied. Suppose the analyst does not have the statistical model forecasts, but does have information on the potential variables that could have been included in such a model. Further, the analyst has expert forecasts, and he wants to analyze their quality. This situation is quite common in airline revenues forecasting, as will be indicated below in the empirical illustration.

It is well documented in the relevant literature that expert forecasts can be biased; see Mathews and Diamantopoulos (1989), Lawrence, O'Connor and Edmundson (2000) and Fildes, et al. (2009). It can also happen, for example by selecting the wrong publicly available variables, that the model forecasts are biased and that the added managerial intuition makes the final forecast unbiased. This paper gives a simple methodology to decompose potential bias or potential expertise in expert forecasts. Section 2 discusses the methodology to do so. Section 3 illustrates it to a unique and detailed database concerning monthly revenues of various regions for KLM Royal Dutch Airlines. Section 4 concludes with a few ideas for further research.

2. Methodology

Consider the availability of actual airline revenues data y_t and consider forecasts for these data made at time t – 1, t – 2, …, t – h denoted as $F_{t|t-h}$. The commonly applied test regression to diagnose bias in these forecasts is

$$y_t = \alpha + \beta F_{t|t-h} + u_t \tag{1}$$

The null hypothesis of no bias corresponds with

$$\alpha = 0, \beta = 1 \tag{2}$$

in (1). In this paper it is assumed that $F_{t|t-h}$ is created by an expert, who may have used a statistical model and/or publicly available variables, but it is unknown how she has done that.

2.1 Decomposing expert forecasts

If the analyst wants to understand what causes bias or success of expert forecasts, then he somehow needs to make assumptions about that expert forecast. Franses, McAleer and Legerstee (2009) recommend to assume that an expert forecast can be decomposed into a replicable part ($F^*_{t|t-h}$) and a non-replicable part e_t, that is,

$$F_{t|t-h} = F^*_{t|t-h} + e_t \tag{3}$$

The non-replicable part can be called the latent (is: unobserved) managerial intuition. The replicable part is the part that the analyst can create. In fact, it is assumed that the replicable part can be approximated by the conditional expectation based on the regression model

$$F^*_{t|t-h} = W_{t-h-1}\delta + \varepsilon_t \tag{4}$$

where W_{t-h-1} contains all kinds of variables that are publicly known at time $t - h$, that is the time when the expert makes the forecast. Note that this excludes information at time $t - h$ itself. The analyst can use (4) to approximate the replicable part. Note that W_{t-h-1} also includes an intercept. Of course, the analyst does not know exactly whether it were these variables that were used by the expert, so (4) is only an approximation.

The relevant test regression for forecast bias in (1) now becomes

$$y_t = \alpha + \beta \hat{F}^*_{t|t-h} + u_t \tag{5}$$

where $\hat{F}^*_{t|t-h}$ follows from applying Ordinary Least Squares [OLS] to (4). As the regressor in (5) is a generated regressor, one needs to use the Newey-West HAC standard errors.

2.2 Strategy

The key idea of the methodology proposed in this paper is to compare the outcomes of the tests for regression models (1) and (5). This means that four distinct situations can occur in practice.

The first is the case where the null hypothesis is rejected both for (1) and (5). This means that the expert forecast is biased and the forecast based on replicable expertise is also biased. This means that the analyst has used the wrong model. Assuming that the expert would not have been able to find better variables, the expert has also used the wrong model, and her managerial intuition could not improve that. This case would be the worst case in practice, as both the analyst and expert make biased forecasts.

The second case is where the expert forecast is biased, but where the replicable forecast is not. This means that the analyst can come up with an unbiased forecast, while it is the added managerial intuition that causes the bias. In other words, experts have replicable expertise, but the added intuition creates problems. It could be that this is done on purpose. For example, the expert sees from the model forecasts that, say, sales go up very steeply or down very rapidly, and she decides to spread forecasts on these rapid changes over a few adjacent months.

The third case is where the null hypothesis is not rejected for (1), but it is for (5). This means that the expert forecast is not biased, but that the replicable forecast is. Hence, the expert may have used the wrong model, as the analyst indicates, but she has sufficient intuition to render the final forecast as unbiased.

Finally, the fourth case entails that the expert forecast and the replicable part of that expert forecast are both unbiased, which means that managerial intuition is not adding much.

Note that the focus here is on bias and not on accuracy. It may well be that biased expert forecasts have much smaller root mean squared prediction errors (RMSPE), as we will see below. On the other hand, one would favor the fourth case, when also the expert forecasts would have smaller RMSPE. Indeed, it is important to understand the causes of bias in expert forecasts, as otherwise one cannot learn from mistakes.

