Setting Accuracy Targets for

Short-Term Judgemental Sales Forecasting

Derek W. Bunn
London Business School
Sussex Place, Regent's Park
London NW1 4SA, UK
Tel: +44 (0)171 262 5050
Fax: +44(0)171 724 7875
Email: dbunn@london.edu

and

James W. Taylor
Saïd Business School, University of Oxford
59 George Street
Oxford OX1 2BE, UK
Tel: +44 (0)1865 288678
Fax: +44 (0)1865 288651
Email: james.taylor@sbs.ox.ac.uk

International Journal of Forecasting, 2001, Vol. 17, pp. 159-169.

Setting Accuracy Targets for Short-Term Judgemental Sales Forecasting

Abstract

Traditionally, the quality of a forecasting model is judged by how it compares, in terms of accuracy, to alternative models. However, by providing a relative measure, no indication is given as to how much scope there might be for improvements beyond the benchmark model. When judgemental methods are used alongside simple forecasting models, the scope for such improvements is considerable and difficult to benchmark. Derivation of targets for forecasting quality is thus not straightforward. The approach taken in this paper is to consider forecast error as consisting of irreducible error due to intrinsic unpredictable uncertainty, and error due to less than perfect modelling, estimation and forecasting. As the intrinsic uncertainty presents a bound on forecast accuracy, our derivation of an accuracy target is based on the measurement of this irreducible uncertainty. The motivation and data for this case-study was taken from the short-term sales forecasting process of a major, international high-technology manufacturer.

Keywords: Accuracy targets; Judgemental forecasting

1. Introduction

This research was motivated by an organisation that wished to implement a quality initiative throughout its production and inventory management. Like many companies, this organisation sought to encourage a quality culture in their operations by the setting of targets for a number of key measurable activities (Juran and Gryna, 1993). However, one important activity that presents special problems for such quality target-setting is short-term sales forecasting. This paper addresses this problem, using the company as a case study.

Benchmarking against industry leaders, and top performing companies in similar functional areas in other industries, is worthwhile for target-setting in many instances of total quality management (Hradesky, 1995). However, cross-company comparisons have not generally been relevant, nor feasible, in the area of setting forecasting quality goals. Company specific and company sensitive market issues often preclude this.

Furthermore, when we look at the research literature on forecasting, it is evident that the focus is more upon *models* than *processes*, and that the quality of a forecasting model tends to be judged by how it compares, in terms of accuracy, to a reasonable alternative statistical model. However, the value of such a comparison clearly depends on the quality of the benchmark model. Moreover, as research has shown that the fit of a model to historical data is not always a good guide to the post-sample accuracy of the model (Makridakis, 1986; Pant and Starbuck, 1990), forecasters have been advised to judge accuracy based on post-sample prediction error. Thus, we have seen many published studies deriving the post-sample forecast errors from a variety of statistical models (such as the M-Competition, Makridakis *et al.*, 1982). However, to the extent that the process of most business forecasting in practice involves considerable well-informed judgemental adjustments to simple time series methods, or may indeed be mostly judgemental, this research is therefore quite limited for the task of quality target-setting.

Indeed, it is clear that in circumstances where judgemental inputs are of proven value in forecasting, the usefulness of statistical model benchmarking is, at best, to provide **lower** quality bounds on performance. This, therefore, still leaves open the issue of assessing **upper** bounds which are theoretically feasible, but strongly challenging and can thereby provide a viable motivation for managerial forecasters.

To address this, we present a methodological framework which considers the forecast error associated with a prediction to consist of two components: the irreducible error due to the intrinsic unpredictable uncertainty in the variable, and the error due to less than perfect modelling and estimation. The intrinsic uncertainty clearly presents a bound on the accuracy of the forecasting process. Hence, our derivation of an upper bound is based on the estimation of this irreducible component of uncertainty in the data. The analogy is with the study of physical systems, where observation noise can be seen as an upper limit to the accuracy of systematic measurements. This concept has been extended in forecasting research. For example, Bunn and Seigal (1983) found that there was an upper bound on the accuracy of minute-by-minute electricity load forecasts due to load measurement problems and used this as a basis for assessing the performance of various short-term predictors. Compared to a measure of *ex post* accuracy, which evaluates the forecast against the actual out-turn, the proposed quality target is clearly more reasonable, but is still an idealised upper bound on performance.

