Direct rolling training for improving judgmental time series forecasting performance

Fotios Petropoulos[1,*], Paul Goodwin[2] and Robert Fildes[1]

[1] *Lancaster Centre for Forecasting, Lancaster University*

[2] *School of Management, University of Bath*

* corresponding author: f.petropoulos@lancaster.ac.uk

Abstract

Several biases and inefficiencies are commonly associated with the judgmental extrapolation of time series. This study examines the effectiveness of using a rolling training approach to improve the accuracy of such forecasts. In an experiment forecasters were asked to make multiple judgmental extrapolations for a set of time series from different time origins. The time series were either stationary, trended or seasonal or trended and seasonal and had either low or high levels of noise. The experiment consisted of two alternating treatments. For each series in turn, the participants were either unaided and received no feedback or they were provided with feedback. In the latter case, following submission of each set of forecasts, the true outcomes and performance feedback were provided. Depending on the group the participants were in, the performance feedback contained information either on the bias or the accuracy of their judgmental forecasts. The objective was to provide a direct training scheme, enabling forecasters to better understand the underlying pattern of the data by learning directly from their forecast errors. Analysis of the results indicated that the rolling training approach is an effective method for enhancing judgmental forecasting accuracy. However, feedback on bias was more effective than feedback on accuracy and, on average it led to more accurate forecasts than unaided judgment on all types of time series.

Keywords: judgmental forecasting, unaided judgments, rolling training, feedback, time series

1 Introduction

Surveys suggest that forecasts based either wholly or in part on management judgment play a major role in company decision making (e.g. [1]). Sometimes the judgmental inputs may take the form of adjustments to statistical forecasts, ostensibly to take into account special factors that were not considered by the statistical forecast [2]. The integration of judgmental and statistical forecasts has been reviewed extensively [3]. However, in some circumstances, judgment may be the only process involved in producing the forecasts. One example of this is the use of judgment to extrapolate from time series data to produce point forecasts, where no other information (except perhaps variable labels such as 'sales' or 'costs') is provided. This type of task has been the subject of much research over the last thirty years and a number of biases associated with judgmental extrapolation have been identified. These include tendencies to overweight the most recent observation (e.g. [4]) to underestimate growth and decay in series [5] and a propensity to see systematic patterns in the noise associated with series.

A number of strategies have been explored to try to mitigate these biases [6]. These include decomposing the time series into its separate elements (e.g. trend and seasonal components) and asking for separate forecasts for each element [7]. Another strategy would be the representation of the data in either tabular or graphical form depending on which is most conducive to accurate forecasting, given the nature for the forecasting task [8]. However, one strategy that has been under explored in the literature is that of providing feedback to forecasters to help them to learn about the task and hence improve their forecast accuracy [5].

This paper reports on an experiment that was designed to explore the effectiveness of providing rolling feedback to forecasters on the outcomes of the variable they are attempting to predict and on their forecasting performance. The objective is to provide a direct training scheme, enabling forecasters to better understand the underlying pattern of the data by learning directly from their forecast errors. Two types of performance feedback were compared: feedback on the bias associated with the submitted forecasts and feedback on their accuracy. The paper is structured as follows. A review of the relevant literature is followed by an outline of the research questions that were investigated. Details of the experiment and presentation of the analysis and results follow this. Finally, the practical implications of the findings are discussed and suggestions made for further work in this area.

2 Literature review & research questions

Goodwin and Wright [6,9] argued that three components of a time series influence the degree of difficulty associated with the judgemental time series forecasting task. These are: (1) the underlying

signal, comprising factors such as its seasonality, cycles and trends and autocorrelation; (2) the level of noise around the signal; and (3) the stability of the underlying signal.

Where there are trends in series studies have consistently found that judgmental forecasters tend to damp them when making extrapolations [10,11,12]. This may because they anchor too closely on the most recent observation [13]. However, damping may also be caused by forecasters bringing non-time series information, based on their knowledge or experience, to the task. For example, experience shows that the sales growth for products tends to be damped. Similarly, in the case of downward trends in sales series, people may expect trend reversals to occur as action is taken to correct the decline [12]. Complex seasonal patterns or cyclical components have also been found to lead to inaccurate judgmental forecasts [14].

Several studies have suggested that judgmental forecasters often confuse the noise in the time series with the signal [15,16,17]. For example, they often adjust statistical forecasts to take into account recent random movements in series which they perceive to be systematic changes that were undetected by the statistical forecast [18]. Conversely, when systematic changes in the signal do occur, forecasters may delay their response to this, perceiving the change to be noise [4]. Also, they may pay too much attention to the latest observation, which will contain an element of noise [13,19]. It seems reasonable to expect that noise can also impair the detection of underlying trends and seasonal patterns, though this was not the case in two studies where series were presented graphically [11,20].

Learning through feedback is one approach that can potentially mitigate these biases. Feedback has been shown to improve the accuracy of point forecasts [18,21,22,23]. However, there are a number of different types of feedback that may be particularly relevant to the time series forecasting task [24,25] and more research is needed to determine the most effective type and how it should be delivered.

The simplest form is outcome feedback, where the forecaster is told the outcome of the variable they have been forecasting when this becomes available. This allows them to make a direct comparison between each forecast and outcome which may help them to improve their forecasting accuracy over time. However, there is evidence that learning through outcome feedback can be slow [26]. One problem is that each outcome will contain an element of noise and by highlighting this outcome it may exacerbate forecasters' tendency to pay too much attention to the latest observation and to overreact to noise in the series. However, this may not be the case when outcome is provided for a set of periods ($n>1$), rather than just one period. In any case, outcome feedback is easy to provide, easily understood and it is not contaminated by older and possibly irrelevant observations (Goodwin et al, 2004). It is also probably something that forecasters would naturally expect to see so it seems reasonable to supply it even when other forms of feedback are being provided as well.

Performance feedback provides the forecaster with information on the quality of their forecasts, such as their accuracy or any bias. Usually it will take the form of an average which reflects performance over a number of periods. Determining the number of periods over which to average performance poses a dilemma: too few and the feedback may be based on too small a sample of forecasts to provide reliable assessments of performance; too many and the performance measure will not adequately reflect recent improvements or deteriorations in performance. Exponentially weighted moving averages of performance may help to solve the dilemma, but they may be less transparent and understandable to the recipients of the feedback. Another option would be simply to supply a set of point errors for *n* recent periods without using any kind of average. This could potentially enable the forecaster to identify specific problematic periods that invite attention (for example seasonality peaks). Moreover, in a rolling origin scheme, this strategy provides a way to check if point errors are reducing over time.

