ELSEVIER

International Journal of Forecasting 16 (2000) 261–275

international journal of forecasting

www.elsevier.com/locate/ijforecast

Correct or combine? Mechanically integrating judgmental forecasts with statistical methods☆

Paul Goodwin*

Faculty of Computer Studies and Mathematics, University of the West of England, Frenchay, Bristol BS16 1QY, UK

Abstract

A laboratory experiment and two field studies were used to compare the accuracy of three methods that allow judgmental forecasts to be integrated with statistical methods. In all three studies the judgmental forecaster had exclusive access to contextual (or non time-series) information. The three methods compared were: (i) statistical correction of judgmental biases using Theil's optimal linear correction; (ii) combination of judgmental forecasts and statistical time-series forecasts using a simple average and (iii) correction of judgmental biases followed by combination. There was little evidence in any of the studies that it was worth going to the effort of combining judgmental forecasts with a statistical time-series forecast – simply correcting judgmental biases was usually sufficient to obtain any improvements in accuracy. The improvements obtained through correction in the laboratory experiment were achieved despite its effectiveness being weakened by variations in biases between periods.

Keywords: Judgmental forecasting; Combining forecasts

1. Introduction

Several studies have found that, in many contexts, both human judges and statistical methods have valuable and complementary contributions to make to the forecasting process (e.g., Blattberg & Hoch, 1990). For example, statistical methods are adept at filtering regular time series patterns from noisy data while judgmental forecasters tend to see false patterns in noise and to overreact to random movements in series (O'Connor, Remus & Griggs, 1993). On the other hand, when it is known that special events will occur in the future, judgment can be used to anticipate their effects, while statistical estimation of these effects may be precluded by the rarity of the events.

The integration of judgmental forecasts with statistical methods can be carried out in several ways. Voluntary integration involves supplying the judgmental forecaster with a statistical forecast, which the forecaster is then free to ignore, accept or adjust. However, a recent study by Goodwin and Fildes (1999) found that judgmental forecasters carried out voluntary integration inefficiently. They made deleterious adjust-

☆An earlier version of this paper was presented at Nineteenth International Symposium on Forecasting, Washington DC, June 1999.

*Tel.: +44-117-965-6261; fax: +44-117-976-3860.

E-mail address: paul.goodwin@uwe.ac.uk (P. Goodwin)

0169-2070/00/$ – see front matter
PII: S0169-2070(00)00038-8

ments to statistical forecasts when they were reliable and ignored these forecasts in periods when they formed an ideal baseline for adjustment. A similar study by Lim and O'Connor (1995) also found that forecasters tended to underweigh statistical forecasts in favour of their own judgments, even when their attention was drawn to the superior accuracy of the statistical forecasts.

In the light of these concerns some researchers have recommended that the integration should be carried out mechanically (Lawrence, Edmundson & O'Connor, 1986; Lim & O'Connor, 1995). Combining and correction are two methods of mechanical integration that have been proposed for situations where the forecasts are expressed as point estimates[1]. In *combining* the forecast is obtained by calculating a simple or weighted average of independent judgmental and statistical forecasts (Clemen, 1989). *Correction* methods involve the use of regression to forecast errors in judgmental forecasts. Each judgmental forecast is then corrected by removing its expected error (e.g. see Theil's optimal linear correction (Theil, 1971)). Correction has received less attention in the literature than combination. However, arguably correction, in its simplest forms, is more convenient in that it does not require the identification, fitting and testing of an independent statistical method in addition to the elicitation of judgmental forecasts.

An obvious concern of using any of these integration methods arises when the judgmental forecaster has access to information about special events that is not available to the statistical method. For example, averaging a judgmental forecast, which reflects the expected high sales that will result from a promotion campaign, with a statistical forecast that takes no account of the campaign may reduce forecast accuracy. Similarly, if correction is employed, estimates of judgmental biases that will occur in forecasts for 'special' periods may be contaminated by the different types of biases observed in 'normal' periods, and vice versa. In practice, information about special events, and the fact that the judgmental forecaster used this information, may not be made explicit or recorded so that it is not possible to remove its effects from the correction model or to suspend averaging with the statistical forecast when special events apply. This is a particular danger because mechanical integration methods are likely to be most appropriate when employed by recipients of judgmental forecasts rather than the forecasters themselves (Goodwin, 1996).

This paper addresses two research questions in circumstances where the judgmental forecaster has exclusive access to non-time series information that will have an impact on the forecast variable.

1. What is the relative accuracy of forecasts obtained through (i) correction, using Theil's optimal linear correction, (ii) combination, using a simple average of judgmental and statistical time series forecasts, and (iii) using both approaches in tandem?
2. To what extent, if any, do these methods improve judgmental forecasts, even though the judgmental forecaster has exclusive access to non-time series information?

To answer these questions data was obtained from two sources. First the integration methods were applied to judgmental forecasts made by subjects in a laboratory experiment. This data allowed the research questions to be explored under a range of controlled conditions. Then, to assess the extent to which the laboratory results can be generalised, the methods were applied to judgmental sales forecasts made by managers in two manufacturing companies.

[1]Note that the discussion here relates to integration of judgment with statistical *methods*, not just statistical *forecasts*.