Measures of actual *ex post* accuracy are, of course, essential for monitoring and, as we have observed, simple model based comparisons can provide a reasonable lower bound on performance. It would seem reasonable, therefore, to evaluate the usefulness of quality target bounds where the upper ones are based upon estimates of irreducible uncertainty, and the lower ones are derived from a simple time series model (e.g. random walk).

We applied this approach to the quality initiatives of our collaborating company. This company operates worldwide in the fast changing, high-technology sector, selling a range of personal computers directly to consumers, either by telephone or internet. They hold no inventories of finished goods, just component parts, and assemble to order. The products have quite short life cycles, with sales very dependent upon pricing and advertising. Their forecasts are mostly judgemental estimates, using sales force knowledge plus market information on product innovations and promotions, against a background of daily monitoring of underlying sales trends per product line. In this respect the case study described here is typical of a more general class of consumer forecasting problems where there is high frequency data (eg EPOS, internet or phone), and a necessarily substantial judgemental component.

The hypothesis of this study is that, in the spirit of TQM, a forecast quality target, implemented with regular monthly feedback, will motivate and monitor improved forecasting throughout the company and that such a target could consist of two bounds. The upper bound could be an estimator which forecasts with error due only to intrinsic uncertainty, whilst the lower bound could be a naïve statistical method based upon a random walk. It is important to understand that this paper is not concerned with the problem of estimating prediction intervals. A prediction interval conveys the interval within which an actual out-turn is likely to fall with a given probability, such as 95%. A quality target is a measure that one must try to attain. Since forecast quality is assessed by accuracy measures, such as MSE, this paper aims to provide a methodology for deriving a value for the measure which would serve as a quality target.

In section 2, we consider the literature on forecast accuracy measures, in order to provide an appropriate metric for these bounds. We present the company's own metric and then, in section 3, we present a framework for deriving bounds on forecast accuracy. Section 4 discusses the limitations of an analytical approach to assessing the accuracy bounds, whilst section 5 presents an alternative approach, which uses Monte Carlo simulation. Section 6 reports the application of the simulation procedure to the company's data and the final section offers some concluding comments.

2. Evaluating Forecast Accuracy

2.1. Forecast Accuracy Measures

In a survey of practitioners and academicians, Carbone and Armstrong (1982) found that the most preferred measure of forecast accuracy is the Mean Square Error (MSE). Chatfield (1992) writes that, for a single series, it is perfectly reasonable to fit a model by least squares and evaluate forecasts from different models by the MSE. However, the MSE has been broadly criticised for use in comparing forecasting methods across series, as it can be disastrous to average MSE from different series since it is scale-dependent. Because of this, and its poor protection to outliers, Armstrong and Collopy (1992) recommend against using the MSE. They found that for selection among forecasting methods, the MdRAE, GMRAE and MdAPE are to be preferred to the MSE. The MdRAE is the median value of the relative absolute error (RAE). This is calculated for a given time series by dividing the absolute forecast error, at a given horizon, for a proposed model by the corresponding error for the random walk. The GMRAE is the geometric average of the RAE values. The MdAPE is the median absolute percentage error which has the advantage of having a closer relationship to decision making. In this context, Fildes (1992) argues that forecast comparisons, based on a population of time series, should not rely on a single time origin for each series but, instead, the error measure should be averaged across time in some way, as well as across different series. Of the alternatives considered, he found that only the geometric root mean squared error (GRMSE) is well-behaved and has a straightforward interpretation. Makridakis and Hibon (1995) disagree, arguing that the GRMSE has very little intuitive meaning as it involves squared terms, products and geometric roots. Their study evaluates various accuracy measures using two statistical and two user-oriented criteria. They conclude that selecting amongst them depends upon the situation involved and the needs of decision or policy makers. They note that such a choice cannot be made without trade-offs. Clearly this issue is still a controversial area in forecasting research.

