Journal of Forecasting, Vol. 9, 305–314 (1990)

Decomposition; A Strategy For Judgemental Forecasting

R. H. EDMUNDSON
University of New South Wales, Australia

ABSTRACT

This paper reports the results of studies concerning the accuracy and efficiency of time-series extrapolation decisions made with the assistance of an interactive graphical tool called GRAFFECT. The tool facilitates the decomposition of the extrapolation task by permitting the serial decomposition of the cue data as the task proceeds. GRAFFECT uses an interactive graphical interface controlled substantially with the use of a mouse. The extrapolation task is divided into the following: (1) trend modelling and extrapolation, (2) seasonal pattern modelling, and (3) extrapolation from the noise residual series. As each component is modelled its effect is stored and the information is washed out of the cue series. The ultimate forecast is produced by automatic recomposition of the judgementally determined components. The results show a significant improvement in forecast accuracy over unaided judgment, resulting in a subjective extrapolation that betters deseasonalized single exponential smoothing.

KEY WORDS Forecasting Judgement Decomposition

OVERVIEW OF THE LITERATURE

There has been only limited research reported concerning the accuracy of judgemental extrapolation, but there is strong evidence of a distrust of judgemental processes to be found in the time-series extrapolation literature. For instance, Makridakis (1981, page 309) stated: 'Complaining, for instance, about the pure predicitive ability of statistical forecasting makes little sense when the alternative, i.e. human judgement, can be worse and is definitely more costly.'

Despite the view attributed to Makridakis there is evidence that practitioners use judgemental processes. Dalrymple (forthcoming) reviews the not insubstantial literature,[1] and concluded that about 70% of commercial forecasts are made subjectively. Armstrong (1985), in commenting on the literature, said:

"Most forecasts are made with subjective methods. It also seems that the more important

[1] See, for example, Cerullo and Avila (1975), Rothe (1978), Sparkes and McHugh (1984), Mentzer and Cox (1984) and Dalrymple (1985).

0277-6693/90/040305-10$05.00

Received April 1990

the forecast, the more likely it is that subjective methods will be used. Yet in many of these situations, objective methods would be more appropriate" (p. 73.)

It may be argued that the commercial forecasts spoken of by Dalrymple are forecasts taking an account of commercial and economic data as well as the time-series information available, while the methods referred to by Armstrong are, in the main, extrapolative time series techniques.

Carbone *et al.* (1983) found that subjective adjustment of statistical forecasts did not give rise to improved accuracy, but the subjects did not make the revision in the light of any extra information. Lawrence *et al.* (1985), using the database of 111 time series used by Makridakis *et al.* (1982), concluded that judgemental extrapolation was as accurate, on average, as the better of the statistical techniques from the so-called 'M-competition' (Makridakis *et al.*, 1982). That result was not confirmed by Carbone and Gorr (1985), who used a sample of time series from the database and concluded that judgemental extrapolation was poorer than statistical extrapolation.

There may be many explanations for the differences and results between the Lawrence *et al.* and the Carbone and Gorr studies. One major difference between the two would appear to lie in the decision strategy adopted by the subjects. In the Lawrence *et al.* study the subjects were required to use a prescribed decomposed approach in which the trend observed in the series was extrapolated, then the seasonal pattern was identified and superimposed upon the extrapolated trend. This approach has been used in subsequent replications (O'Connor, 1986) in which the findings of the Lawrence *et al.* study have been confirmed. The subjects in the Carbone and Gorr study were not required to follow any particular strategy, and the report of the study does not indicate whether a decomposed approach was used by any of the subjects.