The paper is organised as follows. In the next Section, the theory underlying the mechanical integration methods is explored and examples of the applications of these methods that have been reported in the literature are reviewed. In the subsequent Section the laboratory experiment is outlined, and the results of the application of the mechanical methods to the experimental data are presented and discussed. Following this, the application of the methods to the industrial forecasts is outlined and the results compared with those from the laboratory study.

2. Background and theory

2.1. Correcting judgmental forecasts

Theil (1971) showed that the mean squared error (MSE) of a set of forecasts can be decomposed into three elements.

$$\text{MSE} = \underbrace{(\bar{Y}-\bar{F})^2}_{\text{Term 1}} + \underbrace{(S_F-\rho S_Y)^2}_{\text{Term 2}} + \underbrace{(1-\rho^2)S_Y^2}_{\text{Term 3}} \quad (1)$$

here $\bar{Y}$ and $\bar{F}$ are the means of the outcomes and point forecasts, respectively, S_F and S_Y are the standard deviations of the point forecasts and outcomes, respectively and ρ is the correlation between the point forecasts and outcomes.

In this decomposition, Term 1 represents mean (or level) bias. This is the systematic tendency of the forecasts to be too high or too low. Term 2 represents regression bias. This measures the extent to which the forecasts fail to track the actual observations. For example, forecasts may tend to be too high when outcomes are low and too low when outcomes are high. Theil then showed that mean and regression bias can be eliminated from a set of past forecasts (i.e. forecasts for periods where the outcomes have been realised) by using an optimal linear correction. This simply involves regressing the actual outcomes on to the point forecasts and using the resulting intercept and slope estimates, and, to make the correction as shown below:

$$Y_t = \hat{a} + \hat{b}F_t \quad (2)$$

so that

$$P_t = \hat{a} + \hat{b}F_t \quad (3)$$

where Y_t is the outcome at time t, F_t is the point forecast for time t and P_t is the corrected judgmental point forecast for period t.

Ahlburg (1984) found that the correction substantially improved forecasts of US prices and housing starts, while Shaffer (1998) found that correction of commercial forecasts of the US implicit GNP price deflator reduced the MSE of out-of-sample forecasts by either 15% or 25%, depending on the forecast lead time. Similarly, Elgers, May and Murray (1995) applied it to analysts' company earnings forecasts and reported that it reduced the MSEs emanating from systematic bias by about 91%. In a laboratory experiment, Goodwin (1997) found that the correction was most successful where series had high levels of noise. In particular, for white noise series the correction had the effect of smoothing out the variation in the judgmental forecasts which was caused by the forecasters reacting to the random movements in the series.

2.2. Combining judgmental forecasts with statistical forecasts

The effectiveness of combining independent judgmental and statistical forecasts has been examined in several studies (see Clemen (1989) for a review). The general conclusion is that combining improves forecast accuracy because the constituent forecasts are able to capture 'different aspects of the information available for prediction' (Clemen). Although it is possible to use a weighted average to achieve the combination, estimating the appropriate weights when there is only a small data base of past

observations is problematical (Bunn, 1987). This is likely to be a common problem in industrial contexts, particularly in industries where products are subject to rapid change and development (Watson, 1996). In fact, many studies have found that a simple mean of the two forecasts performs relatively well (de Menezes, Bunn & Taylor, 2000). Moreover, Armstrong and Collopy (1998) argue that the simple mean is particularly appropriate where series have high uncertainty and instability because, under these conditions, there will be considerable uncertainty as to which method is likely to be most accurate. (Hereafter, the term 'combination' will refer to the simple mean of two forecasts.)

When the constituent forecasts in a combination are free of mean bias, the MSE of the forecasts is identical to the variance of the forecast errors. In these circumstances, it is easy to show that the variance of the forecast errors of the combined forecasts, σ_c^2, is given by:

$$\sigma_c^2 = 0.25(\sigma_s^2 + \sigma_j^2 + 2r\sigma_s\sigma_j) \quad (4)$$

where σ_s^2 is the variance of the errors of the statistical forecasts, σ_j^2 is the variance of the errors of the judgmental forecasts and r is the correlation between the constituent forecasts' errors

It can be shown that this implies that the variance of the errors of the combined forecasts, and hence the MSE, will only be lower than that of the judgmental forecasts when:

$$\frac{\sigma_j}{\sigma_s} > \frac{r + (r^2 + 3)^{0.5}}{3} = \Phi \quad (5)$$

If it is also the case that:

$$\frac{\sigma_j}{\sigma_s} < \frac{1}{\Phi} \quad (6)$$

then the MSE of the combined forecast will be less than that of both the constituent forecasts. Fig. 1 shows when combination will reduce the MSE of either or both of the constituent forecasts for different values of r. For example, if the two sets of forecast errors are perfectly negatively correlated then combination will improve both forecasts if σ_j/σ_s is greater than $1/3$ and less than 3. Essentially, the vertical axis