3. Illustration

To illustrate the methodology proposed in the previous section, I rely on a unique database. It contains the monthly airline revenues data spanning April 2004 to and including December 2008 for KLM Royal Dutch Airlines for seven distinct regions and for the world. The regions are Europe, Middle East, Africa, North America, Middle and South America, Asia Pacific and India. There are forecasts for these revenue data made 3 months ahead, 2 months and 1 month ahead. For one month (April 2006) the data are missing. The forecasts are all made by experts, who base their final forecasts on input from model forecasts, but to what extent they do so is unknown. That is, statistical model forecasts are not available.

Insert Figure 1 about here

In Figure 1 I display the total revenues. Clearly there is a general upward trend, with a slight decrease at the end of the sample, corresponding with the worldwide economic crisis that started in 2008. Also, one can see a strong seasonal pattern in the data. Further, all forecasts seem to be rather accurate, with slightly larger forecast errors towards the end of the sample.

Insert Table 1 about here

Table 1 reports the estimation results for regression model (1), where the data have been transformed by taking natural logarithms. This is done because the model to be made by the analyst below will also incorporate log-transformed data, as is usual practice. Unreported results on (1) for untransformed series show qualitatively the same results. Comparing the estimates with the standard errors, it is clear that the estimates for α are slightly different from 0, while the estimates for β are slightly different from 1. The Wald test values in the final column of Table 1 indeed suggest that there is some evidence (at the 10% level) of bias in these expert forecasts.

Insert Table 2 about here

Table 2 tells us that the revenues forecasts for Europe and Middle and South America are all unbiased. In contrast, substantial bias for all three horizons is found for the Middle East and India. Forecasts for North America show bias for the horizons 3 and 1, which is a bit odd, as one might wish to have lesser bias with shorter horizons, like it appears to be the case for most expert forecasts. Finally, two-month-ahead forecasts for Africa also show bias. In sum, out of the 7 times 3 cases, 11 show bias.

Insert Tables 3 and 4 about here

Table 3 learns that the bias in the total sales forecasts can be associated with skewed distributions of the forecast errors, that is, the expert forecasts tend to be too high. The negative skewness becomes smaller when the forecast horizon decreases, a result that one would wish to see in practice. The absolute value of the minimum forecast errors is larger than that of the maximum forecast errors, showing that not only the forecasts can be more often too high, also when they are higher they are more so than when they are lower. These results are amplified in Table 4, where the same holds for (almost) all seven regions. In 18 of the 21 cases, the skewness is negative. For Europe, Africa, and Asia Pacific we see a strong and steady decrease in skewness from horizons 3 to 1, but for other regions this is not the case.

A closer look at the forecast errors in the seven regions (unreported due to confidentiality reasons) shows that the high expert forecasts typically occur in 2008, when revenues go down in all areas of the world. One might now be inclined to believe that econometric models would show the same bias, but this is unlikely the case. Indeed, if the analyst has the proper input variables in (4), then unbiased forecasts should follow from the replicable part of the expert forecast.

Insert Table 5 about here

To model the replicable part of the expert forecast, the analyst (who this time is the author of the current article) decides to include in W_{t-h-1} an intercept and the following variables, that is, the harmonic regressors $\cos\frac{2\pi t}{12}$, $\sin\frac{2\pi t}{12}$, Exchange rate Dollar versus Euro (at time t - h-1) ("Dollar/Euro$_{t-h-1}$"), Natural log of USA Industrial Production Index (at time t - h-1) ("IP_USA$_{t-h-1}$"), Natural log of oil price (West Texas crude) (at time t - h-1) ("Oil price$_{t-h-1}$") and Unemployment rate in the USA (at time t - h-1) ("Unemployment$_{t-h-1}$"). The OLS estimation results are given in Table 5, and it is evident that the three models (one for each horizon) all have quite a large value of the R^2, meaning that a substantial fraction of the expert forecast can be replicated. The harmonic regressors are very relevant, but also the Wald tests for the joint significance of the other four economic variables suggest that these also strongly contribute to the fit.