DECOMPOSITION OF THE DECISION

There are a number of pieces of circumstantial evidence, of varying persuasive power, to support the proposition that decomposition of the extrapolation task would improve judgemental performance. There is a general subset of the human information-processing literature which has addressed the issue of decomposition *per se*. Slovic *et al.* (1977) stated the case for decomposition based upon reducing the cognitive load on the decision maker. Armstrong *et al.* (1975) found that twelve out of thirteen responses to almanac-type problems were improved by decomposition of the decision. This finding was not entirely supported by Lyness and Cornelius (1982), who used a simulated decision setting and concluded that holistic judgement may be as accurate as decomposed judgement except in complex circumstances. Lyness and Cornelius further concluded that the algorithmic synthesis of the decomposed judgement outperformed judgemental synthesis.

Given that decomposition as a strategy may lead to improved extrapolation, it is pertinent to consider how well the sub-tasks in the decision (i.e. trend and seasonal identification) would be performed judgementally.

TREND IDENTIFICATION

Eggleton (1982) indicated that the human forecaster may be affected by shortcomings in the identification of trend. He found that subjects underestimated outcomes in a rising trend series

and considered this as evidence of either a parameter error or of an heuristic error in the judgement. Lawrence and Makridakis (1989) examined the extrapolation of trend from seven points in a graphical display and concluded that subjects understated the slope of the series, especially for downward-sloping series. Both the Eggleton and the Lawrence and Makridakis studies compared the *extrapolated* trend with the trend component in the *cue data*, and the results are therefore confounded by the possibility that the extrapolation took account of other factors such as the possibility of a failure in the assumption of constancy. Mosteller *et al.* (1981) showed that subjects were able to perform well in the task of fitting a straight line to a scatter plot of data. This task is somewhat different to identifying trend in the time series but there are some parallels. Despite reservations concerning a slight (and statistically not significant) tendency to overstate the slope of the line and to fit a line to highly scattered data by minimizing the principal component rather than the *y*-wards deviation, the Mosteller *et al.* findings indicate that subjects might perform well in identifying trend in a time series.

SEASONAL IDENTIFICATION

The identification of the seasonal pattern in the data is not so clear cut. Eggleton (1976) showed that subjects were unable to distinguish between alternating sequences and random sequences in contrived time-series data. The link between 'alternating sequences' and seasonal patterns is somewhat strained but it provides the nearest approach available. Eggleton's results are not really generalizable to the normal time series forecasting context because he applied severe time constraints to his subjects. His results are also somewhat at odds with the folklore that ascribes good pattern processing capabilities to mankind (see, for example, Simon and Sumner, 1968, who describe pattern recognition as a survival skill). In other fields such as artificial vision and speech recognition it has certainly been recognized that there are extreme difficulties in achieving even the quite limted pattern recognition capabilities of, say, a four-year-old-child. None of this is an indication, of course, that human judges would perform well identifying a seasonal pattern in time series. Although there might be some persuasive evidence that a judge would detect a pattern that existed in the time series, it could be argued that the pattern-recognition capabilities as a survival skill might tend to an erroneous identification of a seasonal pattern where none existed. There are difficulties in examining human ability to detect seasonal patterns in time series, because even in the case of mathematically contrived data for which the underlying process is known it is possible that random events simulate the effect of a pattern in the short term. The problem is different in the applied setting, where it is not possible to determine the 'true' seasonal component of a time series, and any underlying porcess may not remain constant long enough for the short-term effects to wash out.

A DECISION AID FOR FORECASTING TIME SERIES

In the light of the foregoing it was determined that a decision aid be constructed to assist in the judgemental extrapolation of a time series. The decision aid, called GRAFFECT, runs on IBM PCs or equivalent and uses a Microsoft mouse to facilite data entry for modelling and extrapolation tasks. Thus the model judgemental tasks are supported by graphical cues, and require the production of graphical decision outcomes. The decision aid was designed to enforce a decomposition of the extrapolation task in line with classical time-series analysis, that is, into trend identification, seasonal model identification and extrapolation from the

residual 'noise' component. The trend and seasonal models are stored separately and used in generating the resultant forecast in combination with the extrapolation with the noise residual. The data presentation screens used for trend identification and for extrapolation from the noise residual were conventional graphical plots (Figure 1), whereas the data presentation for seasonal identification superimposed the four years' data shown on a single twelve-month wide plot (Figure 2).