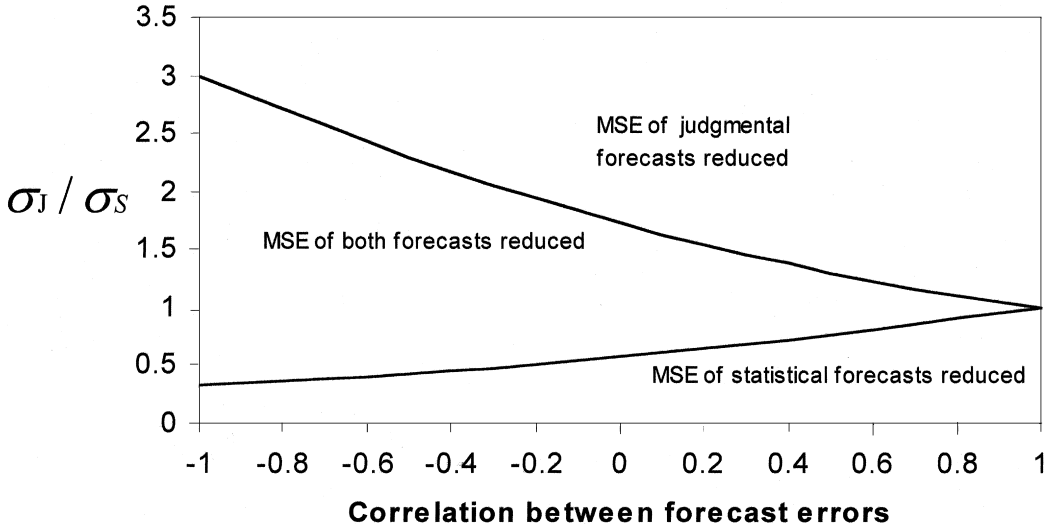

Fig. 1. Where combination improves constituent forecasts.

of the graph represents the relative inaccuracy of the judgmental forecasts when compared to the statistical forecasts, while the horizontal axis can loosely be interpreted as representing the lack of new information brought to the process by the second forecasting method.

2.3. Correcting judgmental forecasts before combining

When the constituent forecasts in a combination suffer from mean bias the benefits of combination will depend on the relative size and sign of the forecasts' mean errors (i.e. $\bar{Y} - \bar{F}$). If the mean errors of the judgmental and statistical forecasts are given respectively by v and w, then the MSE of the combined forecast will be:

$$\text{MSE} = 0.25[(\sigma_s^2 + \sigma_j^2 + 2r\sigma_s\sigma_j) + (v + w)^2] \quad (7)$$

Thus if $v = -w$ the bias of one forecast will cancel out that of the other. However, if the statistical forecasts are unbiased, but the mean bias of the judgmental forecasts is v^2 units then the combination would only remove 75% of this mean bias. Given the propensity of judgmental forecasts to suffer from biases (Bolger & Harvey, 1998), it may be beneficial to apply correction to them before combining them with the statistical forecasts – that is a correct-then-combine strategy. Indeed, in their seminal paper on combination, Bates and Granger (1969) argued that forecasts should be corrected for bias before being combined – although their suggested correction only involved the removal of mean bias. Since Bates and Granger's paper, much of the published theory on combination has been based on the presumption that the constituent forecasts are unbiased (e.g. Bunn, 1987).

There are, however, a number of reasons why applying a correct then combine strategy involving Theil's correction might diminish the potential gains of combination. First, Theil's correction is also designed to remove regression bias from the MSE of the judgmental forecasts, which will reduce the value of σ_j/σ_s in (5). This means that after applying correction to the judgmental forecasts the probability that combination will be exceed the threshold in (5) and thus improve accuracy is reduced. Put simply, the correction might be so successful that subsequent combination cannot lead to further improvements. Secondly, if Theil's correction successfully removes mean bias from future forecasts then it will also remove the potential benefits of mean errors of opposite signs tending to cancel each other out in the combination. Finally, it is possible that the smoothing effect that Theil's method has on the judgmental forecasts (Goodwin, 1997) will increase the correlation of their errors with those of the statistical forecasts. This would again reduce the potential benefits of combination.

Of course, the effectiveness of applying correction to forecasts made for observations that are yet to be realised depends on the validity of the assumption that the pattern of errors is stationary over time (Moriarty, 1985; Goodwin, 1997). In many practical situations the judgmental forecast errors are unlikely to be stationary. For example, the pattern of errors in periods where foreseeable special events will occur may be different from the pattern in 'normal' periods when the judge has access only to time series information.

In order to compare the relative improvements in accuracy that could be obtained from mechanical integration methods, under conditions where the errors are unlikely to be stationary, the three strategies, (i) correct, (ii) combine and (iii) correct then combine, were first applied to judgmental forecasts obtained from a laboratory experiment. The details of this application are discussed in the following Section.

3. Application of methods to experimental data

3.1. Details of experiment

Judgmental forecasts were obtained from one of the treatments in an experiment reported by Goodwin and Fildes (1999). Subjects in this treatment condition saw a computer screen displaying a graph of the last 30 quarterly sales figures of a hypothetical product. These sales were occasionally affected by promotion campaigns and a bar chart showing past promotion expenditures and details of any expenditure in the next quarter was also displayed. The subjects were asked to use their judgment to produce one period ahead sales forecasts for the next 40 periods. After each forecast had been made the graphs were updated and subjects were informed of the sales that had occurred.

The sixteen subjects, who were finalists on a Business Decision Analysis degree course at the University of the West of England, were randomly assigned to one of eight series which were obtained by varying:

(i) *the complexity of the underlying time series signal* – the simple signal had a constant mean of 300 units, while the complex signal had an upward trend of 1.5 units per quarter (starting from sales of 210 units at period 0) with a multiplicative seasonal pattern with seasonal indices of 0.7, 1.1, 1.3, and 0.9 for quarters 1 to 4, respectively;
(ii) *the level of noise around the signal* – this was either 'low' (independently normally distributed with a mean of 0 and a standard deviation of 18.8) or 'high' (as low noise, but with a standard deviation of 56.4);
(iii) *the effectiveness of the promotion expenditure* – in promotion periods this was either 'weak' (extra sales equal to 0.05× expenditure were added to the underlying time series signal) or 'strong' (extra sales = 0.7× expenditure). Promotions occurred in 21 of the 71 quarters (12 in quarters requiring forecasts).