It may now be interesting to examine if the forecasts updates (from horizon 3 to 2 and from horizon 3 to 1) can also be replicated to some extent. For that purpose, I regress $\log F_{t-2} - \log F_{t-3}$ on Dollar/Euro$_{t\text{-}3}$, IP_USA$_{t\text{-}3}$, Oil price$_{t\text{-}3}$, and Unemployment$_{t\text{-}3}$ and on Dollar/Euro$_{t\text{-}4}$ IP_USA$_{t\text{-}4}$ Oil price$_{t\text{-}4}$, and Unemployment$_{t\text{-}4}$. The R^2 is 0.308 and the F-test is 2.238 (p-value is 0.040). A regression of $\log F_{t-1} - \log F_{t-3}$ on the same variables and on Dollar/Euro$_{t\text{-}2}$ IP_USA$_{t\text{-}2}$ Oil price$_{t\text{-}2}$ and Unemployment$_{t\text{-}2}$ gives an R^2 is 0.620 and an F test of 5.037 (p-value is < 0.001). So it seems that the analyst can demonstrate that the experts' updates of the forecasts are also based on replicable expertise.

Insert Table 6 about here

Table 6 is concerned with the test regression (5), where the fit of the models in Table 5 is included. Clearly, the replicable components of the expert forecasts are now found to be unbiased. Hence, here we have encountered the second case mentioned in Section 2.2, that is, the expert forecast is biased, but the replicable forecast is not. This means that the analyst (and thus also the expert) can come up with an unbiased forecast, and it is the added managerial intuition that causes the bias.

Insert Table 7 about here

Given that the expert forecasts show bias, and the replicable component not, would it than be better to only use the model forecast? The results in Table 7 strongly suggest no. The final expert forecasts are much better than the model forecasts, in terms of Root Mean Squared Prediction Error. So, the added non-replicable managerial intuition certainly improves the forecast accuracy.

In sum, for total revenues the experts of KLM give very much better forecasts than that an analyst's model can replicate. Of course, the model could involve more variables. But then still, the model does not give biased forecasts. Also in times of downturns, models keep on giving unbiased forecasts, in contrast to the expert forecasts.

4. Conclusion

In many practical cases in the airline industry, an analyst has to evaluate the quality of expert forecasts, when model forecasts are unavailable, but where it is quite certain that the experts have looked at various variables as input for their forecasts. The part of the expert forecast that cannot be replicated by the analyst can be called the managerial intuition. As it is known that expert forecasts can be biased, it is of interest to examine whether this is due to the variables that were incorporated or to the intuition. This paper introduces a simple methodology to learn how bias in expert forecasts can be understood.

Application to a unique database concerning airline revenue forecasts shows, at least for this case, that the experts of KLM have a tendency to quote too high a forecasts, most likely in times of a downturn. It is shown that econometric models with properly chosen explanatory variables do not have such a tendency. On the other hand, the expert forecasts are much better than the model forecasts, so there is a clear trade-off between accuracy and bias.

Ideally, expert forecasts should contain two components, that is, a replicable part that delivers unbiased forecasts and an intuition part that does not make the final forecast biased, but only better. One way to achieve this is to make the expert to document the replicable part, so that the expert and the analyst can communicate on that part first. Are the proper variables included? Is the estimation routine carried out carefully? From this, her intuition follows from the difference between the expert forecast and the replicable part. In the terms used above, one would want to have $\hat{F}^*_{t|t-h}$ and then get an impression of e_t . The key reason for that is that without knowledge of these two components, one can never learn from past mistakes, nor can one attribute the success of an expert forecast to her intuition. The latter issue is important as by now there is substantial evidence that published forecasts in economics and business frequently involve an expert’s touch to a model forecast.

