This article was downloaded by: [129.93.16.3] On: 10 April 2015, At: 07:56
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Interfaces

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Averaging Model Forecasts and Expert Forecasts: Why Does It Work?

Philip Hans Franses,

To cite this article:
Philip Hans Franses, (2011) Averaging Model Forecasts and Expert Forecasts: Why Does It Work?. Interfaces 41(2):177-181.
http://dx.doi.org/10.1287/inte.1100.0554

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2011, INFORMS

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

Interfaces
Vol. 41, No. 2, March–April 2011, pp. 177–181
ISSN 0092-2102 | EISSN 1526-551X | 11 | 4102 | 0177

DOI 10.1287/inte.1100.0554
© 2011 INFORMS

Downloaded from informs.org by [129.93.16.3] on 10 April 2015, at 07:56 . For personal use only, all rights reserved.

Averaging Model Forecasts and Expert Forecasts: Why Does It Work?

Philip Hans Franses
Econometric Institute, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands, franses@few.eur.nl

This paper addresses a situation in which a manager has forecasts of country-specific stock-keeping unit (SKU)-level sales data from both a statistical model and from a range of experts. Empirical evidence suggests that averaging model forecasts and expert forecasts could give more accuracy than the individual forecasts do. At the same time, empirical evidence demonstrates that expert forecasts tend to be biased. So, why is it that this average could work? This paper assumes that expert forecasts can be decomposed into a judgmental bootstrapping equation, to be created by the manager, and a part based on unobserved intuition. When the bootstrapping equation dominates the intuition, it is this underlying model for the expert that can make the combined average forecast more accurate. An illustration for no less than 1,221 cases supports this conjecture.

Key words: model forecasts; expert forecasts; combining forecasts.
History: This paper was refereed.

This paper addresses the situation in which a manager has forecasts from both a model and a range of experts. The experts are responsible for different country-specific stock-keeping unit (SKU)-level sales. The manager can decide to rely on the model forecasts, the expert forecasts, or any combination of these two forecasts.

Ample empirical evidence suggests that combining forecasts results in more forecasting accuracy; Clemen (1989), Armstrong (2001), and Timmermann (2006) provide surveys supporting this. Typically, taking the simple average (of model-based and expert-based surveys) is sufficient. Analytical results, which conclude that combining model-based forecasts are better, are usually based on the assumption that such forecasts are unbiased or consistent (Bates and Granger 1969, Timmermann 2006).

Interestingly, analytical results for combining model forecasts with expert forecasts do not exist, although substantial empirical evidence is available to show the quality of combined model and expert forecasts; Blattberg and Hoch (1990), Jorgensen (2007), Sanders and Ritzman (2004), Goodwin (2002), Whitecotton et al. (1998), and Webby and O'Connor (1996) include studies. One reason for the lack of availability of such analytical results is that little information exists on how experts create forecasts. Although suggestions for improving expert forecasts are available (e.g., Armstrong and Collopy 1998), precise formulas of the expert forecast schemes usually are not (Boulaksil and Franses 2009). Additionally, expert forecasts are typically biased (Fildes et al. 2009); thus, they violate a key assumption of the analytics.

The key question we try to answer is, what makes combining model forecasts with expert forecasts work? This paper proposes one approach that relies on the so-called judgmental bootstrapping equation to address this issue. The concept behind our approach is based on the Brunswik lens model; see Brunswik (1955) and Tucker (1964), among others. Feedback given to an expert to create better forecasts can be based on a bootstrapping equation, which summarizes the relevant information, perhaps in a simple regression format (O'Connor et al. 2005). One can also use this notion of a bootstrapping equation to single out the model that the expert could have used and separate it from an unobservable part, which we can call intuition. This model, created by the analyst (read: manager) and not by the expert, if the variables are properly chosen, could then deliver unbiased forecasts; this makes the analytics for the combination of forecasts work. If the intuition part is small relative to

RIGHTSLINK

Downloaded from informs.org by [129.93.16.3] on 10 April 2015, at 07:56 . For personal use only, all rights reserved.

the approximate model assumed by the analyst, then the bootstrapping equation contained in the expert forecast could be the element that makes the combination work. Hence, although experts do not document their forecast equation(s), if a statistical model can approximate them, and if the model is a good fit, then the combined forecast might be more accurate.

The next section, *An Illustrative Case Study*, starts with an illustration that shows that the 50%–50% rule (i.e., averaging the model forecasts and expert forecasts) works. Analyzing a huge database with no fewer than 1,221 cases with samples of model forecasts and expert forecasts shows that the 50%–50% rule gives the most accurate forecasts. In the *What Makes the Simple Average Work*? section, we seek to decompose an expert forecast into a replicable (or, model) part and a residual (or, intuition) part such that it becomes possible to examine what generates this success. That section continues with the illustration and demonstrates the relevance of the decomposition. The *Conclusion* section concludes that it is the model part of the expert forecast that makes the 50%–50% rule work.

