Modeling judgmental forecasts under tabular and graphical data presentation formats

Otwin Becker, University of Heidelberg
Johannes Leitner, University of Graz
Ulrike Leopold-Wildburger, University of Graz*

Abstract: Subjects make judgmental forecasts of a time series in two versions of a laboratory experiment: One group is given the time series in a chart, the other group in a table. In both sessions, the past realisations of the time series are the only available sources of information. A heuristic for the explanation of the subjects' average forecasts is presented. The rationale of this simple scheme-oriented forecasting heuristic is based on gestalts characteristics of the time series. We analyse the format-related differences of the forecasts and demonstrate that the model explains the average forecasts very well independently of the presentation format. The average forecasts are also explained by the rational expectations hypothesis in order to compare its the performance to the heuristic.

JEL-classification: C91, C92, C53.
Keywords: Judgemental Forecasting, Time Series, Heuristics, Rational Expectations, Experimental Economics.

*Corresponding author: Ulrike Leopold-Wildburger, Department of Statistics and Operations Research, Karl-Franzens-University of Graz, Universitätsstraße 15/E3, 8010 Graz, Austria. Tel: +43(316)380-3492, Fax: +43(316)380-9560. email: ulrike.leopold@uni-graz.at

Otwin Becker
Prof. em. Dr. Otwin Becker
University of Heidelberg
Tannenweg 21a
D-69190 Walldorf
Tel. ++49 (6227) 61287
Fax. ++49 (6227) 890810
email: o.becker.walldorf@t-online.de

Johannes Leitner
Department of Statistics and Operations Research
Universität Graz
Universitätsstraße 15/E3
8010 Graz (Austria)
Tel: +43(316)380-7245
Fax: +43(316)380-9560
email: johannesleitner@gmx.de

Ulrike Leopold-Wildburger (Corresponding author)
Department of Statistics and Operations Research
Universität Graz
Universitätsstraße 15/E3
8010 Graz (Austria)
Tel: +43(316)380-3492
Fax: +43(316)380-9560
Email: ulrike.leopold@uni-graz.at

Acknowledgements

This work was supported by the project P 17156-N12 of the FWF (Austrian Science Foundation).

Modeling judgmental forecasts under tabular and graphical data presentation formats

1 Introduction

The accuracy of statistical time series forecasts is a critical factor for the situation-specific application of a model. Makridakis and Hibon (1979) were the first to empirically explore the performance of various statistical models in forecasting competitions on a dataset of thousands of real time series. It was found inter alia that simple procedures, such as exponential smooting, perform equivalently to sophisticated models (Makridakis and Hibon (2000)) - a result supported by many other authors. Although statistical models were the initial interest of forecasting competitions probably the most common forecasting approach was incorporated soon: judgmental forecasting. Judgmental forecasts are based on subjective eyeballing of the past realisations of the time series without the support of statistical procedures - a techique which seems to be inferior to statistical procedures at first glance. Lawrence et al. (1985) applied 111 real-life time series of the Makridakis forecasting competition (Makridakis et al. (1982)) in a forecasting experiment and compared the accuracy of judgmental forecasts to statistical models. Judgmental forecasts were at least as accurate as statistical models, and in some cases even superior to them.[1] The authors also identified the influence of data presentation formats on the accuracy: Forecasts of time series presented in tables significantly outperformed graphs for annual time series (long run). They also found table forecasts to be more robust, i.e. smaller standard deviations of the forecasting errors. The authors attribute the differences to the inability of tabular forecasters to recognize short-term trends for the most recent realisations. In a direct experimental comparison of data presentation formats on forecasting accuracy, Harvey and Bolger (1996) tested the forecasts of trended and untrended time series with different noise levels. They find a slight advantage of untrended time series in tabular format, but a clear superiority for the graphical format in all other cases.