The tool allows the subjects to iterate through any number of attempts to model the trend of seasonal components and constructs a multiplicative model of the final forecast. In the case of trend modelling GRAFFECT permits the modelling of linear trend components with a single

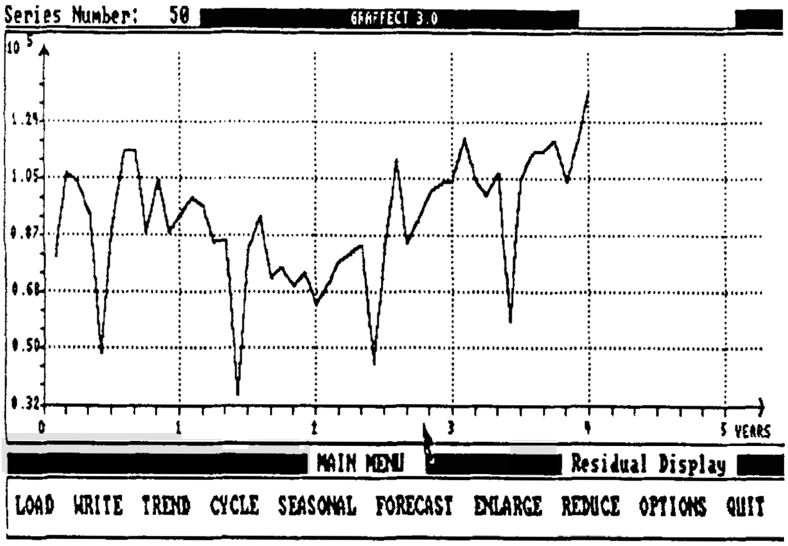

Figure 1. Trend and forecasting screen

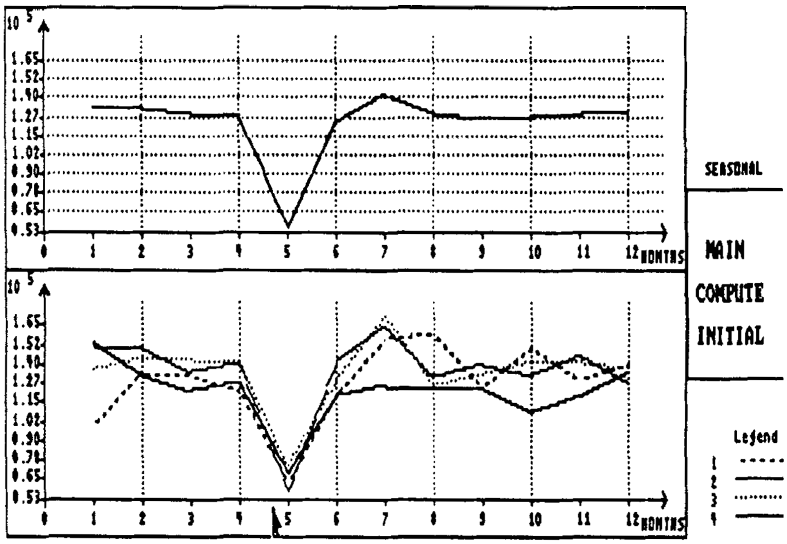

Figure 2. Seasonal identification screen

turning point in the history data. The forecast trend is not constrained to be the same as the 'most recent model', although that is the default condition, but may have a steeper or a shallower slope. Once a satisfactory trend model is identified, the effect of that is washed out of the cue data, so that further modelling procedures can take place in the absence of any confounding effects from the trend information.

The data display used for seasonal identification permits a model to be constructed in the upper half of the screen while in the lower half the data may be displayed net of that model. Thus the judge may iterate towards the model that appears to leave as little signal as possible in the resulting series.