In post-promotion periods the underlying time series observation was reduced by 50% of the previous promotion period's effect. In practice this might occur where consumers simply bring their purchases forward by one period because of the campaign, but reduce purchases in the subsequent period to compensate (Abraham & Lodish, 1987). At the start of the experiment subjects received written instructions which included advice from the 'sales manager'. This informed them (i) whether or not the sales had a seasonal pattern, (ii) that promotion campaigns might not have a strong effect on sales, but any positive effects were restricted to the quarter in which the campaign took place and (iii) that in quarters following effective promotion campaigns a negative effect on sales was to be expected. As an incentive, a prize of £20 was offered for the most accurate forecasts, after taking into account the estimated level of difficulty associated with forecasting each series.

When the judgmental forecasts (JUDGMENTAL) had been obtained their mechanical integration with statistical methods was carried out as follows. For each of the subjects, the first 15 of their forecasts were used to fit an initial Theil regression model (2). This was then used to produce a corrected forecast for the next period. After each period, this model was then recursively updated to take into account the judgmental forecast and sales for that period. In this way, one-period-ahead corrected forecasts were generated for the last 25 periods (CORRECT).

The statistical time series forecasts were obtained automatically by applying the expert system in the *Forecast Pro* package (Stellwagen & Goodrich, 1994) to the first 45 observations and using the selected method to produce one

period ahead forecasts for the remaining 25 periods. *Forecast Pro* always recommended either simple exponential smoothing or the Holt–Winters method.[2] Subsequently, two other sets of forecasts were obtained for the last 25 periods by taking i) the means of the judgmental forecasts and the statistical time series forecasts (COMBINE) and ii) the means of the Theil corrected judgmental forecasts and the statistical time series forecasts (CORRECT THEN COMBINE).

3.2. Results

The forecasts for the last 25 periods were separated into three categories, depending on the type of period that was being forecast: normal, promotion and post-promotion. The evidence of the original Goodwin and Fildes (1999) study was that forecasters tended to forget about the post-promotion reduction in sales. As they were not therefore making use of information that was unavailable to the statistical methods the results for post-promotion periods are of limited interest in this study and will not be discussed here. However, as we shall see, observations for post-promotion periods still had an important influence on the fitting of the statistical models.

Forecast accuracy for the remaining types of period was measured by calculating the median absolute percentage error (MdAPE) in forecasting the time series *signal* (i.e. the underlying time series signal plus any promotion effects – the forecasters and the statistical methods were not expected to forecast the noise in the series). The MdAPE has been recommended as an error measure by Armstrong and Collopy (1992) when the accuracy of forecasting methods needs to be compared over a number of series. The accuracy of the three mechanical integration methods and the original judgmental forecasts was compared by carrying out, separately for normal and promotion periods, a 2 (between series type)$\times$2 (between noise level)$\times$2 (between promotion strength)$\times$4 (within forecasting method) repeated measures ANOVA on the MdAPEs (there were two replications for each treatment).

Table 1 shows the mean MdAPEs for normal periods (the mean MdAPE of the statistical time series forecasts is also shown for comparison). There were no significant interactions involving forecasting methods in the ANOVA, but there was a highly significant main effect ($F_{3,24}=7.1328$, $P=0.0014$). Comparisons of the methods, using Tukey's honest significant difference (HSD) test, indicated that all three integration methods significantly improved on the original judgmental forecasts (all P values were <0.05). However, there were no significant differences between the three integration methods.

The mean MdAPEs for promotion periods are shown in Table 2 (for brevity these have simply been cross-tabulated with promotion effectiveness). When ANOVA was applied to this data there were two significant interactions involving forecasting method: series$\times$noise$\times$method ($F_{3,24}=4.52$, $P=0.012$) and series$\times$promotion-effectiveness$\times$method ($F_{3,24}=4.43$, $P=0.013$). An analysis of these interactions, again using Tukey's HSD method, found that all of the

Table 1
Mean MdAPEs of methods in normal periods

Method	Mean MdAPE
JUDGMENTAL	11.06
CORRECTION	7.78
COMBINE	7.52
CORRECT THEN COMBINE	5.69
Statistical time series	6.61

[2] *Forecast Pro* has a facility for handling special events. This was deliberately not used here in order to simulate situations where non-time series information is available only to the judgmental forecaster.

Table 2
Mean MdAPEs for promotion periods

Method	Weak promotion	Strong promotion
JUDGMENTAL	11.57	19.29
CORRECT	8.08	16.38
COMBINE	6.32	16.28
CORRECT THEN COMBINE	4.43	16.68
Statistical time series	5.84	20.90

integration methods significantly improved on judgment for the trend-seasonal series where there was either high noise or where promotion effects were weak (all $P<0.05$), though Theil's method failed to improve significantly on judgment for the latter series. In all other cases, there was no significant difference between the methods. Thus in promotion periods the integration methods never significantly degraded the accuracy of the judgmental forecasts and in some cases improved accuracy, even though the judgmental forecaster had exclusive access to information about forthcoming promotions.