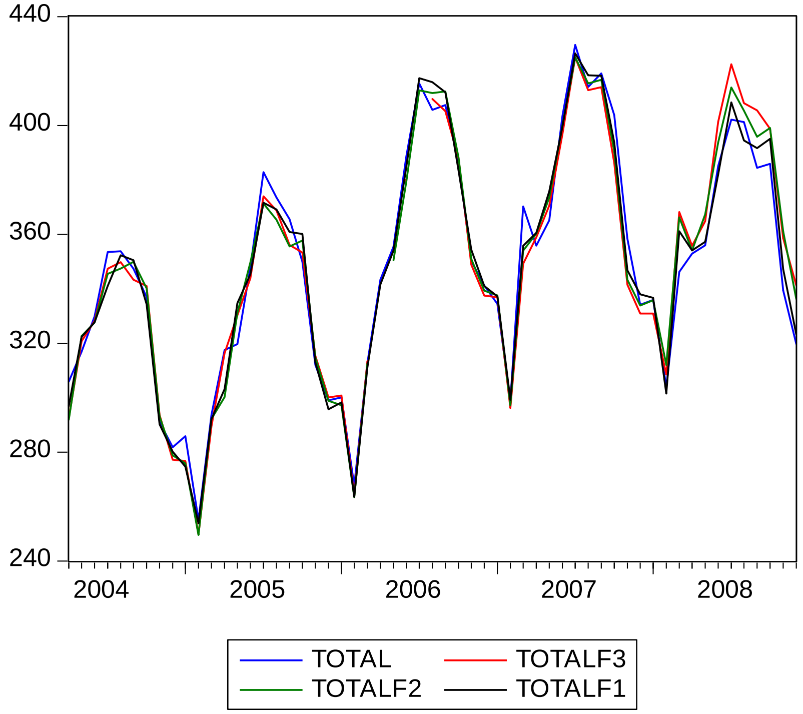

Figure 2:
Total airline revenues and forecasts created 3 months, 2 months and 1 month ago.

Table 1:

Testing for bias in expert forecasts, total revenues (standard error in parentheses)

Horizon	$\hat{\alpha}$	$\hat{\beta}$	Wald test (p-value)
3 months	0.338 (0.164)	0.942 (0.028)	4.285 (0.117)
2 months	0.328 (0.150)	0.944 (0.026)	4.766 (0.092)
1 month	0.245 (0.117)	0.958 (0.020)	4.804 (0.091)

Test regression is:

$$\log y_t = \alpha + \beta \log F_{t|t-h} + u_t$$

and the Wald test concerns the null hypothesis that $\alpha = 0, \beta = 1$. The sample size is 54, 56 and 57, for forecast horizons 3, 2 and 1, respectively.

Table 2:

Testing for bias in expert forecasts, revenues data for seven regions (standard error in parentheses)

Region	Horizon	$\hat{\alpha}$		$\hat{\beta}$		Wald test (p-value)	
Europe	3	0.135	(0.136)	0.970	(0.029)	2.094	(0.351)
	2	0.171	(0.125)	0.963	(0.026)	3.459	(0.177)
	1	0.156	(0.093)	0.967	(0.020)	2.998	(0.223)
Middle	3	0.939	(0.233)	0.694	(0.075)	17.76	**(0.000)**
East	2	0.789	(0.208)	0.743	(0.067)	15.20	**(0.001)**
	1	0.576	(0.173)	0.813	(0.056)	11.37	**(0.003)**
Africa	3	0.213	(0.131)	0.944	(0.035)	4.012	(0.135)
	2	0.307	(0.123)	0.919	(0.033)	10.18	**(0.006)**
	1	0.072	(0.111)	0.982	(0.030)	1.546	(0.462)
North	3	0.323	(0.118)	0.924	(0.028)	10.12	**(0.006)**
America	2	0.256	(0.116)	0.939	(0.028)	5.470	(0.065)
	1	0.283	(0.109)	0.933	(0.026)	8.512	**(0.014)**
Middle	3	0.057	(0.137)	0.986	(0.039)	1.710	(0.425)
South	2	0.042	(0.135)	0.991	(0.038)	1.537	(0.464)
America	1	-0.050	(0.115)	1.017	(0.033)	2.351	(0.309)
Asia	3	0.697	(0.191)	0.833	(0.045)	14.13	**(0.001)**
Pacific	2	0.559	(0.162)	0.867	(0.038)	12.32	**(0.002)**
	1	0.304	(0.126)	0.928	(0.030)	5.804	(0.055)
India	3	0.300	(0.095)	0.835	(0.049)	13.15	**(0.001)**
	2	0.273	(0.082)	0.848	(0.042)	16.15	**(0.000)**
	1	0.251	(0.080)	0.860	(0.042)	14.23	**(0.000)**

Test regression is:

$$\log y_t = \alpha + \beta \log F_{t|t-h} + u_t$$

and the Wald test concerns the null hypothesis that $\alpha = 0, \beta = 1$. The sample size is 54, 56 and 57, for forecast horizons 3, 2 and 1, respectively.

Table 3:

Properties of forecast errors $y_t - F_{t|t-h}$, total revenues data

Horizon	Mean	Minimum	Maximum	Skewness	Kurtosis
3 months	-0.352	-22.00	20.99	-0.493	3.345
2 months	-0.281	-21.70	17.34	-0.180	2.891
1 month	0.424	-15.03	14.46	-0.050	3.000

The sample size is 54, 56 and 57, for forecast horizons 3, 2 and 1, respectively.