Unlike the mentioned studies, the main focus of the present paper is not forecasting accuracy but the modeling of judgmental forecasts of a tabularly and graphically presented time series. In prior experimental setups for the analysis of expectation formation mechanisms the effects of data presentation formats have been widely ignored. Schmalensee (1976) tested the forecasts of subjects on compatibility with the adaptive and the extrapolative models with a chart of a time series. Dwyer et al. (1993) demonstrated that subjects rationally forecast a graphically presented random walk. For instance, Brennscheidt (1993), Hey (1994), Beckman and Downs (1997) tested various models on judgmental forecasts of time series presented in both formats simultaneously. Hey allowed his subjects to switch between formats according to their own convenience. Hence, potential format effects were completely lost

[1]However, not all authors conclude superiority of judgmental over statistical methods. See Webby and O'Connor (1996) for an extensive review.

in the results.
In our experiment, tabular and graphical forecasts of a time series are collected from student subjects. We apply a simple scheme-oriented explanation model, the bounds&likelihood heuristic, for the explanation of the subjects' average forecasts. The heuristic was successfully tested on a sample of about 600 subjects in various experimental versions by Becker et al. (2004a, 2004b) and Becker and Leopold-Wildburger (2000), whereas all these experiments were exclusively based on charts. The primary question is now the extent to which tabular and graphical supported forecasts differ and whether they can be explained - on average - by the heuristic. It is hypothesized that that tabular and graphical supported forecasts are both based on forecasting schemes. It is our motivation to verify whether the rationale of the heurisitic explains the forecasts, i.e. that the average forecasts of both groups follow the same scheme. The performance of the bounds&likelihood heuristic will be compared to the rational expectations hypothesis REH.

2 The experiment

Academic subjects made judgmental forecasts of a time series x_t over 42 periods. The subjects were not provided with any additional information, help from statistical models or any contextual information. The time series was unlabeled. The only utilisable information were the past realisations of x_t. The time series is a realization of the stochastic difference equation

$$x_t = x_{t-1} - INT(\frac{1}{2} \cdot x_{t-2}) + u_t \tag{1}$$

with the endogenous variable x_t and the white noise u_t. The variable u_t is uniformly distributed in the interval [1,6]. All values of x_t and the subjects' forecasts are integer. The forecasts were limited to the interval [0,30]. The start value $x_1 = 7$ was given to the subjects in the first period. No history of realisations was presented to the subjects in the first period. Based on this, subjects made their forecast f_2 and were then informed about the true realisation of x_2. Hence, the information set of the subjects for the forecast of period $t+1$ only consisted of all past values ($\Omega_t = \{x_t, x_{t-1}, \ldots x_1\}$). The experiment was carried out in two versions. The introductions and information given to the subjects, the payment function and the experimental procedure were the same in both versions. The main difference was the presentation format of the time series: In the graphical versions, the values of x_t and f_t were presented in a chart, in the tabular version in a table. Figure 1 and Figure 2 show the time series x_t as it was presented to the subjects in both versions.

Insert Figure 1 about here.

Insert Figure 2 about here.

The tabular experiment was carried out with paper and pencil, the graphical version with computers. In the tabular version, subjects were handed out a table with 42 columns and three rows.[2] The periods were numbered in the first row. The realisations of the time series were inserted in the second row, the subject's own forecasts in the third row. On the handout, the first column of the second row had the value 7. All other fields of the second and third row were empty. When all subjects had made their forecasts, they were informed about the true value verbally and noted it in the table. Then the next forecast was made, and this was repeated for all 42 periods.
The experiments were conducted at the Department of Statistics and Operations Research, University of Graz. Altogether 102 undergraduate subjects participated voluntarily, 72 in the graphical and 30 in the tabular version. The subjects were recruited from undergraduate courses of business administration. They were given a significant financial incentive to forecast the time series accurately. They were paid 60 Cents for an exact forecast, a forecast error of one (two) unit(s) was rewarded with 40 (20) Cents. This simple payment scheme corresponds to a function of absolute forecast errors that is cut off to zero at the value of three. The average payments in the graphical (tabular) version of the experiment were 9.2 (8) Euros at an average duration of about 30 minutes.