2.2. The Company's Accuracy Measure

Returning to the context of our case study and the company involved, after much deliberation concerning ease of interpretation and monitoring, the forecast accuracy measure adopted by the company was

$$(f_t / x_t)\times 100 \quad \text{if} \quad f_t < x_t$$

$$\text{and} \qquad (1)$$

$$(x_t / f_t) \times 100 \quad \text{if} \quad f_t > x_t$$

where x_t is the actual value and f_t is the forecast. We have termed this the similarity percentage (SP). If a summary measure is required for forecast accuracy for several periods, the mean or median of the SP values can be used. The median would be more consistent with the recent forecast evaluation literature as it would be more robust to outliers.

The SP seemed to fit the company's requirements for a quality measure since it varies between zero and 100%, with a higher percentage reflecting more accurate forecasting. However, it is not one of the metrics usually discussed in the literature and this is probably because of its ambiguous interpretation. For example, an SP of 80% can result from two possibilities: either f_t is greater than x_t and the forecast error is 20% of f_t, or f_t is less than x_t and the forecast error is 20% of x_t. Nevertheless, this seemed to make sense to the company in terms of the relative costs of over-stocking versus lost sales.

An advantage of the SP is that, since it is in percentage terms and not scale dependent, it can be averaged across both time periods and time series. The measure seems to have one advantage over the absolute percentage error (APE), which is the most widely used percentage measure. Makridakis (1993) notes that the APE has the problem that equal errors when x_t is larger than f_t give smaller percentage errors than when x_t is smaller than f_t. (For example, when x_t is 150 and f_t is 100, the APE is 33.33%, however, when x_t is 100 and f_t is 150, the APE is 50%.) This difference can create serious problems when the value of x_t is small (close to zero) and f_t is large, as the size of the APE can become extremely large making the comparisons among horizons and among series difficult. This type of asymmetry is not a problem, however, for the similarity percentage (SP) and, furthermore, the SP can never be greater than 100% (for positive values of x_t and f_t). Interestingly, the modified MAPE, proposed by Makridakis (1993) to overcome this type of asymmetry, has recently been shown by Goodwin and Lawton (1999) to be far from symmetric in other respects. This may also be the case for the SP.

The main contribution of our work is not affected by the choice of accuracy measure. Thus, since this paper is essentially a case study, and the company had already established the practice of working with the SP measure, the accuracy targets described below, which we developed and were implemented by the company, were most appropriately expressed in these terms.

3. Framework for Deriving Accuracy Targets

3.1. The Components of Forecast Error

Having chosen an accuracy metric as a measure of forecast quality, we now address the problem of deriving target values for the metric. The way that we approach the problem is to estimate the limits on the accuracy of the company's sales forecasts. Our methodological framework considers the forecast error associated with a prediction as consisting of two components: the irreducible error, e_t, due to the intrinsic unpredictable uncertainty in the variable, and the error, ε_t, due to less than perfect estimation (i.e. forecasting). We can then write the forecast error as

$$x_t - f_t = e_t + \varepsilon_t$$

Improvement in forecasting will reduce ε_t but it cannot reduce the inherent uncertainty, e_t. Consequently, an ideal predictor will have $\varepsilon_t = 0$ and it is clear that e_t implies a bound on forecast accuracy. Suppose we are able to estimate the variance, σ_t^2, of e_t, and we assume that e_t is normally distributed about zero. We can then say that, with 5% probability, e_t will fall outside the interval $[-1.96\sigma_t,1.96\sigma_t]$. Since $\varepsilon_t = 0$ for a perfect predictor, we can say that the forecast error for a perfect predictor would be expected to fall within these bounds 95% of the time. This could then serve as a target for forecast accuracy. It is important to note that these 95% limits are confidence limits for a perfect predictor and are, therefore, not standard prediction interval limits. In section 4, we adapt this idea for estimating targets to obtain bounds for the accuracy metric employed. First, however, we address the problem of assessing the variance, σ_t^2, of the intrinsic uncertainty, e_t, as this is fundamental to the approach.