An Illustrative Case Study

A Netherlands-based pharmaceutical company has sales offices in more than 35 countries. Each office has at least one expert who can create forecasts; model forecasts are also made at the headquarters office, which uses a variant of the ForecastPro™ software. The forecast system provides one-step-ahead (i.e., for the next month) model forecasts each month; the local managers can then give forecasts that deviate from the system's. One unknown factor is whether the local experts in the sales offices take the model-based forecast as their starting point and modify this forecast, or whether they simply create their own model. This shall turn out to be crucial to further analysis.

Each month, the system evaluates a range of models and selects the model with the best in-sample forecasting performance. Parameter estimates are updated each month.

The company sells products in seven distinct categories; however, it does not sell all products in all categories in all countries. The responsibilities of local managers differ greatly; one might be responsible for 5 products, whereas another is responsible for 40 products. The company sells 1,221 products (across all countries and categories). For each product, I have data for the sample from October 2004 to October 2006. These data include the actual sales Y, the model forecast M, and the expert forecast E. Franses and Legerstee (2009) provide details of this unique database.

For each of the 1,221 cases, I compute for M, for E, and for $(\frac{1}{2}M + \frac{1}{2}E)$, the root mean squared prediction error (RMSPE) and the mean absolute deviation (MAD). Because the RMSPE can sometimes be unreliable (see Armstrong and Collopy 1992), I also include the MAD. Additionally, I compute the median values of these criteria over 1,221 cases. The median is preferred over the mean because the data are skewed, with extreme negative values (i.e., sometimes the forecasts are very poor). To make the RMSPE and MAD comparable across the 1,221 cases, in which sales levels can differ greatly, I scale these measures by the standard deviation and by the MAD of the same sales data, respectively, prior to computing the median values.

Table 1 summarizes the relevant median values. For the model forecasts, for example, one can see that the median of the scaled RMSPEs is 1.075, which means that the forecast error variance is larger than the variance of the data. This demonstrates the difficulty in forecasting SKU-level sales data (Fildes et al. 2009). A second conclusion from the data in Table 1 is that

	Median values	
	RMSPE	MAD
Model forecast M	1.075	1.002
Expert forecast E	1.060	0.971
Linear combination $\frac{1}{2}M + \frac{1}{2}E$	1.012[a, b]	0.920[a, b]
Linear combination $\hat{\mu} + \hat{\beta}_M M + \hat{\beta}_E E$	1.573[c]	1.412[c]

Table 1: The data illustrate forecast performance of the statistical and expert models, and models of their averages. In each of the 1,221 cases, the RMSPE values are scaled by the standard deviation of the relevant sales series, and the MAD values are scaled by the MAD of the sales series.

[a] **Smaller than the relevant error measure of the model forecast M, at the 1% level (using various tests for equal medians).**

[b] **Smaller than the relevant error measure of the expert forecast E, at the 1% level (using various tests for equal medians).**

[c] **Larger than the relevant error measure of the forecasts E and M, at the 1% level (using various tests for equal medians).**

Downloaded from informs.org by [129.93.16.3] on 10 April 2015, at 07:56 . For personal use only, all rights reserved.

the linear combination $(\frac{1}{2}M+\frac{1}{2}E)$ significantly outperforms its component forecasts. This finding, based on a very large data set, confirms common wisdom and the findings in some of the literature cited above.

To examine the extent to which the 50%–50% rule is a biased forecast, I ran the regression

$$Y=\mu+\beta_M M+\beta_E E+v \qquad (1)$$

for each of the 1,221 samples, where β_M and β_E are unknown parameters and ν is an error term. Again, the mean value of the parameter estimates (over the 1,221 cases) is heavily influenced by a few outliers; therefore, the median value is more useful. The median value of the 1,221 estimates of $\hat{\beta}_M$ is 0.295 and the median of $\hat{\beta}_E$ is 0.243, whereas the median of $\hat{\beta}_M-\hat{\beta}_E$ is 0.040. In addition, $\hat{\mu}$ in Equation (1) is usually not equal to zero; these factors together imply that $(\frac{1}{2}M+\frac{1}{2}E)$ generally must be a biased forecast. Conversely, the median values of the estimates of $\hat{\beta}_M$ and $\hat{\beta}_E$ are such that M and E receive equal weights—another indication that equal weights of the model and the expert somehow constitute an appropriate combined forecast.

The last row of Table 1 shows that the linear combinations obtained from Equation (1) give poor forecast performance (e.g., compare 1.573 with 1.012). Hence, although the 50%–50% rule may lead to biased forecasts, it does give more accurate forecasts than least-squares based combined forecasts.