EVALUATION OF THE DECISION AID

Pilot studies conducted with early versions of the GRAFFECT indicated a likely improvement in accuracy over forecasts supported by hard-copy plots (GRAPH). It was therefore possible to state the first two research hypotheses as:

- H1 Forecasts made using the GRAFFECT decision aid will not be more accurate than those produced using GRAPH.
- H2 Forecasts made using the GRAFFECT decision aid will not be more accurate than those produced using deseasonalized single exponential smoothing.

The comparative accuracy of experienced and novice subjects has implications for the use of GRAFFECT, and for the development of the tool for inexperienced users. The work of Lawrence *et al.* (1985) did not find a significant difference in the accuracy of experienced and novice forecasters using the GRAPH method. Again, pilot studies conducted during development indicated that novice forecasters did not perform differently to experienced forecasters. The research hypothesis is therefore:

- H3 Experienced subjects will perform no better using the GRAFFECT decision aid than inexperienced subjects.

The costs of forecasting are not limited to the costs of the errors in the forecast. The use of resources must also be considered in determining the desirability of a particular forecasting method. In that light, it was necessary to establish what effect the use of the GRAFFECT decision aid had on the time taken to forecast a time series.

It is not possible to compare the use of resources in the deseasonalized single exponential smoothing method with that in a judgmental method. Deseasonalized single exponential smoothing is a simple, computer-based method which may utilize little or no human resources for the extrapolation phase. The trade-off between human and computer resources is complex, and must include consideration of a number of qualitative factors such as the implications for management understanding of the results and the effect on the political scenario.

This study was directed, primarily, at the evaluation of the decision aid against the hard-copy graphical technique. It was limited to data from experienced subjects in order to avoid confounding the results with effects from any learning curve experienced by novice subjects.

There was no *a priori* basis for assuming that the GRAFFECT decision aid would permit faster or slower forecasting than would the provision of hard-copy plots. The hypothesis addressed is therefore expressed in 'two-tailed' form:

- H4 The use of the GRAFFECT decision aid will have no effect on the time taken to forecast time series.

DESCRIPTION OF THE EXPERIMENT

The time series data

The time series used came from the database of time series provided by the authors of the 'M-Competition'. In this case, all 68 of the monthly series from that database were used.

Data collection

The forecasts for deseasonalized single exponential smoothing were obtained from the 'M-Competition' and the judgmental forecasts using hard-copy graphical data presentation were obtained from the Lawrence *et al.* (1985) study. The 'experienced' subjects used in the studies were three postgraduate students with a knowledge of time-series analysis, and with prior experience in the use of GRAFFECT. The subjects had undertaken postgraduate courses in operations research, and had general commercial experience. Each subject forecast all 68 time series in the database, and their forecasts for each were randomly assigned (without replacement) to the three studies conducted to compare the accuracy with judgement using hard-copy plots (Graph), with deseasonalized single exponential smoothing (DSE) and with novice forecasters using GRAFFECT.

The novice subjects were 35 postgraduate students with no prior experience of time-series analysis, but they had general commercial experience. The subjects were enrolled in a decision support systems course at the University of New South Wales. They were volunteers who agreed to use the GRAFFECT decision aid following a brief presentation of the tool as part of their course work.

The subjects were given an introduction to the terms and basic concepts of time-series analysis. That is, the idea that a time series is expressible as trend, cycle and noise was introduced. The tool was demonstrated to the subjects. The instruction and demonstration was carried out in a single session of 20 minutes.

The 68 time series were allocated to the 35 students taking part according to the following rules:

(1) Five time series were allocated to each subject.
(2) It was ensured that every time series occurred at least once in the set of time series numbered 3 to 5 in the subjects' lists.
(3) Subject to the above, and the requirement that no series would occur more than once in each list, the time series were 'randomly' allocated to the lists of five.

The subjects forecast the series in the order in which they appeared on their particular 'assignment' list. The lists were randomly assigned to subjects.