3 Two explanation models

3.1 The bounds&likelihood heuristic

The bounds&likelihood heuristic (b&l heuristic) by Becker and Leopold-Wildburger (1996, 2000) models average forecasts. It is assumed that two features of the time series are essential for the forecasts: the average variation and the turning points. The average absolute variations of the time series $b_t = \frac{1}{t-1}\sum_{j=2}^{t}|x_j - x_{j-1}|$ are the bounds for the predicted change based on the actual time series value x_t. The maximum predicted change is supposed to be in the interval [-b_t, b_t]. The actually predicted change depends on the likelihood that x_t is a turning point. For $x_t > x_{t-1}$, an upswing case, $l_{t(peak)}$ is the probability that x_t is a local maximum. The total number of local minima observed so far (N_t) and the number of local minima $\leq x_t$ (n_t) are considered. If all local maxima are below x_t, i.e. $n_t = N_t$, it is very likely to be a turning point. For a downswing case ($x_t < x_{t-1}$), the total number of local minima (M_t) and the number of local minima $\geq x_t$ are considered for the calculation of $l_{t(trough)}$. This is shown in equation (2).

[2] We know from our database of experimental results that no significant differences between computerized and paper-based settings exist. The differences reported here can accordingly be ascribed only to the data presentation format.

$$
\begin{aligned}
l_{t(peak)} &= \frac{1+n_t}{2+N_t} \\
l_{t(trough)} &= \frac{1+m_t}{2+M_t}
\end{aligned} \tag{2}
$$

In the case of no change ($x_t = x_{t-1}$), it is assumed that the upswing and downswing cases are combined linearly. At a high level of the time series, subjects will forecast a downswing; at a low level, an upswing. Based on these assumptions, the values of the heuristic $f_{t,b\&l}$ are described by equation (3).

$$
f_{t+1,b\&l} = \begin{cases} x_t + b_t(1 - 2l_{t(peak)}) & for\ x_t > x_{t-1} \\ x_t + b_t(l_{t(trough)} - l_{t(peak)}) & for\ x_t = x_{t-1} \\ x_t - b_t(1 - 2l_{t(trough)}) & for\ x_t < x_{t-1} \end{cases} \tag{3}
$$

3.2 The Rational Expectations Hypothesis

The rational expectations hypothesis (REH) suggests that agents form their expectations consistent with economic theory. They should derive their forecasts from the true economic model that generates the variable to be forecasted. The subjective distributions about future realizations should be the same as the actual distributions, conditional on the available information set (Muth (1961)). The information set of a rational forecaster contains the true model and its parameters and all the realizations of the time series observed so far. The knowledge of the true model (1) that generated the experimentally applied time series allows the calculation of the values of rational expectations. In our experiment, the REH values can simply be calculated by replacing u_t in (1) with its expected value 3.5:

$$
f_{t,REH} = x_{t-1} - INT(\frac{1}{2} \cdot x_{t-2}) + 3.5 \tag{4}
$$

With these values it can be tested whether the rational expectations hypothesis gives a valid explanation of the subjects' average forecasts.

4 Results

In this section we analyze the forecasts of the subjects. The differences between the tabular and graphical forecasts are explored on the collective and the individual level. The performance of the b&l-heuristic and the REH in explaining the subjects' average forecasts will be tested.[3]

[3] In both experimental versions, the first six periods serve as a phase for familiarization and practice. Hence, periods 1 to 6 are not taken into account within the statistical analysis and only periods 7-42 are considered.