We might expect different types of performance feedback to vary in their effectiveness. Feedback on bias can provide a direct message that one's forecasts are typically too high or too low and hence suggest how they might be improved. This is likely to be beneficial for untrended series or series with monotonic trends. However, it may lead to unwarranted confidence in one's current forecasting strategy when series have alternating patterns or seasonal patterns because biases in different periods will tend to cancel each other out if an average across the signed errors is to be used. Feedback on accuracy, in contrast, provides no such direct message and its implications may be difficult to discern. If forecasters are to learn from accuracy feedback they will need to experiment with alternative approaches, not specified by the feedback, and then establish if these have led to improved accuracy. This will require comparisons of accuracy across different periods adding to the forecaster's cognitive burden. Thus accuracy feedback seems unlikely to be conducive to rapid learning. This may explain the ineffectiveness of performance feedback in a study by Remus et al. [21] which consisted only of an accuracy measure (the mean absolute percentage error).

Other forms of feedback seem likely to be less relevant to practical judgmental time series forecasting contexts. Cognitive process feedback aims to provide forecasters with insights into their own forecasting strategy, causing them to reflect on the possible deficiencies of this strategy [27]. For example, a regression model may be used to attempt to capture their strategy so that the weights implicitly being attached to different items of available information, or cues, can be identified. In time series forecasting it will clearly take time to obtain sufficient information to estimate these weights reliably, thereby reducing the speed of learning by forecasters. Also, identifying the relevant cues to include in a model from the huge number of potential cues that are present in the time series forecasting task is problematical (e.g. typical cues might be the last observation, the mean of last n observations, the last difference between observations, the range of the last n observations and so on).

In addition, many of these cues will be serially correlated so multicollinearity is likely to reduce the precision with which weights can be estimated.

Task properties feedback relates to the provision of statistical information on the nature of the task to forecasters. In time series forecasting this might, for example, involve providing to the forecaster the current estimates of level, trend and seasonal indices obtained from the Holt-Winters method. However, this would essentially modify the task into one of accepting or adjusting statistical forecasts. This task has been widely researched elsewhere (e.g. [18,22,28,29]) and is not the topic of the current paper.

Ultimately, any form of feedback, regardless of its type, is likely to be most effective if it is easily and quickly understood [27], and salient, accurate and timely [5]. However, when forecasts are made in real time the intervals between successive forecasts can be lengthy (e.g. a year may elapse between the use of judgment to produce annual forecasts) so that any lessons learned at the time the previous forecast are likely to be forgotten by the time the next forecast is due. Moreover, it is possible that the forecasts for the next period are requested before the actual values for the current period become available. Such delays between decisions and feedback can hurt performance [30].

Another important aspect is repetition of the task. O'Connor el al. [31] note that updating judgmental forecasts leads to improved accuracy for trended series. Lurie and Swaminathan [32] suggest that more frequent feedback can lead decision makers to focus on feedback provided by round (and not across rounds), where rounds in the case of forecasting would refer to the calculation of new sets of forecasts from different origins. We expect this to be even more evident when interchanging between time series. For this reason, we suggest a direct rolling training scheme that focuses on each series separately, providing at the same time feedback across all rounds. We expect that such a scheme will motivate the forecaster to focus on each series and its patterns separately, increasing forecasting performance.

3 Experimental design

3.1 Forecasting approaches

Two judgmental forecasting approaches were evaluated in the current research. Each subject provided judgmental estimates with both approaches, using a fully symmetric experiment as we will discuss in sub-section 3.4.

Unaided Judgment: This is the simplest judgmental forecasting approach, while being quite popular. Humans are requested to provide their point forecasts all at once for all lead times (h), without

receiving any kind of guidance, other than the past data points. This approach will act as the benchmark in our study and, hereafter, is referred to as UJ.

Direct Rolling Training: We propose a direct rolling training approach. Letting N denote the number of available observations for a series and h the periods ahead to be estimated, $k>1$ blocks of h periods each are withheld ($N>kh$). At the first stage, only the first $N-kh$ periods are presented to the forecaster, while h forecasts ahead are requested. Upon submission of the participants' forecasts, the actual values of these h observations are presented, along with performance feedback in terms of percentage errors for each period (signed or not). This procedure is repeated for k times, with h data points being added in each repetition. Hence the completion of each training loop is followed by the submission of the h estimates for the future, unknown, periods. As such we, therefore, perform h-steps-ahead rolling evaluation [33], which is common practice in automatic forecast model selection [34]. However, in this case, instead of selecting the best model based on out-of-sample performance, we assume that this procedure will assist the subjects to better understand the time series patterns, thus providing more accurate forecasts. Hereafter, this approach is referred to as RT.

3.2 Time series

Most relevant studies that have focused on the impact of feedback for judgmental forecasting tasks made use of simulated series (e.g. [35,8]). Moreover, many studies did not examine seasonal series, confining their attention to stationary and trended ones [32,8]. Therefore, in the current research we focus on real time series that collectively demonstrate a variety of characteristics (stationary, only trended, only seasonal and both trended & seasonal). More specifically, 16 quarterly series were manually selected from the M3-Competition data set [36] as to have the required characteristics. These were confirmed by ACF plots, Cox-Stuart/ Friedman tests and /or by fitting an appropriate exponential smoothing model, using the Akakie Information Criterion. In addition, half of the trended and the seasonal series did not exhibit any significant pattern (trend or seasonality respectively) in the first two years, but did so later on. This selection was made in order to examine subjects' adaption and ability to recognise developing series characteristics.

The 16 series were grouped in two categories, each containing 8 series. These sets of series allowed for the implementation of a symmetric experimental design, which will be described in sub-section 3.4. Each set contained exactly two series with the same characteristics, as displayed in Table 1. For analysis purposes, the 16 series were further split into two sets of equal size in terms of noise (low and high), as measured by the standardised random component of classical decomposition. Lastly, 4 additional series were used at a first (warming-up) stage of the experiment, in order to familiarise the participants with the system.