3.2. Estimating the Intrinsic Uncertainty

The problem, as posed by the company, is to establish statistical target bounds for calendar *month* sales forecast accuracy measures based upon some measure of the intrinsic randomness in the data. Thus, we wish to estimate the irreducible component of uncertainty in the monthly values. The conceptual framework that we present for estimation is new and relies upon the availability of data with frequency that is higher than monthly data, for example, daily or weekly observations.

The general proposition is that measurements taken at higher frequency than that postulated for the underlying structural process can provide a means of estimating intrinsic randomness. The analogy is with estimating noise in physical systems, from repeated measurements with the same control variables. In our case study, the structural process proposed for the data by the company is that of monthly shifts in demand which need to be forecasted, but weekly or daily variations within the month which exhibit intrinsic randomness about that monthly mean. Clearly, more complex structural models could be developed imputing within month trends, but the data series investigated here did not seem to warrant the extra parameterisation, at least for a first analysis. Product life-cycles were short and the monthly mean shift seemed to be an appropriate structural assumption.

Thus, if average sales per week is correctly forecast, there would still be statistical error in the monthly total due to random week-by-week variations about this average. This variation could be used to estimate the intrinsic uncertainty in the monthly totals, and this estimate could then be used to provide statistical bounds for monthly accuracy. The key underlying "stability" assumption is that each week's value is varying independently about the same weekly average within a calendar month, although we would expect small average changes from one calendar month to another.

The variance, σ_t^2, of the intrinsic uncertainty, e_t, in the calendar month total, x_t, is then calculated from the four weekly values, x_{1t}, x_{2t}, x_{3t} and x_{4t}, thus

$$\sigma_t^2 \;=\; var(\, x_{1t} + x_{2t} + x_{3t} + x_{4t}\,) \;=\; 4\sigma_{wt}^{\;2} \tag{2}$$

where $\sigma_{wt}^{\;2}$ is the variance in a weekly sales value. This variance, $\sigma_{wt}^{\;2}$, is estimated as the variance of the four values x_{1t}, x_{2t}, x_{3t} and x_{4t}. Expression (2) relies on the assumption that the four weekly sales values are not correlated. This was indeed true for our company's data which implies that time series forecasting methods will be ineffective. It is therefore not surprising that the company uses judgemental methods to produce sales forecasts for these short life-cycle products. Although this framework is most suitable for judgemental forecasting, it is worth noting that it can be adapted to the situation where stable or sustained patterns do exist in the data. In such cases, the underlying pattern should be estimated and then the weekly randomness from it used to estimate intrinsic uncertainty in the monthly totals.

Whilst only weekly data was available for the individual products, daily data was available for the total sales. We calculated the intrinsic variances for the total monthly sales using weekly data and compared the results to those using daily data. For monthly totals, the intrinsic variances computed from weekly and daily data seemed sufficiently close to justify calculation of the monthly intrinsic variances for the individual products using the weekly data. We also investigated the significance of inter-correlations between sales of products on a weekly basis, to see the extent to which statistical dependencies needed to be taken into account when looking at averages of the accuracy measure across products. Our analysis indicated that the inter-correlations could be omitted from the analysis as they did not appear to have a major effect.

The idea of using higher frequency data to estimate the variance in lower frequency data is not entirely original. For example, it has been used in financial applications to evaluate volatility forecasts (e.g. Day and Lewis, 1992). As volatility is unobservable, it is not obvious what to use as the *actual* with which to compare the forecast. Suppose we wish to evaluate forecasts of weekly volatility. If daily observations of the returns are available, a proxy for the actual weekly-return variance is created by computing the daily variance from the daily returns, and then multiplying this by the number of trading days in the week.