Table 4:

Properties of forecast errors $y_t - F_{t|t-h}$, sales data for seven regions

Region	Horizon	Mean	Minimum	Maximum	Skewness
Europe	3	-0.518	-10.00	7.800	-0.052
	2	-0.587	-10.30	6.500	-0.161
	1	-0.171	-5.800	5.850	0.043
Middle East	3	-0.192	-3.200	2.880	0.055
	2	-0.128	-2.400	2.180	0.186
	1	-0.070	-2.600	1.760	-0.250
Africa	3	0.238	-3.840	3.270	-0.573
	2	0.384	-2.800	3.000	-0.264
	1	0.191	-3.440	3.800	-0.018
North America	3	0.475	-5.800	5.020	-0.402
	2	0.214	-6.000	5.610	-0.482
	1	0.360	-8.400	4.990	-1.151
Middle South America	3	0.345	-3.810	3.890	-0.086
	2	0.341	-4.710	4.330	-0.382
	1	0.356	-3.270	3.020	-0.236
Asia Pacific	3	-0.550	-12.80	6.160	-0.999
	2	-0.347	-10.10	8.690	-0.593
	1	-0.101	-7.400	4.760	-0.584
India	3	-0.151	-1.900	1.400	-0.364
	2	-0.159	-1.911	1.000	-0.862
	1	-0.141	-2.010	0.950	-0.809

The sample size is 54, 56 and 57, for forecast horizons 3, 2 and 1, respectively.

Table 5:

Modeling (the natural log of) replicable expertise, total revenues data (Newey-West HAC standard errors in parentheses)

	Forecast horizon (in months)		
Variable	3	2	1
Intercept	-1.077 (4.567)	-0.267 (4.431)	-6.334 (2.841)
$\cos\frac{2\pi t}{12}$	-0.049 (0.008)	-0.053 (0.009)	-0.056 (0.008)
$\sin\frac{2\pi t}{12}$	0.128 (0.009)	0.127 (0.011)	0.128 (0.012)
Dollar/Euro$_{t\text{-}h\text{-}1}$	0.096 (0.131)	0.056 (0.160)	-0.096 (0.149)
IP_USA$_{t\text{-}h\text{-}1}$	1.466 (0.984)	1.273 (0.974)	2.604 (0.635)
Oil price$_{t\text{-}h\text{-}1}$	0.041 (0.052)	0.074 (0.064)	0.024 (0.064)
Unemployment$_{t\text{-}h\text{-}1}$	-0.049 (0.045)	-0.048 (0.041)	-0.001 (0.026)
R^2	0.897	0.873	0.861
p-value (Wald test for all variables, except intercept and cycle)	0.000	0.000	0.000

Table 6:

Testing for bias in replicable expert forecasts (total revenues)

(Newey-West HAC standard errors in parentheses)

Horizon	$\hat{\alpha}$	$\hat{\beta}$	Wald test (p-value)
3 months	0.264 (0.395)	0.955 (0.068)	0.446 (0.800)
2 months	0.206 (0.413)	0.965 (0.071)	0.252 (0.882)
1 month	0.147 (0.361)	0.975 (0.061)	0.167 (0.920)

Test regression is:

$$\log y_t = \alpha + \beta \log \hat{F}^*_{t|t-h} + u_t$$

and the Wald test concerns the null hypothesis that $\alpha = 0, \beta = 1$. The model for the log of $\log F^*_{t|t-h}$ is presented in Table 5. In this regression, the fitted value for logF_{t-h} is included as explanatory variable. The sample size is 50, 53 and 55, for forecast horizons 3, 2 and 1, respectively.

Table 7:

Would an econometric model do better than the experts? (total revenues)

Horizon	RMSPE	
	Model	Expert
3 months	13.39	9.513
2 months	11.57	8.659
1 month	13.88	6.661

The econometric model for horizon *h* is given in Table 5.

References

Boulaksil, Y. and P.H. Franses (2009), Experts' stated behavior, *Interfaces*, 39, 168-171.

Fildes, R., P. Goodwin, M. Lawrence and K. Nikolopoulos (2009), Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning, *International Journal of Forecasting*, 25, 3-23.

Franses, P.H., M. McAleer and R. Legerstee (2009), Expert opinion versus expertise in forecasting, *Statistica Neerlandica*, 63, 334-346.

Lawrence, M. M. O'Connor and B. Edmundson (2000), A field study of sales forecasting accuracy and processes, *European Journal of Operational Research*, 122, 151-160

Mathews, B.P. and A. Diamantopoulos (1989), Judgmental revision of sales forecasts – longitudinal extension, *Journal of Forecasting*, 8, 129-140.