Table 1. Sets of series

	Stationary	Trended	Seasonal	Trended & Seasonal	Total
Set A	2 series	2 series	2 series	2 series	8 series
Set B	2 series	2 series	2 series	2 series	8 series

The required length of all series was set to N=28 points (7 years), with longer series being truncated. In both UJ and RT approaches, the last 4 observations (last year) were withheld and used only for the out-of-sample evaluation and comparison of the two approaches. The length of this sample matches the required forecasting horizon, thus h=4. In addition, 12 more observations were used for the RT procedure, thus k=3. The forecasting performance was tested on the last 4 observations (last year), where forecasts for both approaches (UJ and RT) were produced.

3.3 Participants & incentives

Two groups of subjects participated in this experiment. The first group consisted of 105 undergraduate students enrolled in the *Forecasting Techniques* module of the School of Electrical & Computer Engineering at the National Technical University of Athens. The experiment was introduced to these students as an elective exercise, giving bonus credit for the 50% of the participants who produce the most accurate forecasts. The second group consisted of 60 participants, a mix of undergraduate (11) and postgraduate students (12), researchers (15), practitioners (6) and others (16). The second phase of the experiment was advertised through the LinkedIn network and personal communication, giving as an incentive five £20 vouchers. In both cases, only the participants with the best forecasting performance received rewards (bonus credit and vouchers respectively).

In order to attract a large number of participants, we decided not to perform a laboratory experiment, but to build a custom web-based system. Subjects could connect remotely through their personal computers via any web browser.

3.4 Process of the experiment

Instead of splitting the participants into two groups, control and test, we adopt a symmetric experimental design, where each participant submitted forecasts for both UJ and RT. The sets of series A and B alternated randomly between UJ and RT, so that half of the participants forecast some series with UJ and the other half forecast the same series with RT and vice-versa. In order to avoid familiarity with the task, UJ and RT were interchangeably presented to the subjects. When feedback was provided, each participant was randomly assigned to either the signed or unsigned percentage errors treatment (so that either accuracy or bias feedback was provided).

All series were presented in a line graph format, using the color blue for the actual values and green for the submitted forecasts. Historical data points were kept unlabeled in terms of the exact values, so that the subjects could not export these values into a spreadsheet and use statistical approaches. However, grid lines were provided in order to accommodate numerical estimations. Four text boxes were used for the input of judgmental forecasts for each lead time, while an *update* button could be used to refresh the graph, so that the subject could check his or her judgmental estimates graphically before submitting. Figure 1 presents two typical screens of the implemented system, before (a) and after (b) the input of the four point forecasts.

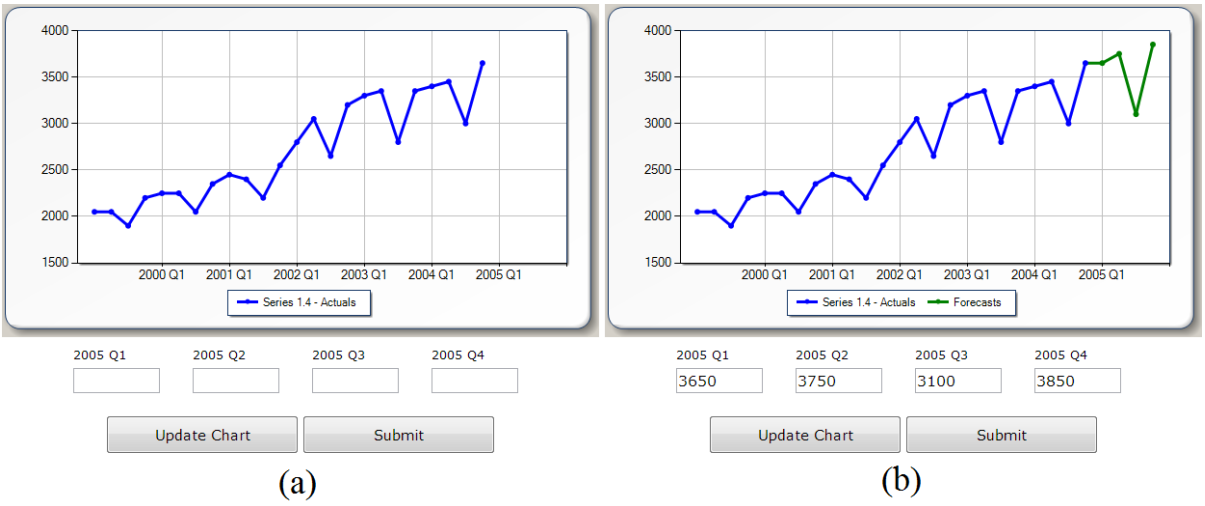

Figure 1. Screenshots of the system's graphical representation and input features.

Including the warming-up up round, the experiment was completed after three rounds. Each round is described in detail below. As noted, rounds 2 and 3 were presented in a reverse order for half of the participants.

Round 1 – Warming-up: Each of the first four series was presented to the participants, withholding the last four observations. The participants were requested to provide judgmental point forecasts for the next four quarters (one year). A short description of each series was provided, describing any historical patterns. Upon submission of the forecasts for each series, forecast errors for each point (signed or not) were automatically calculated and displayed in bar charts, using the color red. As this round was a 'warm-up' the forecasts elicited were not taken into account when the results of the study were analysed. Figure 2 presents the screen with the information provided to the participants after the submission of the four point forecasts for a series.

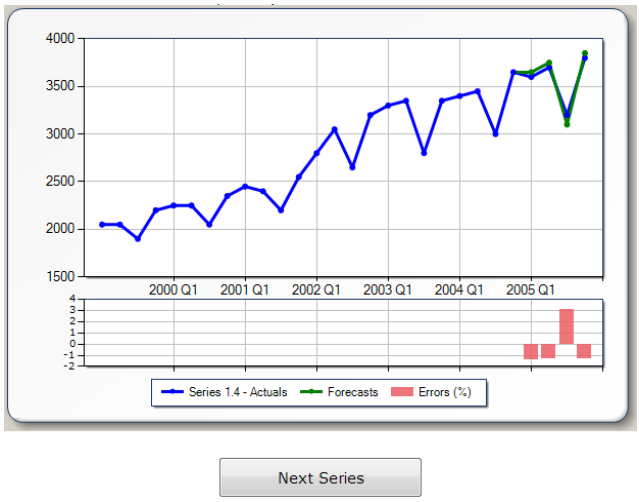

Figure 2. Screenshot of the system's feedback report features in terms of outcome (out-of-sample actual values) and performance (error bars).