4.1 The forecasts of the subjects

The first crucial question for the modeling of average forecasts is whether the distributions of the forecasts in both groups differ, and if so, what reasons for these differences can be assigned. In each of the 36 considered periods a Kolmogorov-Smirnov test was performed to test for differences in the distributions of the individual forecasts. In 29 periods, graphical and tabular forecasts do not differ significantly at the 99%-level of significance. The remaining seven periods (8, 11, 23, 32, 36, 40 and 41) are local extrema or periods before/after local extrema. The graphical group overestimates the time series especially in periods of local maxima. A possible explanation for the overestimation bias is the presentation of the time series in the lower half of the chart (see Figure 1). Both groups were told that the realisations of the time series are within the interval [0,30] but only the graphical group was permanently aware of this fact by the scale of the ordinate. Despite the cyclical structure of the time series, the graph group could expect the time series to reach higher values than the local maxima observed so far. The local minima are less relevant since they occur close to the abscissa of the chart. In the tabular experiment, subjects need longer to notice the low level of the time series (see Figure 5). The subjects forecast much higher values in the first periods, which explains the large deviation in period 8. There are indications for systematic differences between the forecasts of both groups in these seven periods, but the small sample does not allow the test for significance. In Figures 3 and 4, the frequencies of forecasted values are represented by circles of different sizes.

Insert Figure 3 about here.

Insert Figure 4 about here.

The differences in the individual and collective forecasts of both groups are analysed by their forecasting errors. The collective forecasts f_t^{avg} are calculated as arithmetic means of the forecasts in each period. The error measurement categories are reported in Table 1. At values of 0.808 (graphical) and 0.886 (tabular) the Theil's U of both average forecasts are below the critical value of 1. Consequently, both groups outperform the naive random walk forecasts. The MdAPE, MSE and the MAE consistently favor the graphical group, but the ME is lower for the tabular forecasts. The latter implies that the graphical group overestimates the time series. More detailed insight into the structure of the forecasting errors can be brought by a decomposition of the MSE (see Theil (1966))

$$MSE = \frac{1}{T}\sum_{t=7}^{T}(x_t - f_t)^2 = (\overline{x} - \overline{f})^2 + (SD_x - SD_f)^2 + 2(1 - r_{xf})SD_x SD_f \quad (5)$$

where $\overline{x}$ is the arithmetic mean of the time series x_t, $\overline{f}$ is the arithmetic mean of the corresponding forecasts f_t, SD_x and SD_f are their standard deviations and r_{xf} denotes the correlation between the time series and the forecasts. In Table 1, the

MSE components are reported. As expected from the mean error, the deviation from $\overline{x}$ is larger in the graphical group, but its standard deviation is lower and its correlation coeffcient is higher. The $\overline{f}$ of the tabular forecasts is closer to $\overline{x}$, which compensates their lower correlation and worse accordance with the standard deviation of x_t. Despite the reported differences, according to a Wilcoxon signed rank test of the forecast errors of both groups, the differences are not significant (z=-1.384, p=0.166). Thus, neither the distributions nor the forecasting errors of average forecasts of both groups differ significantly.
The same measurement categories are applied to the forecasts of the individuals. The average/median measures of all individuals in both groups are presented in Table 1. MSE, Theil's U and MAE are lower in the graphical group, but the ME is higher. Furthermore, the same conclusions as for the individuals can be drawn from the MSE components. The individuals in the graphical version overestimate the time series, the standard deviations of their forecasts are closer to the actual standard deviations and their forecasts show a significant higher correlation (Mann-Whitney U Test, p<.001) to the actual values.
It can be concluded that there are some significant differences between the individual forecasts which can be attributed to the correlation of the forecasts. However, the distributions of the individual forecasts and their averages do not differ significantly. These results are the basis for the modeling of average forecasts reported in the next section. Another interesting observation is that the combination of the forecasts of both groups results in much higher forecasting accuracy. The arithmetic mean of 30 tabular subjects has a Theil's U of 0.886 while the average indvidual Theil's U is at 1.305, far above the critical value of 1. Thus, by combining 30 forecasts with a simple arithmetic mean the accuracy can be improved substantially.