3.3. Discussion of laboratory experiment results

Three main results emerge from this laboratory experiment. First, even though data used by the statistical methods was contaminated by observations for special periods, all of the integration methods were still effective in improving on the judgmental forecasts for normal periods. Second, in promotion periods, despite judgmental forecasters having exclusive access to non-time series information, the use of integration still led either to improvements over unaided judgment, or at worst, did not diminish the accuracy of the forecasts. Third, the absence of significant differences between the three integration methods means that there was no evidence to suggest that there was anything to be gained by combining judgment with statistical time series forecasts – simply correcting judgment appeared to be sufficient. The mechanics underlying these results are discussed next.

In normal periods, when the judgmental forecasts tended to vary randomly around the signal – as forecasters reacted to each random movement in the series, the integration methods succeeded by 'averaging out' some of this random variation (as Theil's method did in Goodwin (1997)). This improved the consistency of the forecasts.

However, the Theil-corrected forecasts for normal periods still had slight mean bias, with a predominant tendency to forecast too high (e.g. the mean percentage error for flat series as −2.6%). This bias resulted from contamination of the regression model by observations for 'non-normal' periods. Although this bias tended to be reduced by subsequent combination with the statistical time series forecast (to −1.3% for flat series), the improvements were not sufficient to be significant.

In promotion periods, although the integration methods did not degrade the judgmental forecasts, they were also less successful in improving them for series where the promotion effects were strong. This appears to be a result of a combination of two factors – biased judgmental forecasts and integration methods weakened by the effect of observations for non-promotion periods. Making forecasts for promotion periods will have been particularly difficult where the underlying time series was complex or subject to high levels of noise (Goodwin & Wright, 1993) and in these cases biases were likely to occur. For example, Table 3 shows the median percentage errors in forecasting the signal in promotion periods for series where the promotion effect was strong. It can be seen that,

Table 3
Median percentage error on signal of judgmental forecasts for promotion periods when promotion effect was strong (note only promotion periods occurring in the last 25 periods are considered here)

Subject	Series type	Median percentage error
1	Flat, low noise	−0.29%
2	Flat, low noise	5.81%
3	Flat, high noise	11.86%
4	Flat, high noise	20.16%
5	Trend, seasonal, low noise	22.05%
6	Trend, seasonal, low noise	10.53%
7	Trend, seasonal, high noise	25.50%
8	Trend, seasonal, high noise	36.76%

for the more complex and high noise series, there was a substantial tendency to under forecast. Theil's method is, of course designed to correct this type of bias, while the CORRECT THEN COMBINE strategy should have ensured that the time series pattern was represented in the forecast.

Despite this, there was evidence that the success of the mechanical integration methods in promotion periods, where promotion effects were strong, was blunted by contamination of the models by observations for non-promotion periods – in particular by observations for post-promotion periods. Recall that in post-promotion periods a dip in sales was expected. It seems that, not only did subjects forget about this effect, but they also tended to make higher forecasts for post-promotion periods than for normal periods. On average, for series with strong promotion effects, judgmental forecasts for post-promotion periods were 11.5% higher than those for normal periods! It appeared that subjects tended to anchor on the high sales observed in the preceding promotion period. This meant that judgmental forecasts of high sales were associated both with high actual sales in promotion periods and low actual sales in post-promotion periods, thereby reducing the explanatory power of the Theil regression model. Furthermore, in promotion periods where promotion effects were strong, the statistical time series forecasts were relatively less accurate and, since both these forecasts and the Theil-corrected forecasts tended to underestimate sales, their errors were positively correlated. All of these factors were detrimental to the CORRECT THEN COMBINE method.

The laboratory experiment allowed the effectiveness of the mechanical integration methods to be assessed under controlled conditions. However, the forecasting task employed in the experiment may be atypical of many practical forecasting situations. For example, it only involved a single contextual cue (promotion expenditure) and hard data relating to this cue was supplied to the forecaster. In practice, managers may base their forecasts on a multiplicity of cues from many sources (Lim & O'Connor, 1996), while much of the information relating to these cues may be 'soft', in that it is of questionable reliability, or presented in an informal verbal manner. Furthermore, in 'normal periods' the pattern of sales in the laboratory experiment followed a regular time series pattern, undisturbed by external events. In some practical situations the entire time series may be disturbed by these events to the extent that the time series pattern explains a relative small percentage of the variation in the series. Finally, the laboratory forecasts were only made for one period ahead (many organisations adopt a rolling forecast procedure) and the forecasters had no expert product knowledge or prior information on sales (e.g., as a result of contracts already agreed).

In order to test the integration methods in the more complex circumstances that may apply in many practical contexts judgmental sales forecasts were obtained from two companies. The

next Section describes the application of the methods to this data.

4. Analysis of industrial data

4.1. European textile company

Data was obtained on the monthly forecasts and sales of each of 15 products sold by a European textile manufacturer for the period January 1995 to May 1997 (29 observations). The company manufactures a large number of soft furnishing products for both small and large UK retailers, including one in-house customer. Because the large customers usually specify exact details of their requirements well in advance, sales forecasting is only required for smaller customers and the in-house customer.