4. Analytical Approach to Assessing Accuracy Bounds for SP

We can work towards a bound on accuracy by considering the limit on the forecasting performance of an ideal predictor. The forecasts, p_t, of an ideal predictor have $\varepsilon_t = 0$, so that

$$x_t - p_t = e_t$$

If we assume that e_t is normally distributed, we can say that with 5% probability, e_t will fall outside the interval $[-1.96\sigma_t, 1.96\sigma_t]$. Using this, and recalling the definition of the similarity percentage (SP) given in expression (1), we can make the following probability statements for the forecast of the ideal predictor:

Since $p_t > x_t + 1.96\sigma_t$ with probability of 2.5%, then with probability of 2.5%:

$$x_t < p_t \qquad \text{and} \qquad \frac{x_t}{p_t} < \frac{x_t}{x_t + 1.96\sigma_t}$$

Since $p_t < x_t - 1.96\sigma_t$ with probability of 2.5%, then with probability of 2.5%:

$$x_t > p_t \qquad \text{and} \qquad \frac{p_t}{x_t} < \frac{x_t - 1.96\sigma_t}{x_t}$$

We can thus say that, with probability less than 5% and greater than 2.5%:

$$\text{SP of the Perfect Predictor} < \min\left(\frac{x_t}{x_t + 1.96\sigma_t}, \frac{x_t - 1.96\sigma_t}{x_t}\right)\times 100 = \left(\frac{x_t - 1.96\sigma_t}{x_t}\right)\times 100 \tag{3}$$

The value of SP for the ideal predictor will thus be greater than the expression in (3) between 95 and 97.5% of the time. If the value of SP for our forecast is found to be greater than the expression in (3), then we could conclude that it is not significantly less accurate than the ideal predictor. The expression in (3) could then be used as a forecast accuracy target.

However, this analytical approach to assessing bounds for the accuracy measure is unsatisfactory for two reasons. Firstly, we may require an upper bound corresponding to a particular percentage. However, the analysis is only able to supply the imprecise probability statement for the SP of the perfect predictor with probability “less than 5% and greater than 2.5%”. The second reason concerns the company’s additional requirement that accuracy targets are also derived for the weighted average (*WSP_t*) of the similarity percentage for month *t*, averaged across different products.

$$WSP_t = \sum_i w_{it}\, SP_{it} \qquad \text{where} \qquad w_{it} = \frac{f_{it}}{\sum_i f_{it}}$$

f_{it} is the company’s forecast for sales of product *i* in month *t*, and SP_{it} is the similarity percentage for f_{it}. We require a 5% bound for WSP_t for a particular month. Unfortunately, it is impossible to form probability statements, as in the previous section, for a weighted average of the accuracy metric. Consequently, it is impossible to form a probability statement for WSP_{it}. With the complexity of the problem limiting the usefulness and feasibility of theoretical analysis, recourse to simulation provides a practical alternative.

5. Simulation Approach to Assessing Accuracy Targets for SP and WSP

Upper bounds may be derived for the accuracy measure by simulating the actual sales, x_{it}, for product *i* in month *t*. We proceed by considering the observed actual as being a random variable consisting of a non-stochastic expectation component, E(x_{it}), plus an intrinsic error term, e_{it}. Having estimated the standard deviation, σ_{it}, of e_{it}, we are then in a position to simulate values of x_{it} as

$$x_{it} = \mathrm{E}(x_{it}) + e_{it}$$

where e_{it} is a value derived by Monte Carlo sampling from a normal distribution with zero mean and standard deviation σ_{it}. We use the observed actual as $\mathrm{E}(x_{it})$. The ideal predictor, p_{it}, is then modelled as a perfect predictor of $\mathrm{E}(x_{it})$ so that the only error is due to the intrinsic uncertainty.

$$p_{it} = \mathrm{E}(x_{it})$$

For each simulated value of x_{it}, we record the value of the similarity percentage, SP_{it}:

$$(\, p_{it} / x_{it}\,)\% \quad \text{if} \quad p_{it} < x_{it}$$

and

$$(\, x_{it} / p_{it}\,)\% \quad \text{if} \quad p_{it} > x_{it}$$

We also record the weighted similarity percentage for the ideal predictor using the simulated actuals for a particular month,

$$WSP_t = \sum_i w_{it}\, SP_{it} \qquad \text{where} \qquad w_{it} = \frac{p_{it}}{\sum_i p_{it}}$$