The procedure ensured that there was little or no bias effect expected as a result of the position of time series in the sequence of the forecasts. The first two forecasts from each subject were eliminated as practice, and the remainder of the completed forecasts were accepted as candidate forecasts. The file of candidate forecasts contained a minimum of one forecast for each series, and for some series there were up to four forecasts.

In order to permit the use of a paired t-test a single forecast attempt was required for each series. Multiple observations for individual time series were eliminated by random selection.

The subjects followed the procedure laid down in the GRAFFECT User Manual, forecasting each series by first considering the trend component, then the seasonal component, and finally forecasting from the residual noise series. This was similar to the sequence used by Lawrence *et al.* (1985), who required subjects to first consider the trend component followed by the seasonal pattern.

No time restriction was placed on the completion of the task. There was no outcome feedback provided until all series had been forecast. As with the Lawrence *et al.* (1985) study, the subjects were not given any information outside the values in the time series. Thus, they had no indication of the source of nature of the time series, and they did not know the time periods from which the experimental series were drawn. This placed the subjects in an identical position, with regard to information about the time series, as those in the Lawrence *et al.* (1985) study. It also placed the decisions of those subjects on the same footing as the extrapolations of the statistical methods which cannot take any external data into account.

Analysis of the data

Following Lawrence *et al.* (1985), the data were manipulated to generate the Mean Absolute Percentage Error of forecast (MAPE) for the two forecast horizons, months 1–6 and 7–12. MAPE was chosen as the error measure in order to ensure consistency with Lawrence *et al.*, and because it has been shown that measures based on perecentage error are widely accepted.[2] The use of MAPE also avoids the heavy emphasis that a squared error measure such as Mean Squared Error places on extreme errors. The results of Lawrence *et al.* (1985) indicated that judgmental forecasts avoided the very large errors associated with statistical processes for certain series. To apply a squared measure in those circumstances would be seen to give advantage to the judgmental processes under investigation.

In order to reduce the effect on results from individual time series that was evident in the previous studies on this database, it was necessary to use a 'paired' method of comparison. In the light of prior results, hypotheses 1 to 3 had been stated in 'one-tailed' form. Thus the method of analysis chosen was the one-tailed, paired *t*-test.

RESULTS

The forecasts of the experienced subjects had been randomly assigned to three data sets. It had been determined before analysis took place that the first set was to be used for significance testing in the comparison with GRAPH, the second with DSE, and the third with the forecasts from the novice subjects. Table I contains the MAPE's for each of the data sets.

The results show that human judgment using the GRAFFECT decision aid provides error rates lower than any single forecasting method reported in the M-Competition. Table II displays the error rates for the methods under test, and for several other statistical methods reported in the 'M-Competition'. For simplicity, the error rate for expert forecasters using GRAFFECT is reported as the mean of the rates reported in Table I.

Table I. GRAFFECT error rates for the three samples

Data Set	MAPE for Months 1–6	7–12	1–12
GRAFFECT versus GRAPH set	10.8	13.7	12.3
GRAFFECT versus DSE set	9.5	12.7	11.1
EXPERT versus NOVICE set	9.9	13.4	11.7
Mean Error	10.1	13.3	11.7

[2] See Armstrong (1985, p. 360).

Table II. Average MAPE over 68 time series

Method	MAPE for months 1–6	7–12	1–12
GRAFFECT			
Expert (mean)	10.1	13.3	11.7
Novice	10.0	14.0	12.0
DSE	11.0	14.2	12.6
GRAPH	11.6	16.5	14.1
Box–Jenkins	11.3	16.3	13.8
D ARR EXP	10.8	13.8	12.3
Bayes F	10.7	14.5	12.6

As can been seen, the MAPE for GRAFFECT was approximately 1 percentage point lower than deseasonalized single exponential smoothing (when rounding errors are taken into account), for both the 1–6 and the 7–12 forecast horizons. In the 'M-Competition' the deseasonalized single exponential smoothing method was one of the best methods overall. However, for the 1–6 horizon it was bettered by deseasonalized adaptive response rate exponential smoothing and the Bayesian F method. GRAFFECT shows a 6% improvement in the 1–6 case, and a 4% improvement in the 7–12 case, over the best previously reported methods for each forecast horizon. Those results must be interpreted in the light of Table III and the discussion that follows.