Insert Table 1 about here.

4.2 Modeling average forecasts

The main interest of the analysis is the extent to which average forecasts of both groups can be explained by the b&l heuristic and the REH. We estimate a simple linear regression with the average forecast as a predicted variable and the two models as predictors:

$$f_t^{avg} = \alpha + \beta f_{t,\theta} \qquad \text{with } \theta = b\&l, REH. \tag{6}$$

Both models are tested over the forecasting horizon of periods 7-42. The results are presented in Table 2. The most important result is that the heuristic explains 93.1% of the variance of the average forecast in the graphical version and only slightly less (89.9%) in the tabular version. While the heuristic performs worse than the REH in the graphical version, it outperforms rational expectations in the tabular version (89.9% vs. 86.4%). These results hold at lower autocorrelation of the residuals for both b&l estimates as indicated by the Durbin-Watson statistics. While the estimated slope coefficients in the tabular experiment are not significantly different from 1, the intercepts are significantly larger than 0. This is a drawback compared to the graphical version. In order to test potential learning processes and the time invariance of these results, a half split analysis is performed and further regressions are estimated by considering periods 7-24 and 25-42. The results reported in Table 2 support the conclusions from the analysis of the total subset. The coefficients of determination hardly vary with the exception of REH in the tabular version in which it increases from 83.3% to 92.1%. The heuristic explains the average behavior of the subjects over the whole considered time horizon. Figure 5 shows the average forecasts of the subjects f_t^{avg} and the two models in the two experimental verisons.

Insert Figure 5 about here.

These results demonstrate that the b&l-heuristic explains the average forecasts of the subjects very well in both experimental versions. The rationale of the bounds and likelihoods can be applied to tabularly presented time series, since there are no remarkable differences between the tabular and the graphical average forecasts. The heuristic explains the forecasts to the same degree as the REH. This is a remarkable fact since the REH works with strong assumptions: It assumes the knowledge of the true model, whereas the heuristic is only based on the gestalts characteristics of the time series. Furthermore this means that the efficiency of the scheme-oriented forecasting procedure is remarkably high.

Insert Table 2 about here.

5 Summary and Conclusion

In this study we reported on a forecasting experiment with the first application of the bounds&likelihood heuristic on judgmental forecasts of a tabularly presented time series. The average forecasting behavior of the subjects can be explained surprisingly well by the heuristic. A comparison with an experiment in graphical format shows hardly any performance differences. It was also shown that the heuristic performs equivalently to the REH.

This result can be attributed to the fact that no significant differences in the distributions of the two samples could be found. Why is this the case? The psychological background of the bounds&likelihood heuristic is the schema-theory. This psychological theory explains human behavior with the application of categorical rules that are used to interpret the environment. New information is processed according to how it fits into this schema. Schemes are not only used to interpret, but also to predict future events in our environment. In both presentation formats, the only source of information is the history of past realisations of the time series. Based on these values, the past experience is transferred to the forecast of the next situation. The presentation format does not affect the schemes, on average. However, some significant differences on the individual level are observed. Future research will therefore focus on the explanation of individual behavior.

6 References

Becker, O., Leitner, J., Leopold-Wildburger, U., 2004a. Expectation Formation in a Complex Information Environment. Working Paper of the European University Institute Fiesole, October 2004.

Becker, O., Leitner, J., Leopold-Wildburger, U., 2004b. Modeling Expectation Formation Involving Several Sources of Information. Working Paper of the European University Institute Fiesole, May 2004.

Becker, O., Leopold-Wildburger, U., 2000. Erwartungsbildung und Prognose - Ergebnisse einer experimentellen Studie. Austrian Journal of Statistics, 29, 7-16.

Becker, O., Leopold-Wildburger, U., 1996. The bounds and likelihood-procedure - A Simulation Study Concerning the Efficiency of Visual Forecasting Techniques. Central European Journal of Operations Research and Economics, 4, 223-229.