Round 2 – UJ: The series from Set A (or Set B) were used, holding out the last four observations in each series. The series were presented in a random order. The participants were requested to provide judgmental point forecasts for the next four quarters. No description of the series and no information on the accuracy of the forecasts were provided.

Round 3 – RT: Series from Set B (or Set A, the opposite from Round 2) were used, holding out the last 16 observations of each series. Series were again presented in random order. Each participant was requested to provide judgmental point forecasts for the next 16 quarters (4 years) in a rolling origin manner. At first, he or she was asked to submit just the first four point forecasts (next year). Upon submission, the actual data points were appeared and forecasts errors (signed or not, the same with Round 1) were given in bar charts. Next, the second set of forecasts was requested, followed by outcome and performance feedback. Then, the third set of forecasts was requested, again followed by outcome and performance feedback. Finally, the participants submitted their final four forecasts. In order to be directly comparable with the UJ, only the last set of forecasts were used for the analyses of the current study.

The completion of rounds 2 and 3 was followed by a questionnaire which included questions on the participants' confidence in the accuracy of their submitted forecasts, their expected forecasting performance, the extent to which they had examined the graphs and series patterns, and the time they spent in producing their forecasts. In addition, a final questionnaire was used to ask participants about their familiarity with forecasting tasks, their level of forecasting expertise, their perceptions of the effectiveness of RT and their motivation to provide accurate estimates. The two sets of questions posed are displayed in Table 2. All questions were accompanied by 5-step ranked multiple selection answers.

Table 2. Questions posed to the participants

	Questions
After Round 2 and Round 3	How confident are you that the forecasts you submitted in this round on average, be within 10% of the actual values? Please, rate your expected forecasting performance in the series of this round. Did you examine carefully the time series graphs? Did you take into account any historic patterns in the series when making your forecasts during this round? How much time (on average) did you spend for each series of this round? How likely is it that taking more time would change your forecasts?
After completion of the experiment	How familiar are you with such forecasting exercises? How would you describe your level of expertise? Please, rate the effectiveness of rolling training as a tool to increase your accuracy. Please, indicate how motivated you were to provide accurate estimates.

Table 3 presents how the number of participants are distributed across the different cases of the experiment, in terms of set of series (A or B) assign to an approach (UJ or RT) and the type of the feedback provided.

Table 3. Distribution of participants across the different treatments

Approach	Set of Series	
	A	B
UJ	81	84
RT (Total)	84	81
RT (Bias)	43	38
RT (Accuracy)	41	43

4 Analysis

4.1 Forecasting performance

Table 4 presents the percentage improvements in accuracy that were achieved by using RT when compared with UJ. These percentage improvements are measured as:

$$100\left(1-\frac{MAPE_{\text{RollingTraining}}}{MAPE_{\text{UnaidedJudgment}}}\right)\ (\%)$$

where the UJ in the denominator is acting as the benchmark for this study. Negative values denote that RT performed worse than UJ. In both the numerator and the denominator, mean absolute percentage error is calculated across the participants, series and horizons, as:

$$MAPE = \frac{100}{S \cdot H} \sum_{s=1}^{S} \sum_{h=1}^{H} \frac{1}{P_S} \sum_{p=1}^{P_S} \left| \frac{y_{s,h} - f_{p,s,h}}{y_{s,h}} \right| \ (\%)$$

where P_S denotes the number of participants, S the number of series, H the number of out-of-sample lead times, $y_{s,h}$ the actual value of series s at time h and $f_{p,s,h}$ the forecast of participant p for series s at time h. Note that the number of participants (P_S) is not the same for all series, as a results of the slightly unequal sample sizes (Table 3). Results are analysed by columns in terms of series characteristics (stationary, trended, seasonal, trended & seasonal, low noise and high noise). Major rows indicate all (1 to 4), short (1 to 2) or long horizons (3 to 4), as well as results analysed in terms of the two groups of subjects considered. Minor rows provide additional analysis of the results based on the type of feedback (in the case of RT) provided to the subjects. As mentioned in section 3.1, two types of feedback have been considered: bias feedback in the form of signed percentage errors (PE) and accuracy feedback in the form of absolute percentage errors (APE).

Table 4. Accuracy improvements (%) of RT approach over UJ

	Type of Feedback	All Series	Stationary	Trended	Seasonal	Trended & Seasonal	Low Noise	High Noise
All	**ALL**	1.99	4.50[2]	3.25	-3.86	5.51	0.03	2.48
	PE	5.90[1]	5.70	4.01	7.18	5.29	0.99	7.13[1]
	APE	-1.48	3.23	2.62	-13.46[1]	5.86	-0.71	-1.68
Short Horizons (1-2)	**ALL**	0.59	4.32	0.35	-3.12	0.47	-2.69	1.37
	PE	4.39	6.40	-0.25	6.24	-2.13	-4.10	6.42
	APE	-2.79	2.18	0.99	-11.40[2]	3.31	-0.99	-3.23
Long Horizons (3-4)	**ALL**	2.99	4.62[1]	5.64	-4.61	8.03	1.86	3.29
	PE	6.98[1]	5.27[2]	7.51	8.13	8.99	4.40	7.65[1]
	APE	-0.54	3.88	3.95	-15.54	7.14	-0.52	-0.55
Group 1 (NTUA Students) Sample: 105	**ALL**	4.20[1]	5.97[2]	4.17	0.77	5.91	2.39	4.65
	PE	5.20[1]	5.78	4.51	5.81	3.14	2.07	5.98[1]
	APE	3.23	6.15[2]	3.85	-4.12	8.62	2.74	3.35
Group 2 (General) Sample: 60	**ALL**	-1.42	2.93	1.89	-11.80	4.87	-3.24	-0.96
	PE	7.93[2]	6.46	3.12	10.72	9.38	0.00	9.94[2]
	APE	-8.56	-1.06	0.85	-27.42[2]	1.71	-5.72[2]	-9.28

[1]Differences are statistically significant at 0.05 level.