By repeatedly sampling from the distribution for e_{it}, we produce a distribution for SP_{it} and for WSP_t. The 5th percentiles of the resultant probability distributions can then be interpreted as respective upper bounds for the similarity percentage and weighted similarity percentage of the company’s sales forecasts. They have been computed at the 95% confidence level, in the sense that, with ideal forecasting of the monthly means, intrinsic variation would imply that one’s forecast metric would be less than the bounds only 5% of the time. The bound is therefore the limit of the acceptance region corresponding to the one-sided hypothesis test with null that the forecast is at least as accurate as the perfect predictor.

It is worth reiterating that the simulation methodology described in this section aims to generate the distribution for the similarity accuracy measure for a perfect predictor. We then use this probability distribution to provide quality targets for the similarity accuracy measure for the company’s forecasts. Of course, these forecasts have not been used to construct the accuracy targets which, therefore, remain fixed over time. It is hoped that these targets will motivate improvement in the company’s forecast accuracy over time.

Although the main aim of this paper is to derive upper bounds for forecasting accuracy, which can be interpreted as quality targets, it seems useful to also consider a general approach to the derivation of lower bounds for accuracy. The random walk is often taken as a benchmark against which to test the success of other forecasting methods. Rather than simply calculating the accuracy metrics (SP and WSP) for random walk forecasting for each month, we used the simulated actuals to build up a distribution for the accuracy metrics for the random walk.

The company’s forecasts are made with a lead time of three months. In view of this, we used the simulated actual for month *t*-3 as the random walk forecast for product *i* in period *t*. By recording the SP for each simulated actual and the WSP for each month, we generated a distribution for the SP and WSP for random walk forecasting for each product in each month. A sensible probability statement regarding a lower bound would be that the forecast should not be significantly worse than the random walk. In other words, the accuracy metric for the company’s forecast should not be significantly lower than the accuracy measure for the random walk. The lower bound for accuracy would then be the 5th percentile of the probability distribution of the simulated accuracy measure for the random walk.

6. Case Study Results

We had data for 11 months. We used 1,000 iterations in the Monte Carlo simulations. In other words, we produced a thousand simulated actuals from which we calculated the SP and WSP for the ideal predictor and for the random walk. The 5th percentiles of the resultant distributions were then used as upper and lower bounds, respectively, on accuracy. The upper bound can serve as a target for forecast quality.

The company groups all of its products into one of three different categories. The weighted accuracy metrics are calculated separately for each category. The results of the simulations for the WSP for category 1 are given in Table 1. Figure 1 shows the distribution generated for the WSP of perfect predictor forecasting for the products in category 1 in month 1. It is easy to confirm that the 5th percentile is approximately 91%, as reported in Table 1 for month 1. The negative skewness of the distribution was typical for all the WSP distributions.

Briefly, the results indicated that the upper bound or target value for the weighted similarity percentage (WSP) varied between about 80 and 90%. The company's forecasts in this data set generally had a WSP which was about 30% lower. A noticeable difference between the recorded accuracy and the target is to be expected as the targets are ultimate and are not realistically achievable.

Rather more surprising are the values of the company's WSP relative to the lower bounds. The results indicated that in more than half the cases the company's WSP is below the corresponding lower bound and thereby appeared to be no better than a random walk model. However, in restricting the data set to stable months, and requiring a three month lead time, the sample size was small, already being judgementally filtered. Nevertheless, since performance on this data set was closer to the bottom than the top of the target bands, it did cause the company to think about more systematic ways of using the subjective sales and marketing information.