Beckman, S.R., Downs, D., 1997. Forecasters as imperfect information processors: Experimental and survey evidence. Journal of Economic Behavior and Organization, 32, 89-100.

Brennscheidt, G., 1993. Predicitve Behavior - An Experimental Study. Lecture Notes in Economics and Mathematical Systems Vol. 403.

Dwyer, G. P., Williams, A.W., Battalio R.C., Mason T.I., 1993. Tests of rational expectations in a stark setting. The Economic Journal, 103, 586-601.

Harvey, N., Bolger, F., 1996. Graphs versus tables: Effects of data presentation format on judgemental forecasting. International Journal of Forecasting, 12, 119-137.

Hey, J.D., 1994. Expectations Formation: Rational or Adaptive or...?. Journal of Economic Behavior and Organization, 25, 329-349.

Lawrence, M.J., Edmundson, R.H., O'Connor, M.J., 1985. An examination of the accuracy of judgmental extrapolation of time series. International Journal of Forecasting, 1, 25-35.

Makridakis, S., Hibon, M., 1979. Accuracy of forecasting: an empirical investigation (with discussion). Journal of the Royal Statistical Society A 142, 97145.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton J., Parzen E., Winkler R., 1982. The accuracy of extrapolative (times series) methods: the results of a forecasting competition. Journal of Forecasting, 1, 111-153.

Makridakis, S., Hibon, M., 2000. The M3-Competition: results, conclusions and implications. International Journal of Forecasting, 16, 451-476.

Muth, J.F., 1961. Rational Expectations and the Theory of Price Movements. Econometrica, 29, 315 - 335.

Schmalensee, R., 1976. An Experimental Study of Expectation Formation. Econometrica, 44, 17-41.

Theil, H., 1966. Applied Economic Forecasting. Chicago: Rand McNally.

Webby, R., O'Connor, M., 1996. Judgemental and statistical time series forecasting: a review of the literature. International Journal of Forecasting, 12, 91-118.

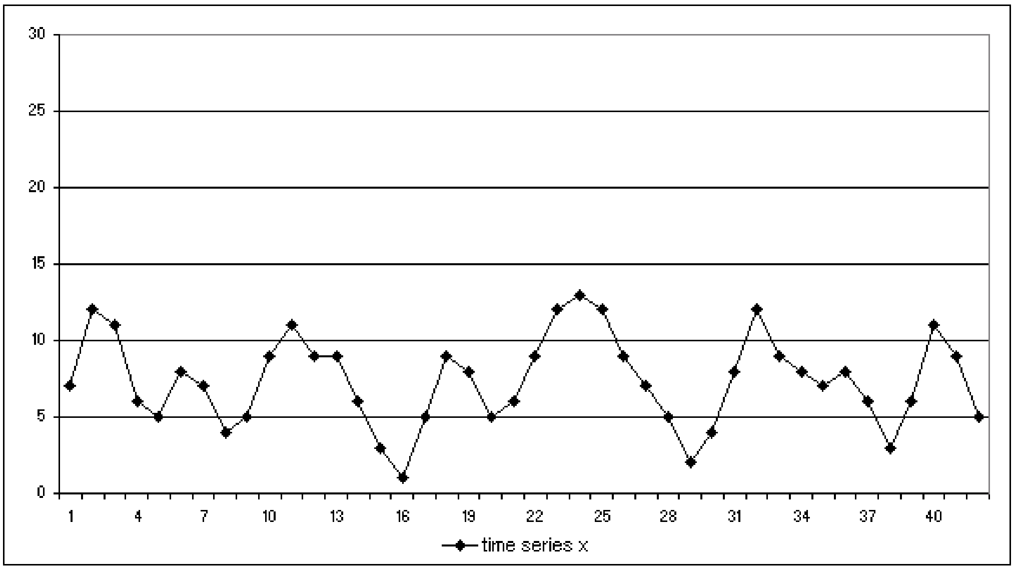

Figure 1: The time series in the graphical presentation format

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
7	12	11	6	5	8	7	5	4	9	11	9	9	6	3	1	5	9	8	5	6	9	12	13	12