[2]Differences are statistically significant at 0.10 level.

Overall, there is evidence that the RT approach results in greater forecasting performance, as measured by MAPE. Improvements are more prominent for series exhibiting stationarity (4.50%), trend (3.25%), trend with seasonality (5.51%) and high noise (2.48%). Moreover, the difference of the

average performance between UJ and RT has, overall, the same direction for these categories across short and long horizons.

Focusing on the very first row of the results, where all subjects and horizons are considered, the only negative value comes from the seasonal series. In an attempt to understand the reason behind this result, a further split was performed, examining separately series with evident seasonality for the very first years or not, as discussed in section 3.2. For series with constant seasonal behaviour UJ and RT have very similar performance. However, RT is not suitable for series with developing seasonality, as in this case UJ is on average better (by 7%). This outcome was not confirmed for the trended series. In fact, when trended behaviour is developing, the RT approach works even better (improvements of 4.84% in contrast to 1.13% for series with almost constant trend).

In terms of the type of feedback provided to the subjects, it is apparent that bias feedback demonstrates significant improvements for most of the cases (5.90% overall, statistically significant differences in accuracy at 0.5 level), while this is not true for accuracy feedback (-1.48% overall). One could argue that providing errors in an absolute format may lead to confusion, as the participants may not be able to correctly evaluate this kind of information. On the other hand, bias feedback for each point in the form of signed bar charts is easier to interpret and understand and indicates a clear strategy for improving one's forecasts. It is notable that bias feedback, which involved the provision of signed percentage errors for each individual period, improved accuracy for seasonal series. It is unlikely that providing a mean of these percentage errors would have been as effective because any tendency to over forecast for some seasons and under forecast for others would have been masked by the averaging process.

Another very important observation is that improvements for RT are greater for series with high noise (in contrast to low noise ones) as well as when longer horizons are examined (in contrast to shorter horizons). This is precisely where unaided judgmental forecasters are most likely to have difficulty with the task so RT may have an important role to play here. This suggests that RT is suitable for forecasting and decision making under low levels of predictability (ie. where there is a high degree of uncertainty). Lawrence et al. [37] suggested that, when the forecasting task is based on graphs, judgmental forecasts can be as good as statistical models at least for the shorter horizons. The use of a direct rolling training scheme improves graph-based judgmental long term forecasting, building on the efficiency of judgmental over statistical approaches. Although the two groups (NTUA students versus general) differed in the accuracy improvement they obtained when APEs were provided to them, both groups generally benefited from PE feedback.

The exact same analysis was performed as to examine any differences in the two approaches (UJ and RT) in terms of out-of-sample forecasting bias (measured across horizons). The results (available

upon request) did not offer statistically significant additional insights and, therefore, for brevity are not presented in this paper.

4.2 Questionnaire responses analysis

Figure 3 presents in spider graphs the mean values of the responses for each question separately, analysed in UJ and RT for the first set of questions. Figure 4 presents graphically the relationships between the participants' responses to the first set of questions (x-axis) with their mean performance (y-axis), as measured by MAPE Separate lines are presented for UJ (blue) and RT (green). The size of the circle on each data point reflects the number of participants who provided the respective response. As this first set of questions was posed twice (after UJ and RT respectively), we can examine how the subjects alternate their responses after each forecasting approach. In the same manner, Figure 5 presents the relationships for the second set of questions. Separate lines are again presented for UJ and RT.

The moderate negative association of confidence level with MAPE in UJ changes to no correlation for RT. Moreover, as seen from Figure 3, subjects tend to be less confident for their submitted forecasts when using RT over UJ. This outcome is very important, as it is obvious that RT leads the subjects to be more cautious in their expectations, thus potentially mitigating a well know problem of judgemental forecasting , namely the underestimation of uncertainty (e.g. [38]). However, in both cases (UJ and RT) participants are able to estimate their expected forecasting performance well, as there are strong associations between their observed and expected accuracy.

As expected, both examination of graphs and patterns have strong negative associations with the MAPE, suggesting that, as subjects devote more time to these tasks, improvements in forecasting accuracy are recorded. According to Figure 3, no differences are observed between the two approaches (UJ and RT). One would have expected that RT would encourage the participants to examine the graphs and series patterns more carefully; however this was not the case.

The forecasting performance achieved with both UJ and RT is associated with the time the participants reported spending to produce the forecasts for each series – the more time they spent the greater the accuracy they achieved. Also, there is evidence that participants who were less accurate recognised that spending more time on the task might has improved their performance (this is particularly the case for the RT group), as a result of the strong observed association between forecasting accuracy and possibility to perform changes in the submitted forecasts, given more time.

Familiarity with forecasting exercises, self-reported level of expertise and motivation were only weakly associated with forecasting accuracy (Figure 5). However, the majority of the subjects (73%)

found the RT approach to be either effective or very effective, and its perceived effectiveness has a moderate association with forecasting performance.

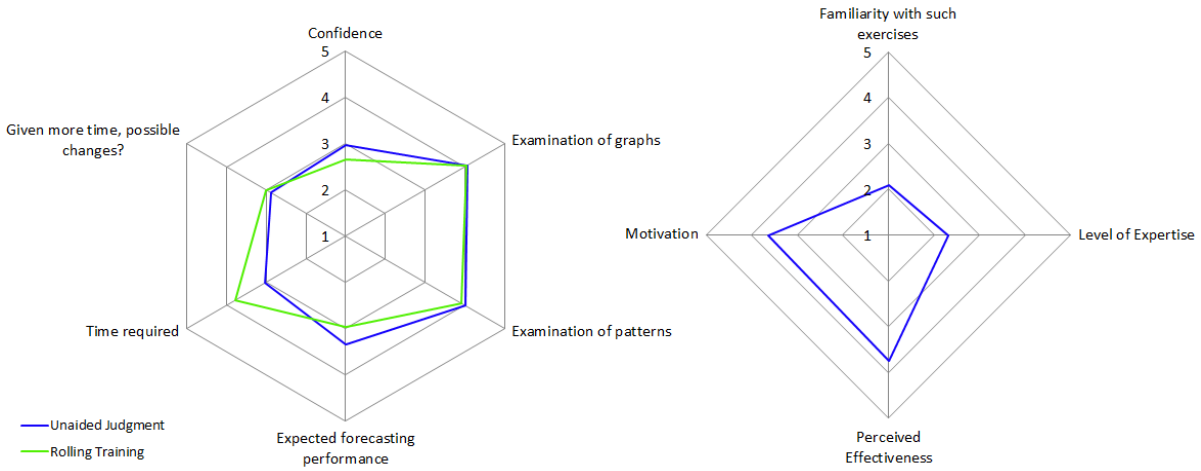

Figure 3. Average values of the responses for each question.