The forecasts are produced by the company's sales department, but used by the operations department to plan production. Preliminary forecasts are made six months ahead, but these are regularly fine-tuned as the forecast period approaches. However, because manufacture of the products takes six weeks, the 'final' forecasts, which are the ones analysed here, also have this lead time.

The company usually runs promotion campaigns for its products twice a year in May/June and October/November, but customers also run their own campaigns. Sales staff meet regularly with customers to obtain details of their promotions and other sales information. The forecasters indicated that they used both this market information and past sales history (i.e. time series information) to arrive at their forecasts.

The three forecast integration methods were applied to the data as follows. Because of the six-week production time, the statistical methods could only have access to data up to month t when a forecast for month $t+2$ was required. The methods were fitted to the data for the first 17 months, allowing months 19 to 29 (the last 11 months) to be used for out-of-sample comparisons of the two-period ahead forecasts. As before, the expert system on the *Forecast Pro* package was used to obtain statistical time series forecasts automatically and the package selected simple exponential smoothing for all 15 series. To allow Theil's method some flexibility to adapt to possible changes in judgmental biases over time to the regression equation used in (2) was again recursively updated after each month's sales figure was known. The estimates of a and b at time t were then used to correct the judgmental forecast made for the sales in month $t+2$.

Note that the judgmental forecasters had several advantages over the statistical methods. Not only did they have access to non-time series information, but they could also delay their forecasts until 6 weeks before the forecast period and so make use of informal and preliminary sales information that was available within the statistical method's two month lead time.

The out of sample MdAPEs of the forecasting methods, averaged over the 15 products, are shown in Table 4. The use of significance tests should be treated with caution here as the products were not randomly selected and there may be some dependence between the observations for the different products. Nevertheless, in the light of the laboratory results, significance tests were used to assess (i) whether Theil's correction significantly improved the judgmental forecasts and (ii) whether COMBINE or CORRECT THEN COMBINE led to any greater accuracy than Theil's correction. A one-tailed paired t-test[3] showed that Theil's correction had a significantly lower mean MdAPE than the

[3] A one-tailed test on Theil's method was considered to be justified in the light of the evidence from the laboratory study.

Table 4
Accuracy of textile company sales forecasts

Method		Mean MdAPE
Management judgment	(JUDGMENTAL)	23.8%
Exponential smoothing		25.7%
Theil's method	(CORRECT)	20.9%
Mean of judgment and exponential smoothing	(COMBINE)	21.2%
Mean of Theil and exponential smoothing	(CORRECT THEN COMBINE)	23.3%

judgmental forecasts ($t_{14}=1.97$, $P=0.034$). Table 4 shows also that Theil's method led to the most accurate forecasts of all the methods so there was clearly nothing to be gained by using either form of combination.

4.2. UK-based engineering company

The UK headquarters of an American company, which manufactures and sells drill bits to the international oil industry, provided data on its one-month-ahead judgmental sales forecasts. These are made by personnel who have access to information provided by the sales force. The manager responsible for forecasting estimated that, at the time of making the forecast, on average between 10% and 20% of next month's sales are already known, because of contracts already agreed. Forecasts and outcomes were obtained for each of seven of the company's sales regions for the period January 1993 to December 1994 (24 months).

The first 18 months were used to fit the statistical methods and the last 6 were used for out of sample comparisons. As before, *Forecast Pro* was used to generate the statistical time series forecasts automatically (it always selected simple exponential smoothing), and Theil's regression equation was recursively updated after each sales value was known. Table 5 shows the MdAPEs of the forecasting methods, averaged over the seven series.

Once again a one-tailed paired *t*-test was used to investigate whether Theil's correction led to a lower mean MdAPE than the judgmental forecasts. This suggested that Theil's method did not lead to significant reductions in the mean MdAPE ($t_6=1.35$, $P=0.11$). However, this result should be treated with caution for the reasons stated earlier and also because a sample of only seven sales areas were involved. In fact, Theil's method outperformed both the original judgmental forecasts and the exponential smoothing forecasts in six of the seven series. Once again, combining did not appear to be worthwhile.

Table 5
Accuracy of engineering company sales forecasts

Method		Mean MdAPE
Management judgment	(JUDGMENTAL)	15.4%
Exponential smoothing		22.2%
Theil's method	(CORRECT)	11.8%
Mean of judgment and exponential smoothing	(COMBINE)	14.9%
Mean of Theil and exponential smoothing	(CORRECT THEN COMBINE)	14.9%

4.3. Discussion of industrial forecasting results

As in the laboratory study, Theil's correction method (CORRECT) played a valuable role in improving the accuracy of the judgmental forecasts. It improved the judgmental forecasts for 15 out of the 22 industrial series and rendered the use of combination redundant. This result is consistent with other studies of forecasters in the field which have shown the effectiveness of correction, relative to statistical forecasts or combination, albeit by employing slightly more complex, correction methods (Fildes, 1991; Lawrence, O'Connor & Edmundson, in press). Why was combination not useful in the company forecasts presented here?