It is clear from Table II that provision of the GRAFFECT decision aid did lead to a significant improvement in performance in judgmental forecasting. The improvement obtained by providing experienced subjects with GRAFFECT over the results from similarly experienced subjects with GRAPH is 13% for the 1–6 case and 20% for the 7–12 case. The table also indicates that there is little difference between experienced and novice subjects for months 1–6, but that the novice subjects were somewhat less accurate over the longer forecast horizon. The paired t-tests carried out are reported in Table 3.

Given the above, there is no problem in rejecting Hypothesis H1 with a confidence level of 99% or better for both time horizons considered. That is, the GRAFFECT decision aid clearly produces results that are better than the Graph technique by 13% or more, and that the improvement is statistically significant.

The comparison with deseasonalized single exponential smoothing was not quite so clear. The level of confidence with which hypothesis H2 could be rejected (94%) bordered on a commonly adopted lower bound of significance of 95%. For the 7–12 case there was no opportunity to reject hypothesis H2 with a reasonable level of confidence (the significance level being 89%). The sample data set used for the examination of hypothesis H2 had the lowest error rate of the samples. This observation would tend to indicate that claims that GRAFFECT

Table III. Paired T-test results

GRAFFECT expert with	One-tailed probability 1–6	7–12
GRAPH	0.005	0.001
DSE	0.060	0.110
GRAFFECT novice	n.s.	n.s.

Table IV. Standard deviations of MAPE error

Method	1–6	7–12
	0.074	0.112
GRAFFECT	0.079	0.113
	0.085	0.112
DSE	0.115	0.150
GRAPH	0.095	0.142

Table V. Time for forecast (seconds)

Method	Mean time	Std dev.
GRAFFECT	161.7	42
Graph	265.4	62

has a lower error rate than DSE would be suspect. On the other hand, all three sample data sets had lower MAPEs than DSE, and this would tend to lend support to the supposition that GRAFFECT is more accurate than DSE. The difficulty in interpreting the results of the study without equivocation centres on the normative question of what level of significance should be adopted. The conservative view of the results would be that any improvement was not statistically significant, but that view generates a real possibility that the null hypothesis is mistakenly accepted in marginal cases such as this.

Hypothesis H3 cannot be rejected. There is no statistical difference between the results obtained by experienced and novice forecasters. The higher error rate of the novice forecasters over the longer forecast horizon, though not statistically significant, raised an implication that they might not be dealing with trend in the same way as the experienced forecasters. *Post hoc* analysis revealed that in only two of the 68 series did the subjects damp the trend functions that they had identified.[3] The experienced forecasters damped the trend in 10 series.

The analysis of the error data revealed that the GRAFFECT method obtained its low levels of MAPE error without losing the advantage of the relatively low standard deviation of error that judgmental forecasting has exhibited in the Lawrence *et al.* (1985) study.

Table IV displays the standard deviations of error for the three GRAFFECT data sets, DSE, and GRAPH for the two forecast horizons. This would indicate that there may be incentive to use judgmental techniques, and especially methods such as GRAFFECT, for the forecasting of critical time series in which there is a desire to avoid very large error rates.

The evaluation of Hypothesis 4, concerning the time taken to make a judgmental extrapolation was carried out using a two-tailed *t*-test, paired on time series. Table V summarizes the results which revealed that the time taken with the GRAFFECT decision aid was about 60% of that taken with the hard-copy graph. The difference between the methods was significant at the $p = 0.000$ level.

The Hypothesis H4 can be rejected with high confidence, and the conclusion made that the data presentation methods in GRAFFECT resulted in substantial time savings.