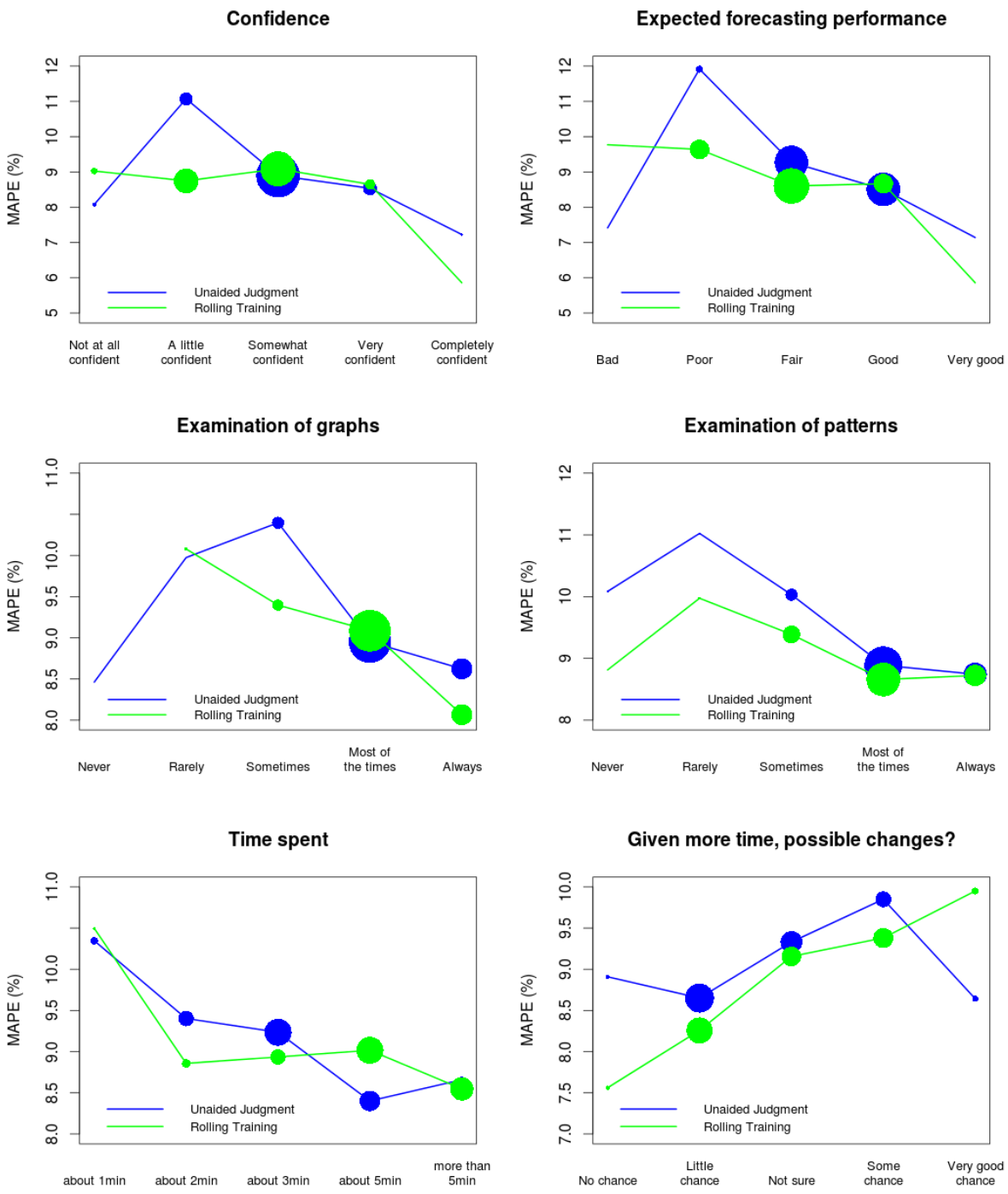

Figure 4. Association between questionnaire responses and forecasting performance for the first set of questions