Overall, the main result of this study was to suggest that a target of 85% accuracy in the WSPs should be used for performance monitoring in the company and that a quality initiative should sustain a steady improvement towards that goal. Indeed in the six months that followed the setting of a quality target, the company did manage to move its average monthly WSPs to above 80%. Feedback on actual accuracy had been given to forecasters for several years, but it seems that being associated with a target provided the extra motivation for quality improvement. Furthermore, it does seem that quality improvement is essentially in the forecasting of demand, rather than in the sales teams managing their out-turns. The forecasts were estimated centrally, whereas the sales were almost completely made through telephone call centres during the period of this study. Thus, given the short-lead times involved, the lack of personal contact between the sales team at the call centres and most of the customers and the fact that they had no discretion on discounting list prices, there would appear to have been relatively little scope for the sales team to have influenced the attainment of the forecasts.

----- TABLE 1 -----

----- FIGURE 1 -----

7. Concluding Comments

Based upon the assumptions of intrinsic randomness, we have presented a simulation framework for deriving an upper and lower bound for the weighted accuracy measure. In the actual case study, the approach assumed that there is unforecastable week-by-week variation within each month, but that the average, from which these weeks are statistical outcomes, is predictable.

The upper bound was computed at the 95% confidence level, in the sense that, with ideal forecasting of the monthly means, the forecast metric should achieve this bound 95% of the time. It is, of course, unreasonable to assume that, 3 months out, product transition effects, media influences, logistics and other events can be anticipated perfectly, and so we have to see the upper bound as an ideal target against which to motivate quality improvements in forecasting; it should not be seen as a benchmark that should be regularly attainable. Juran (1988) discusses how it should not be assumed that quality goals can always be met. Nevertheless, they do provide a basis for motivation. In this case, they succeeded where simple feedback had failed, and quality improvements in forecasting were achieved relative to the target.

The quality target bounds are, therefore, ultimate and are not realistically achievable every month. In order to position forecasting performance within the range of attainable accuracies, we calculated lower bounds. The random walk model is often taken as a benchmark against which to test the success of other forecasting methods. It essentially assumes that future changes from the current level are unpredictable and that the best forecast is the latest value. The lower bound was specified so that if the accuracy measure falls below that value, one can conclude with 95% confidence that the forecast has been outperformed by the random walk. As we saw in this example, historical performance close to the lower bound can also instigate greater efforts at forecast improvements.

Finally, it should be observed that this is a "bounding" exercise, looking at statistical variation from two extremes. The upper assuming that the data is predictable 3 months out except for some short-term, intrinsic week-by-week variation. The lower assuming that changes from the latest value are not at all predictable. It is not, therefore, a “benchmarking” exercise in terms of seeing what sort of accuracy other commonly used techniques, such as exponential smoothing, could give. Nor does it address how a structured process of forecasting quality improvement might be implemented. Nevertheless, the statistical bounds derived here did provide a framework within which to motivate improvements in forecasting. The principle is generalisable to more complex structural models, such as trending and life-cycle models, and may well be the most appropriate way to set quality targets for forecasts which involve substantial judgemental inputs.

Acknowledgements

We would like to acknowledge the helpful comments of an associate editor and two anonymous referees.

References

Armstrong, J. S. & Collopy, F. (1992). Error Measures for Generalising About Forecasting Methods: Empirical Comparisons, *International Journal of Forecasting 8,* 69-80.

Bunn, D. W. & Seigal, J. P. (1983). Forecasting the Effects of Television Programming upon Electricity Loads, *Journal of the Operational Research Society 34,* 17-25.

Carbone, R. & Armstrong, J. S. (1982). Evaluation of Extrapolative Forecasting Methods: Results of a Survey of Academicians and Practitioners, *Journal of Forecasting 1*, 215-217.

Chatfield, C. (1992). A Commentary on Error Measures, *International Journal of Forecasting 8,* 100-102.

Day, T. E. & Lewis, C. M. (1992). Stock Market Volatility and the Informational Content of Stock Index Options, *Journal of Econometrics 52*, 267-287.

Fildes, R. (1992). The Evaluation of Extrapolative Forecasting Methods, *International Journal of Forecasting 8,* 81-98.

Goodwin, P. & Lawton, R. (1999). On the asymmetry of the symmetric MAPE, *International Journal of Forecasting 15,* 405-408.