What Makes the Simple Average Work?

To answer the question of why $(\frac{1}{2}M+\frac{1}{2}E)$ gives better forecasts than its components, one must understand what experts do to arrive at their forecasts. For example, one must know whether they look at the model forecasts prior to making up their minds. Boulaksil and Franses (2009) survey a range of experts and report that only half look at the model forecast; the other half ignore it completely. This difference in behavior somehow must be incorporated in the analysis of the success of the 50%–50% rule.

Franses et al. (2009) propose one way to think about how an expert makes a forecast—by assuming that E is the sum of a reproducible part E^* and a nonreproducible part e. Therefore,

$$E=E^*+e, \qquad (2)$$

where an analyst can assume that E^* can be a function of M and of other factors W, such that

$$E^*=\mu+\lambda M+\delta' W, \qquad (3)$$

where W can contain prior data on Y, previous model forecast errors $(Y-M)$, previous expert forecast errors, and previous differences between E and M. The reproducible part can then be approximated by regressing E on an intercept, M and W. The expert forecast can thus be based partly on a model that differs from the model used to create M. In addition, each expert can be treated differently, because this follows from different parameter estimates in Equation (3).

Franses (2009) proposes to call e in Equation (2) intuition, because by definition it is not predictable, and because it matches the definition of intuition—understanding without apparent effort. The literature on human decision making would call Equation (3) the judgmental bootstrap equation (Dawes 1971).

In summary, one can decompose the expert forecast E as

$$E=\hat{E}+\hat{e}, \qquad (4)$$

where $\hat{E}$ is obtained from the linear regression (although other models are also possible) and where $\hat{e}$ is estimated intuition.

$$E=\mu+\lambda M+\delta' W+e, \qquad (5)$$

and the estimated residuals of Equation (5) reflect the estimated intuition. By using Equation (5), one can also adapt for the possibility that experts do not look at the model, because in that case the λ parameter will be estimated to be equal to zero.

Table 2 summarizes the results of comparing $(\frac{1}{2}M+\frac{1}{2}E)$ with $(\frac{1}{2}M+\frac{1}{2}\hat{E})$, where it is noted that the W in Equation (5) for each case covers two-month lagged sales, one-month lagged model-forecast errors, one-month lagged expert-forecast errors, and one-month lagged differences between model and expert forecasts. The key result is that $(\frac{1}{2}M+\frac{1}{2}\hat{E})$ is on average equally as good as $(\frac{1}{2}M+\frac{1}{2}E)$. Hence, it is the model that the expert uses and that the analyst approximates reasonably, thus making the linear combination $(\frac{1}{2}M+\frac{1}{2}E)$ work as an adequate forecast.

The empirical finding in this paper is that most forecasting content in an expert forecast draws upon

Downloaded from informs.org by [129.93.16.3] on 10 April 2015, at 07:56 . For personal use only, all rights reserved.

	Median values	
Linear combinations	RMSPE	MAD
$\frac{1}{2}M + \frac{1}{2}E$	1.012	0.920
$\frac{1}{2}M + \frac{1}{2}\hat{E}$	0.998	1.007[a]
All cases:		
Cases with $0.50 < R^2 < 0.75$:	0.986	0.995[a]
Cases with $R^2 > 0.75$:	0.725[b]	0.734[b]

Table 2: The forecast performance of alternative averages of forecasts also depends on the fit (R^2) of Equation (5). In each of the 1,221 cases, the RMSPE values are scaled by the standard deviation of the sales series, and the MAD values are scaled by the MAD of the sales series.

[a] Larger than the relevant error measure of the combined forecast ($\frac{1}{2}M + \frac{1}{2}E$), at the 1% level (using various tests for equal medians).

[b] Smaller than the relevant error measure of the combined forecast ($\frac{1}{2}M + \frac{1}{2}E$), at the 1% level (using various tests for equal medians).

the judgmental bootstrap equation and not upon the intuition part. Much relevant variation in the expert forecast can be covered by the bootstrap model; resultant intuition merely seems to divert the forecast and leads to inconsistency.

If this argument is valid, then one would expect to see differences across the cases with low and high R^2 (fit) values for Equation (5). When R^2 is larger, then $(\frac{1}{2}M + \frac{1}{2}\hat{E})$ should become even more accurate than average. Table 2 gives empirical evidence to support this implication. The insignificant improvement when $0.50 < R^2 < 0.75$ could be because of experts' forecasts being largely based on intuition; it could also result if the bootstrap model is incorrectly specified.