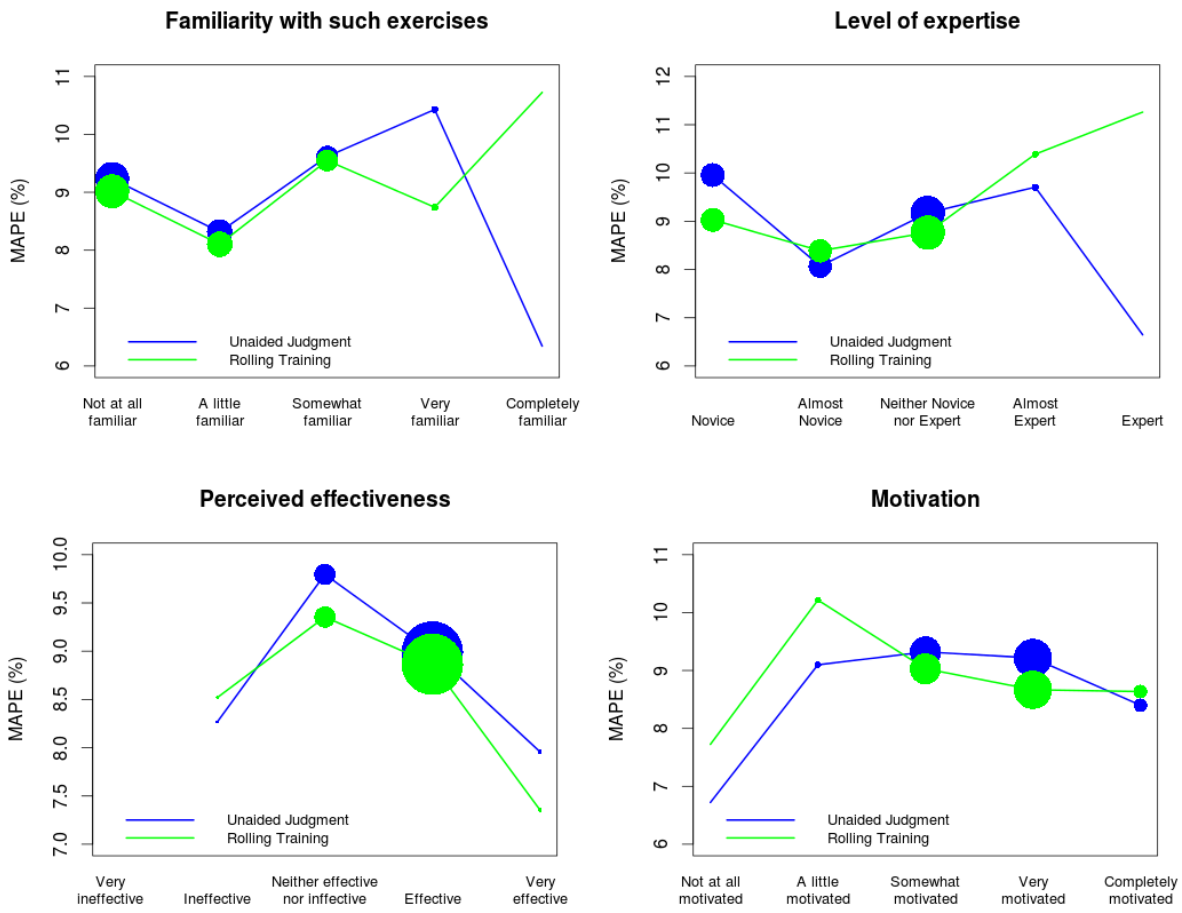

Figure 5. Associations between questionnaire responses and forecasting performance for the second set of questions

5 Discussion & implications

The key finding of this study is that, in tasks involving time series extrapolation where no contextual information is available, judgmental forecasting accuracy can be substantially improved by providing simple, understandable performance feedback to the forecasters. A number of characteristics of this feedback appear to be crucial. First, to be most effective, the feedback should relate to bias, rather than accuracy. As discussed earlier, feedback on bias provides a clear indication of how future forecasts might be improved. In contrast, feedback on accuracy does not provide any indication of possible improvement strategies. Nor does it provide an indication of whether accuracy improvement is even possible. For example, does an APE of 10% represent the limit of the accuracy that can be achieved, given the noise level, or is their scope for further improvement?

Second, the attribute of the bias feedback that appeared to contribute to its effectiveness was the feedback of a set of individual errors, rather than an average of these errors. In series where the signal

has cyclical elements such as seasonal series, judgmental biases may lead to positive errors at some stages of the cycle (e.g. where sales are increasing) and negative errors at other stages (e.g. where sales are decreasing). The presentation of individual errors allows each observed bias to be associated with individual periods and avoids the cancelling out of opposing biases that would be a feature of any averaging. Also, the need for appropriately selecting a length for averaging the point forecast errors is now removed.

Third, the presentation of the bias feedback as bar charts may have enhanced its effectiveness, though further research would be needed to establish this. For example a set of four negative bars would be a strong, simple and clear indication that the previous set of forecasts was too high (error = actual – forecast). A table of four numbers would probably provide a less salient message.

Fourth, the rolling nature of the feedback enabled it to reflect improvements in performance quickly, while at the same time avoiding the danger of confining attention to the performance of the most recent forecast (which is a danger of outcome feedback). Moreover, rolling across origins for one series, before moving on to the second one, helped the participants to focus on each series separately and better understanding the improvements (or deteriorations) in their performance over time.

Recent research suggests that the focus on enabling people to learn how to avoid bias is appropriate. A study by Sanders and Graman [39] found that when translating forecast errors into costs (such as excessive inventory or labor costs) accuracy was less important than bias and costs rose exponentially as the bias in the forecasts increased. In their survey of forecasters Fildes and Goodwin [1] expressed surprise at the number of company forecasters who never checked the accuracy of their forecasts. The current study and the findings of Sanders and Graman [39] suggest that monitoring and feeding back levels of bias may be just as, if not more, important than checking accuracy levels if the objective is to foster improved forecasts and minimize the costs of errors.

6 Conclusions & perspectives

Judgmental forecasting is widely employed in many contexts for estimating future values of time series. The current study examined the efficiency of a rolling training scheme that provides direct feedback by reporting to participants their levels of performance in such a task. To implement performance feedback, signed or absolute point errors for each period were reported on a rolling basis. Real time series featuring a number of characteristics were used, while a large number of participants were employed for the experiment. In order to maximise exploitation, participants provided estimates for both the control case (unaided judgment) and the test case (rolling training). This was achieved by a symmetric experimental design.

Analysis of the judgmental estimates indicates that a rolling training scheme can improve the performance of judgmental extrapolation, especially when combined with feedback in form of signed errors. Effectively, providing feedback with regards to the bias in the forecasts, the forecasting ability of participants is enhanced. This is particularly obvious in non-stationary series. On the other hand, an absolute form of errors is found to be more difficult to interpret, leading to worse performance in the case of series exhibiting seasonality.

One very interesting outcome is that improvements achieved by using a rolling training procedure are higher for longer forecasting horizons and noisy series. However, despite the improvements achieved through the rolling training procedure, it made the participants less confident in their forecasts. This may be an advantage as there is evidence that people tend to under estimate the levels of uncertainty associated with their forecasts.

The current paper focused on analysing the performance over the final set of periods (final year) and contrasting unaided judgment with rolling training. However, a further objective would be to analyse how the forecasting performance changes over time within a single series, as a direct result of the application of the rolling training procedure. Moreover, policy capturing regression models may be developed to provide insights of the forecasting strategy employed by the participants. This can include a large number of potential cues linked with time series forecasting. Of course, often the time series forecasting task is carried out in situations where contextual information (such information from market research or information on advertising strategies) is available to the forecaster, in addition to time series data and it would be interesting to test the effectiveness of rolling training in this context.

References

[1] R. Fildes, P. Goodwin, Against your better judgment? How organizations can improve their use of management judgment in forecasting, Interfaces 37 (2007) 570-576.

[2] R. Fildes, P. Goodwin, M. Lawrence, K. Nikolopoulos, Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning, Int J. Forecast. 25 (2009) 3-23.