CONCLUSIONS

The above analysis has demonstrated that the provision of interactive graphical support for the forecasting decision can improve judgmental forecasting by between 13% and 20% depending upon the forecast horizon. There is also evidence that a similar, but smaller,

[3] The two series with damped trend came from the same subject. Examination of the data revealed that the subject had damped the trend for the second to fifth series in the sequence forecast. This might therefore have been associated with experimenting with the tool rather than a real decision to damp the trend.

improvement is obtainable over deseasonalized single exponential smoothing. There is no evidence, however, that this improvement is statistically significant, especially for the 7–12 forecast horizon. Again, judgmental processes have been shown to be somewhat more controlled (see also Lawrence *et al.* 1985) in that they have lower standard deviations of error.

The results obtained in the studies reported raise further questions of interest. The major issues arising concern the process by which the improvement in judgment came about, and the effect of trend on the accuracy of judgmental extrapolations.

REFERENCES

Armstrong, J. S., Denniston, W. B. Jr, and Gordon, M. M., 'The use of the decomposition principle in making judgments', *Organizational Behavior and Human Performance*, **13** (1975), 257–63.

Armstrong, J. S., *Long-Range Forecasting: from Crystal Ball to Computer*, 2nd edn, New York: John Wiley (1985).

Carbone, R., Anderson, A., Corriveau, Y. and Corson, P. P., 'Comparing for different time series methods the value of technical expertise, individual analysis, and judgmental adjustment', *Management Science*, **29**, (1983), 559–66.

Dalrymple, D. J., 'Sales forecasting practices: results from a United States survey', *International Journal of Forecasting*, forthcoming.

Eggleton, I. R. C., 'Intuitive time series extrapolation', *Journal of Accounting Research*, **20** (1982), 68–102.

Eggleton, I. R. C., 'Patterns, prototypes, and predictions: an exploratory study', *Selected Studies on Human Information Processing in Accounting* (1976), Supplement to *Journal of Accounting Research*, **14**, 68–131.

Lawrence, M. J. and Makridakis, S., 'Factors affecting judgmental forecasts and confidence intervals', *Organizational Behavior and Human Decision Processes*, **42** (1989), 172–87.

Lawrence, M. J., Edmundson, R. H. and O'Connor M. J., 'An examination of judgmental time series extrapolation', *Journal of Forecasting*, **2** 2, (1985), 169–179.

Lyness, K. S. and Cornelius, E. T., III, 'A comparison of holistic and decomposed strategies in a performance rating simulation', *Organizational Behavior and Human Performance*, **29** (1982), 21–38.

Makridakis, S., 'Forecasting accuracy and the assumption of constancy', *Omega*, **9**, 3 (1981), 307–11.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R., 'The accuracy of extrapolative (time series) methods: results of a forecasting competition', *Journal of Forecasting*, **1**, No. 2 (1982).

Mosteller, F., Siegel, A. F., Trapido, E. and Youtz, C., 'Eye fitting straight lines', *The American Statistician*, **35**, 3 (1981), 150–52.

O'Connor, M. J., *An examination of the accuracy and determinants of confidence levels in judgmental time series forecasting*, PhD Dissertation, University of New South Wales (1986).

Simon, H. A. and Sumner, R. K., 'Pattern in music, in Kleinmuntz (ed), *Formal Representation of Human Judgment*, New York: John Wiley (1968), 219–50.

Slovic, P., Fischhoff, B. and Lichtenstein, S., 'Behavioral decision theory', *Ann. Rev. Psychol.*, **28** (1977), 1–39.

Author's biography:

Bob Edmundson is an Associate Professor of Information Systems and Director of Business Information Technology at the University of New South Wales. His research interests include DSS design and information systems design architecture.

Author's address:

R. H. Edmundson, School of Information Systems, University of New South Wales, PO Box 1, Kensington NSW 2033, Australia.