Figure 2: The time series in the tabular presentation format

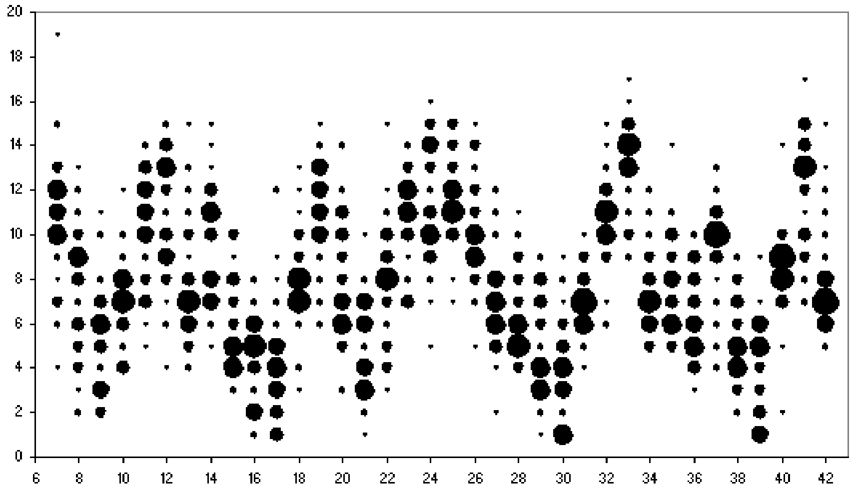

Figure 3: The distribution of the forecasts in the graphical presentation format

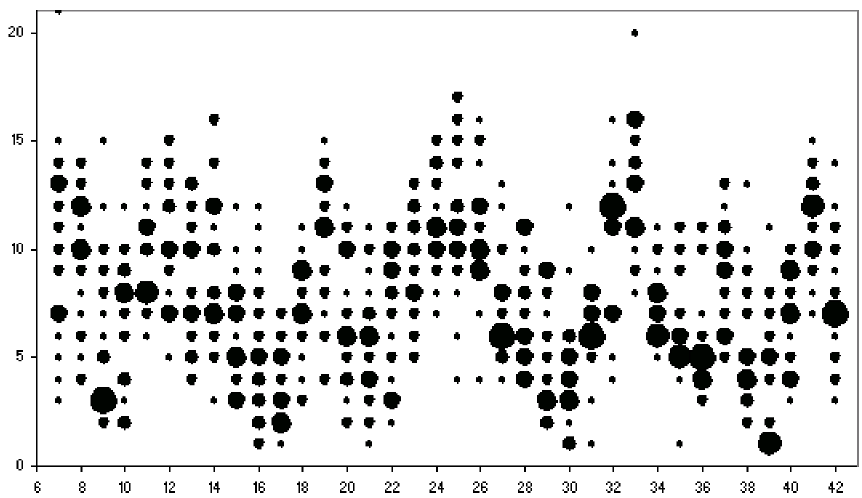

Figure 4: The distribution of the forecasts in the tabular presentation format

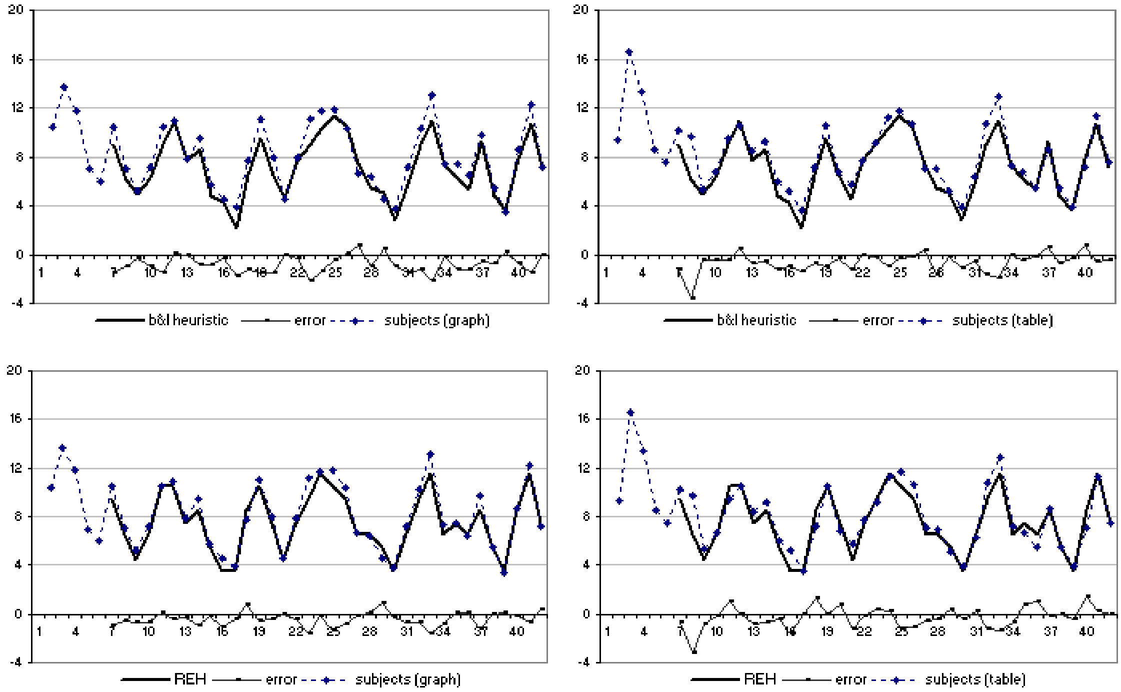

Figure 5: The subjects' average forecasts compared to REH and b&l

Forecasts	Individual Forecasts		Average Forecasts	
Format	Graph	Table	Graph	Table
MdAPE	26.136%	32.86%	21.407%	25.556%
ME	0.684	0.506	0.694	0.539
MAE	2.360	2.740	1.856	2.021
MSE	8.820	12.594	4.73	5.689
Theils' U	1.093	1.305	0.808	0.886
$(\overline{x} - \overline{f})^2$	0.483	0.426	0.481	0.291
$(SD_x - SD_f)^2$	0.078	0.219	0.096	0.283
r_{xf}	0.604	0.447***	0.731	0.639

Table 1: Accuracy of the average individual forecasts (***p<.001)

Model	**Format**	**Periods**	α	β	R^2	DW
b&l	Graph	7-42	0.479 (0.368)	1.046 (0.049)	0.931	1.811
		7-25	0.704 (0.505)	1.034 (0.068)	0.936	1.831
		25-42	0.247 (0.542)	1.58 (0.071)	0.933	1.824
	Table	7-42	1.079 (0.41)	0.94 (0.054)	0.898	1.822
		7-25	1.725 (0.609)	0.609 (0.081)	0.878	1.940
		25-42	0.478 (0.536)	1.000 (0.070)	0.927	1.917
REH	Graph	7-42	-0.146 (0.309)	1.078 (0.039)	0.957	1.902
		7-25	0.585 (0.406)	0.988 (0.051)	0.959	2.271
		25-42	-0.960 (0.406)	1.181 (0.052)	0.970	1.646
	Table	7-42	0.755 (0.504)	0.938 (0.064)	0.864	1.471
		7-25	1.861 (0.714)	0.803 (0.090)	0.833	1.679
		25-42	-0.475 (0.626)	1.091 (0.080)	0.921	1.407

Table 2: Regression results for both models