[3] P. Goodwin, Integrating management judgment and statistical methods to improve short-term forecasts, Omega: Int. J. Manag. Sci. 30 (2002) 127-135.

[4] M. O'Connor, W. Remus, K. Griggs, Judgemental forecasting in times of change, Int. J. Forecast. 9 (1993) 163-172.

[5] M. Lawrence, P. Goodwin, M. O'Connor, D. Onkal, Judgmental forecasting: A review of progress over the last 25 years, Int. J. Forecast. 22 (2006) 493-518.

[6] P. Goodwin, G. Wright, Improving judgmental time series forecasting: A review of the guidance provided by research, Int. J. Forecast. 9 (1993) 147-161.

[7] R.H. Edmundson, Decomposition; a strategy for judgmental forecasting, J. Forecast. 9 (1990) 305-314.

[8] F. Bolger, D. Önkal Atay, The effects of feedback on judgmental interval predictions, Int. J. Forecast. 20 (2004) 29-39.

[9] P. Goodwin, G. Wright, Heuristics, biases and improvement strategies in judgmental time series forecasting, Omega: Int. J. Manag. Sci. 22 (1994) 553-568.

[10] I.R.C. Eggleton, Intuitive time-series extrapolation, J. Account. Res. 20 (1982) 68-102.

[11] M.J. Lawrence, S. Makridakis, Factors affecting judgmental forecasts and confidence intervals, Organ. Behav. Hum. Decis. Process. 42 (1989) 172-187.

[12] M. O'Connor, W. Remus, K. Griggs, Going up-going down: How good are people at forecasting trends and changes in trends? J. Forecast. 16 (1997) 165-176.

[13] F. Bolger, N. Harvey, Context-sensitive heuristics in statistical reasoning. Q. J. Exp. Psychol. 46A (1993) 779-811.

[14] M. Lawrence, M. O'Connor, Scale, randomness and the calibration of judgemental confidence intervals, Organ. Behav. Hum. Decis. Process. 56 (1993) 441-458.

[15] P.B. Andreassen, Explaining the price volume relationship - the difference between price changes and changing prices, Organ. Behav. Hum. Decis. Process. 41 (1988) 371-389.

[16] N. Harvey, Why are judgments less consistent in less predictable task situations? Organ. Behav. Hum. Decis. Process. 63 (1995) 247-263.

[17] L.L. Lopes, G.C. Oden, Distinguishing between random and nonrandom events. Journal of Experimental Psychology-Learning Memory and Cognition, 13 (1987) 392-400.

[18] P. Goodwin, R. Fildes, Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? J. Behav. Decis. Mak. 12 (1999) 37-53.

[19] M.J. Lawrence, M.J. O'Connor, Exploring judgemental forecasting. Int. J. Forecast. 8 (1992) 15-26.

[20] F. Mosteller, A.F. Siegel, E. Trapido, C. Youtz, Eye fitting straight lines, The Am. Stat. 35 (1981) 150-152.

[21] W. Remus, M. O'Connor, K. Griggs, Does feedback improve the accuracy of recurrent judgemental forecasts? Organ. Behav. Hum. Decis. Process. 66 (1996) 22-30.

[22] N.R. Sanders, The impact of task properties feedback on time series judgmental forecasting tasks, Omega: Int. J. Manag. Sci. 25 (1997) 135-144.

[23] E. Welch, S. Bretschneider, J. Rohrbaugh, Accuracy of judgmental extrapolation of time series data - Characteristics, causes, and remediation strategies for forecasting, Int. J. Forecast. 14 (1998) 95-110.

[24] W.K. Balzer, M.E. Doherty, R. O'Connor, Effects of cognitive feedback on performance. Psychol. Bull. 106 (1989) 410-433.

[25] D. Önkal, G. Muradoglu, Effects of feedback on probabilistic forecasts of stock prices, Int. J. Forecast. 11 (1995) 307-319.

[26] J. Klayman, Learning from experience. In: B. Brehmer, C.R.B. Joyce (Eds.), Human Judgment. The SJT View, North Holland, Amsterdam, 1988, pp. 281-304.

[27] M. O'Connor, W. Remus, K. Lim, Improving judgmental forecasts with judgmental bootstrapping and task feedback support, J. Behav. Dec. Mak. 18 (2005) 247-260.

[28] T.R. Willemain, Graphical adjustment of statistical forecasts, Int. J. Forecast. 5 (1989) 179-185.

[29] T.R. Willemain, The effect of graphical adjustment on forecast accuracy, Int. J, Forecast. 7 (1991) 151-154.

[30] J.D. Sterman, Modeling managerial behaviour: misperceptions of feedback in a dynamic decision making experiment, Manag. Sci. 35 (1989) 321-339.

[31] M. O'Connor, W. Remus, K. Griggs, Does updating judgmental forecasts improve forecast accuracy?, Int. J. Forecast. 16 (2000) 101-109.

[32] N.H. Lurie, J.M. Swaminathan, Is timely information always better? The effect of feedback frequency on decision making, Organ. Behav. Hum. Decis. Process. 108 (2009) 315-329.

[33] L.J. Tashman, Out-of-sample tests of forecasting accuracy: an analysis and review, Int. J. Forecast. 16 (2000) 437-450.

[34] R. Fildes, F. Petropoulos, An evaluation of simple versus complex selection rules for forecasting many time series, J. Bus. Res. forthcoming.

[35] I. Fischer, N. Harvey, Combining forecasts: What information do judges need to outperform the simple average?, Int. J. Forecast. 15 (1999) 227-246.

[36] S. Makridakis, M. Hibon, The M3-Competition: results, conclusions and implications, Int. J. Forecast. 16 (2000) 451-476.

[37] M.J. Lawrence, R.H. Edmundson, M.J. O'Connor, An examination of the accuracy of judgmental extrapolation of time series, Int. J. Forecast. 1 (1985) 25-35.

[38] S. Makridakis, R.M. Hogarth, A. Gaba, Forecasting and uncertainty in the economic and business world, Int. J. Forecast. 25 (2009) 794-812.

[39] N.R. Sanders, G.A. Graman, Quantifying costs of forecast errors: A case study of the warehouse environment, Omega: Int. J. Manag. Sci. 37 (2009) 116-125.