Hradesky, J. (1995). *Total Quality Management Handbook,* McGraw-Hill, New York, 645-655.

Juran, J. M.. (1988). *Juran on Planning for Quality*, The Free Press, New York, ch. 8.

Juran, J. M. & Gryna, F. M. (1993). *Quality Planning and Analysis - From Product Development Through Use*, 3rd Ed, McGraw-Hill, USA, ch. 8.

Makridakis, S. (1986). The Art and Science of Forecasting: An Assessment and Future Directions, *International Journal of Forecasting 2*, 15-39.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting 9*, 527-529.

Makridakis, S. & Hibon, M. (1995). Evaluating Accuracy (or Error) Measures, Working Paper, INSEAD, Fontainebleau, France.

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

Pant, P. N. & Starbuck, W. H. (1990). Innocents in the Forest: Forecasting and Research Methods, *Journal of Management 16*, 433-460.

Months	1	2	3	4	5	6	7	8	9	10	11	Average
WSP Lower Bound	-	-	-	67	56	58	68	48	66	57	59	60
Company’s WSP	64	54	68	61	51	46	49	63	51	67	50	57
WSP Upper Bound	91	89	88	87	87	89	84	90	92	86	90	88

Table 1: Bounds for the WSP for product category 1

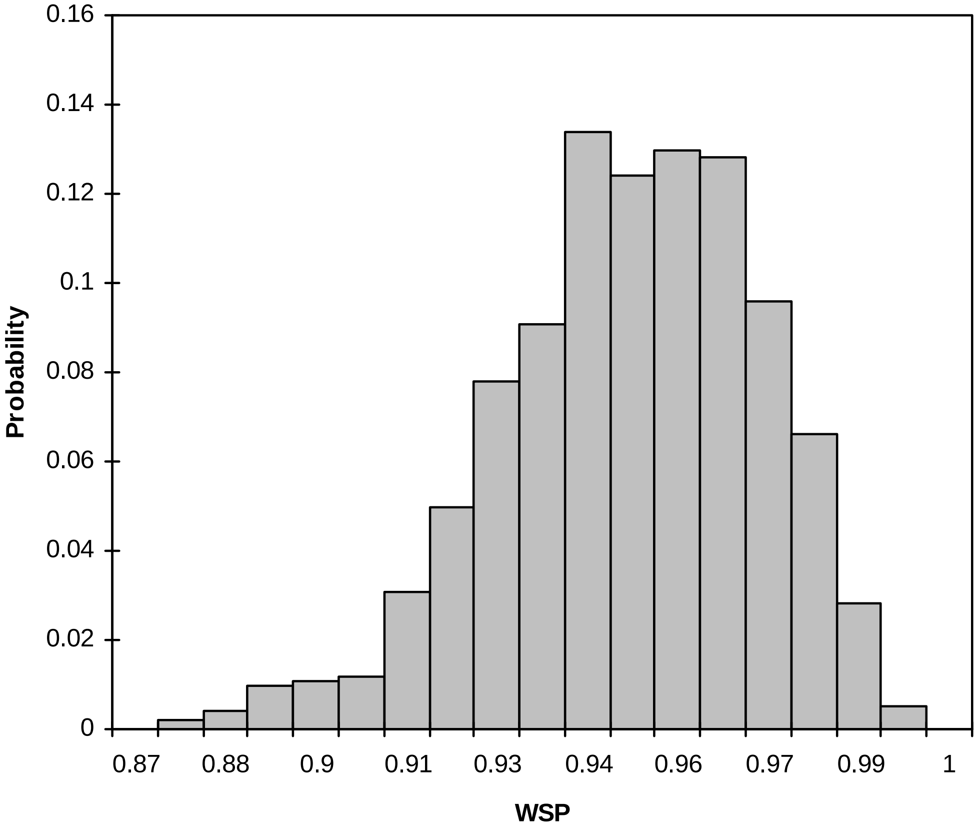

Figure 1: Distribution for the WSP of the ideal
predictor for the products in category 1 in month 1