Conclusion

This paper uses an extensive SKU-level database, involving monthly forecasts for 1,221 products, to demonstrate that the simple average of model and expert forecasts yields the highest forecast accuracy. Without knowledge of exactly how these experts do forecasting, their behavior is approximated by a judgmental bootstrapping equation designed by the analyst. This equation allows for a variety of expert behaviors.

The major conclusion is that the replicable part (i.e., replicable by the analyst)—not the intuition (i.e., nonreplicable for the analyst) part—of the expert forecast makes the 50%–50% rule work. The bootstrap model averages the inconsistencies in the expert forecast—it eliminates intuition. This finding is consistent with the findings in the literature on combining model forecasts, that is, a simple average of model forecasts is often best.

One issue for further research concerns the inherent predictability of the sales series and the success rate of the linear combination $(\frac{1}{2}M + \frac{1}{2}\hat{E})$. An expert is likely to be more consistent in easier-to-forecast situations (Harvey 1995); therefore, the judgmental bootstrap model would then be a better fit. The question to be answered is how one should define predictability. The R^2 of the bootstrapping equation might (or might not) provide some guidance. Hence, seeing how combining model forecasts with expert forecasts can be mediated by the time-series properties of the sales data is a topic for additional research.

References

Armstrong, J. S. 2001. Combining forecasts. J. S. Armstrong, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Norwell, MA, 417–439.

Armstrong, J. S., F. Collopy. 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. *Internat. J. Forecasting* **8**(1) 69–80.

Armstrong, J. S., F. Collopy. 1998. Integration of statistical methods and judgment for time series forecasting: Principles from empirical research. G. Wright, P. Goodwin, eds. *Forecasting with Judgment*. John Wiley & Sons, New York, 269–293.

Bates, J. M., C. W. J. Granger. 1969. The combination of forecasts. *Oper. Res. Quart.* **20**(4) 451–468.

Blattberg, R. C., S. J. Hoch. 1990. Database models and managerial intuition: 50% model+50% manager. *Management Sci.* **36**(8) 887–899.

Boulaksil, Y., P. H. Franses. 2009. Experts' stated behavior. *Interfaces* **39**(2) 168–171.

Brunswik, E. 1955. Representative design and probabilistic theory in functional psychology. *Psychol. Rev.* **62**(3) 193–217.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* **5**(4) 559–583.

Dawes, R. M. 1971. A case study of graduate admissions: Application of three principles of human decision making. *Amer. Psych.* **26**(2) 180–188.

Fildes, R., P. Goodwin, M. Lawrence, K. Nikopoulos. 2009. Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *Internat. J. Forecasting* **25**(1) 3–23.

Franses, P. H. 2009. Can managers' judgmental forecasts be made scientifically? *Foresight, Internat. J. Appl. Forecasting* **2009**(15) 32–36.

Franses, P. H., R. Legerstee. 2009. Properties of expert adjustments on model-based SKU-level forecasts. *Internat. J. Forecasting* **25**(1) 35–47.

Franses, P. H., M. McAleer, R. Legerstee. 2009. Expert opinion versus expertise in forecasting. *Statistica Neerlandica* **63**(3) 334–346.

Goodwin, P. 2002. Integrating management judgement and statistical methods to improve short-term forecasts. *Omega Internat. J. Management Sci.* **30**(2) 127–135.

Harvey, N. 1995. Why are judgments less consistent in less predictable task situations? *Organ. Behav. Human Decision Processes* **63**(3) 247–263.

Jorgensen, M. 2007. Forecasting of software development work effort: Evidence on expert judgement and formal models. *Internat. J. Forecasting* **23**(3) 449–462.

O'Connor, M., W. Remus, K. Lim. 2005. Improving judgmental forecasts with judgmental bootstrapping and task feedback support. *J. Behav. Decision Making* **18**(4) 247–260.

Sanders, N. R., L. P. Ritzman. 2004. Integrating judgmental and quantitative forecasts: Methodologies for pooling marketing and operations information. *Internat. J. Oper. Production Management* **24**(5) 514–529.

Timmermann, A. 2006. Forecast combinations. G. Elliott, C. W. J. Granger, A. Timmermann, eds. *Handbook of Economic Forecasting*, Vol. 1. Elsevier, Amsterdam, 135–196.

Tucker, L. R. 1964. A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychol. Rev.* **71**(6) 528–530.

Webby, R., M. O'Connor. 1996. Judgemental and statistical time series forecasting: A review of the literature. *Internat. J. Forecasting* **12**(1) 91–118.

Whitecotton, S. M., D. E. Sanders, K. B. Norris. 1998. Improving predictive accuracy with a combination of human intuition and mechanical decision aids. *Organ. Behav. Human Decision Processes* **76**(3) 325–348.

Downloaded from informs.org by [129.93.16.3] on 10 April 2015, at 07:56 . For personal use only, all rights reserved.