An analysis of the forecast errors showed that, after correction, the judgmental forecasts tend to be more accurate than the exponential smoothing forecasts and the errors of the two types of forecast were highly correlated (e.g., for the textile company the mean value of r was 0.84). As Fig. 1 shows, both of these factors were to the detriment of combination. The underlying mechanics can be seen in Fig. 2, which shows the out-of-sample forecasts for one of the products. While there is clear evidence that the judgmental forecaster is using non-time series information to anticipate movements in sales, these forecasts tend to be too high. (Structured interviews with the forecasters provided no evidence that this bias was deliberately created for political reasons or because the forecast loss function was perceived to be asymmetric.) However, once the over forecasting bias in the judgmental forecasts has been mitigated by Theil's correction, it can be seen

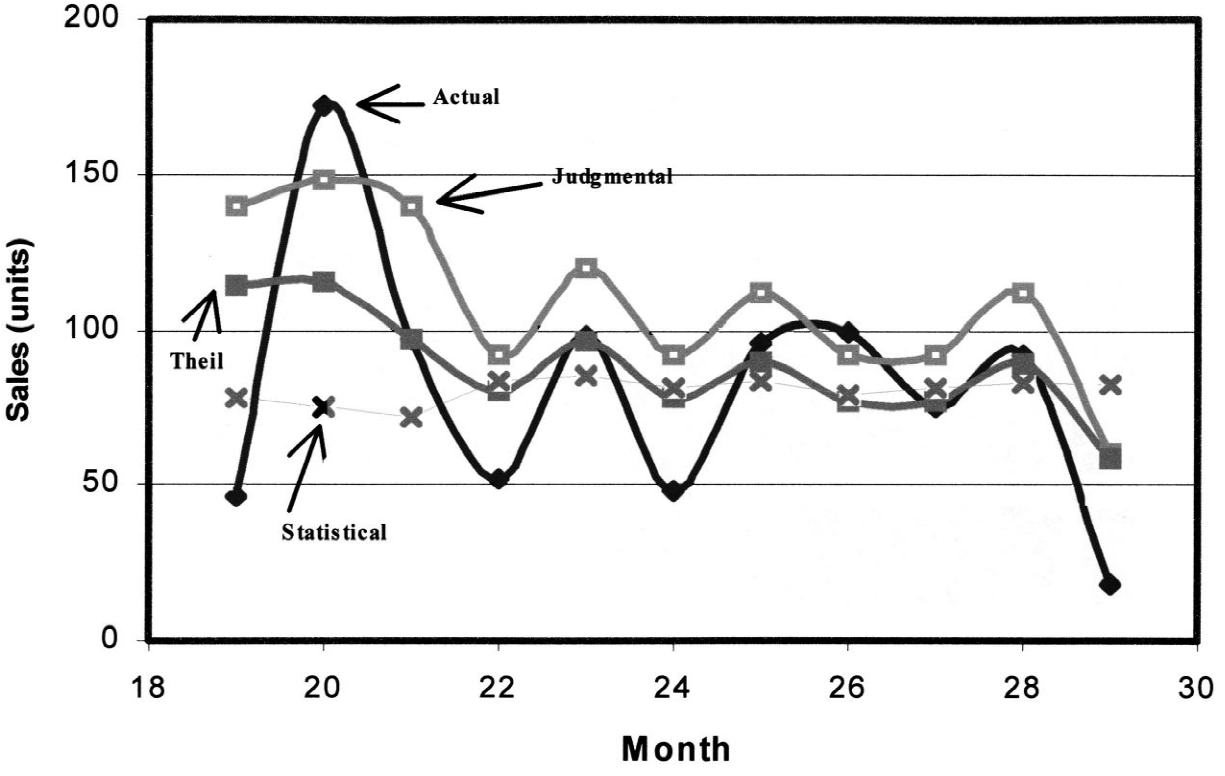

Fig. 2. Actual sales and forecasts for textile company product.

that exponential smoothing has nothing to add to the ability of the forecasts to explain movements in the sales series.

Clearly, there are important differences between the laboratory and industrial data. Unlike the laboratory task, the industrial forecasts were characterised by access to continuous non-time series information (from multiple sources) about events whose effects tended to submerge the relatively 'weak' time series pattern. For the engineering company this non-time series information included prior knowledge of some sales. In both companies the forecasters had expert product knowledge and experience of the forecasting tasks so they were able to make good use of the non-time series information. In the case of the textile company the judgmental forecasters had a shorter lead time than the statistical method and so were able to use more recent non-time series information. Contrast this with the laboratory study where the series had a strong time series pattern, non-time series information that was only available sporadically and inexperienced, non-expert forecasters who made inefficient use of this information. Nevertheless, in both of these very different contexts the use of correction appeared to be effective and there was no evidence that greater accuracy could be achieved through combination.

5. Conclusions

This paper is based on the premise that the use of judgment in forecasting is justified when non-time series information, which may be difficult to model statistically, has high predictive power. However, the limitations of judgment mean that integration with a statistical method may be desirable. The results presented here suggest that, where useful, but difficult-to-model, non-time series information is available, the most appropriate role of statistical methods is to correct judgmental forecasts. The simple correction method used in the study was, in many cases, sufficient to obtain significant improvements in accuracy and there was little to be gained by obtaining independent statistical time-series forecasts and then combining these with the judgmental forecasts (or corrected judgmental forecasts). Moreover, the correction method appears to be robust in that it can still improve forecasts, or at worst not degrade them, even when different biases apply in different types of period – though its effectiveness is reduced by these variations. (The method may be less robust when the nature of the biases changes in a non-reversionary way over time (Goodwin, 1997)).

Of course, the extent to which these conclusions can be generalised is limited by the conditions which applied in the laboratory experiment and in the two companies studied. In particular, the relatively small sample size used in the laboratory study may have meant that the effectiveness of the CORRECT THEN COMBINE strategy was underestimated, though, of course, even this strategy involved correction. Indeed, taken together, the results presented here suggest that, relative to combination, correction may have been under represented as a recommended technique for harnessing the complementary strengths of judgment and statistical methods.

Appendix

Note that, in this study, the correction method was applied indiscriminately to all of the series in order to compare their performance. An alternative approach would have involved testing the in-sample judgmental forecasts for bias before deciding whether to apply Theil's correction. To achieve this an F-test can be employed to test the joint hypothesis that $a = 0$ and $b = 1$ in (2) (Johnston, 1972, p. 28). However, evidence from Goodwin (1997, 1998) suggests that

this test has little value in predicting whether the correction will improve judgmental forecasts in the out-of sample periods. The limitations of this test have also been discussed in the economics 'rational expectations' literature (Liu & Maddala, 1992; Lopes, 1998). Research is currently being undertaken to try to develop improved methods for identifying when Theil's correction is appropriate. In the absence of these methods, the evidence of this study is that indiscriminate correction of judgmental forecasts is likely to be worth carrying out.

References

Abraham, M. M., & Lodish, L. M. (1987). PROMOTER: An automated promotion evaluation system. *Marketing Science 6*, 101–123.

Ahlburg, D. A. (1984). Forecasting evaluation and improvement using Theil's decomposition. *Journal of Forecasting 3*, 345–351.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting 8*, 69–80.

Armstrong, J. S., & Collopy, F. (1998). Integration of statistical methods and judgment for time series forecasting: principles from empirical research. In: Wright, G., & Goodwin, P. (Eds.), Forecasting with judgment, John Wiley, Chichester, pp. 269–293.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly 20*, 451–468.

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

Bolger, F., & Harvey, N. (1998). Heuristics and biases in judgmental forecasting. In: Wright, G., & Goodwin, P. (Eds.), Forecasting with judgment, John Wiley, Chichester, pp. 113–137.

Bunn, D. (1987). Expert use of forecasts: bootstrapping and linear models. In: Wright, G., & Ayton, P. (Eds.), Judgmental forecasting, John Wiley, Chichester, pp. 229–241.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting 5*, 559–583.

de Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research 120*, 190–204.

Elgers, P. T., May, H. L., & Murray, D. (1995). Note on adjustments to analysts' earning forecasts based upon systematic cross-sectional components of prior-period errors. *Management Science 41*, 1392–1396.

Fildes, R. (1991). Efficient use of information in the formation of subjective industry forecasts. *Journal of Forecasting 10*, 597–617.

Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: a review of the guidance provided by research. *International Journal of Forecasting 9*, 147–161.

Goodwin, P. (1996). Statistical correction of judgmental point forecasts and decisions. *Omega: International Journal of Management Science 24*, 551–559.

Goodwin, P. (1997). Adjusting judgmental extrapolations using Theil's method and discounted weighted regression. *Journal of Forecasting 16*, 37–46.

Goodwin, P. (1998). Interfacing judgmental forecasts with statistical methods. Unpublished PhD thesis, University of Lancaster, U.K.

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making 12*, 37–53.

Johnston, J. (1972). Econometric methods, 2nd ed., McGraw-Hill, New York.

Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science 32*, 1521–1532.

Lawrence, M. J., O'Connor, M. and Edmundson, R. (in press). A field study of sales forecasting accuracy and processes. European Journal of Operational Research.

Lim, J., & O'Connor, M. (1995). Judgmental adjustment of initial forecasts – its effectiveness and biases. *Journal of Behavioral Decision Making 8*, 149–168.

Lim, J., & O'Connor, M. (1996). Judgmental forecasting with time series and causal information. *International Journal of Forecasting 12*, 139–153.

Liu, P. C., & Maddala, G. S. (1992). Rationality of survey data and tests for market efficiency in the foreign exchange markets. *Journal of International Money and Finance 11*, 366–381.

Lopes, A. S. (1998). On the 'restricted cointegration test' as a test of the rational expectations hypothesis. *Applied Economics 30*, 269–278.

Moriarty, M. M. (1985). Design features of forecasting systems involving management judgments. *Journal of Marketing Research 22*, 353–364.

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgmental forecasting in times of change. *International Journal of Forecasting 9*, 163–172.

Shaffer, S. (1998). Information content of forecast errors. *Economics Letters 59*, 45–48.

Stellwagen, E. A., & Goodrich, R. L. (1994). Forecast Pro for Windows, Business Forecast Systems Inc, Belmont, MA.

Theil, H. (1971). Applied economic forecasting, North-Holland Publishing Company, Amsterdam.

Watson, M. C. (1996). Forecasting in the Scottish electronics industry. *International Journal of Forecasting 12*, 361–371.

Biography: Paul GOODWIN is Principal Lecturer in Operational Research at the University of the West of England. His research interests focus on the role of judgment in forecasting and decision making and he received his PhD from Lancaster University in 1998. He is the co-author of *Decision Analysis for Management Judgment* (2nd edition) published by Wiley and co-editor of *Forecasting with Judgment*, also published by Wiley. He has published articles in a number of academic journals including the *International Journal of Forecasting*, the *Journal of Forecasting*, the *Journal of Behavioral Decision Making* and *Omega*.