Heuristic Modeling of Expectation Formation in a Complex Experimental Information Environment

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

*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

Abstract: Academic subjects made judgmental forecasts of a graphically presented time series in a laboratory experiment. Besides the past realizations of the time series itself, the only available information for the forecasting task was provided by leading series, i.e. indicators with a constant lead period of one. The number and the quality of the leading series were varied systematically between seven versions of the experiment resulting in different levels of information complexity. We present a heuristic that explains the subjects' average forecasting behavior better than the rational expectations hypothesis in all versions of the experiment. Furthermore, we find that the forecasting accuracy of the subjects increases with the number of reliable indicators but their efficiency declines with increasing complexity.

JEL-classification: C91, C92, C53.
Keywords: Heuristics, Judgmental Forecasting, Expectation Formation, Experimental Economics, Rational Expectations.

Acknowledgements

The financial support of the Austrian Science Foundation (FWF) for the project *"Heuristics for the Subjective Prediction of Time Series"* P17156-N12 is gratefully acknowledged.

1 Introduction

Experimental studies of the forecasting behavior of subjects basically have one feature of their experimental setup in common: Subjects are given past realizations of a time series in graphical or tabular format and have to forecast the next value, period by period. One direction of the research is focused on how the forecasting accuracy of statistical models can be improved by human judgment. This addresses the aspects of performance depending on trends, forecasting horizon, available data points, the influence of contextual information, structural breaks and the development of appropriate procedures to combine judgmental and statistical forecasts. For an extensive review of the literature see Webby/O'Connor (1996).

Despite the practical relevance of forecasting situations with several sources of information, the presentation of additional information in a time series forecasting task has only been explored by a few authors. Lim and O'Connor (1996) gave subjects low-correlated and high-correlated causal information for the judgmental forecast of sales figures. Subjects were able to determine the reliability of causal information and to improve their forecasting accuracy. However, they used this information inefficiently. Similar results were also found in more recent studies by Sanders (1997) and Goodwin and Fildes (1999).

Another approach of the research in this context is the experimental analysis of forecasting behavior with the application of economic and psychological hypotheses of expectation formation. For instance it was shown that subjects are able to behave rationally when forecasting random walks but fail to do so with more complex time series (see e.g. Dwyer et al. (1993), Hey (1994) and Beckman/Downs (1997)). As above, the majority of this research is limited to the judgmental forecast of a time series without any additional information. Only Garner (1982) and Brennscheidt (1993) tested the rational expectations hypothesis in an environment of more than a single time series. In Garner's experiment, subjects had to predict the price of a fictitious agricultural commodity which was determined by an autoregressive model with three exogenous variables. Garner had to reject the hypothesis of rational expectations for almost all individuals and their average forecasts. Brennscheidt experimented with 14 different settings, of which all included exogenous variables. He found biases in individual and group forecasts which caused him to reject the rational expectations hypothesis.

We report on an experiment in which academic subjects make judgmental one-span forecasts of a graphically presented unlabelled time series over 42 periods. The subjects are not supported by statistical methods or any contextual information but they are given several partially contradictory and variably reliable time series (indicators) to support their forecasts. The stochastic characteristics of the indicators result in high and low correlation to the time series to be forecasted. Consequently, we call the indicators "weak" and "strong". All indicators have a constant lead of one period and they are available to the subjects when they make a forecast for the next period. Results of four prior versions of this experiment are available (see Becker/Leopold-Wildburger (2000) and Becker et al. (2004)): In version 0 no indicator is presented, in version 1 (2) one strong (weak) indicator is available and in version 3 both indicators are provided.

Becker et al. presented a simple heuristic based on visually perceived gestalts characteristics of the time series and applied the model to the average forecasts of the subjects. It was demonstrated that this bounds & likelihood heuristic (described in section 3) explains the subjects' average forecasts surprisingly well and outperforms the rational expectations hypothesis. In the study at hand, three further versions of the experiment are analyzed. In version 4 two strong indicators and one weak indicator are presented to the subjects when they make a forecast for the next period. In version 5 one strong indicator and two weak indicators are given and in version 6 two strong indicators and two weak indicators are given to the subjects. The crucial question is whether the subjects change their forecasting strategy in the complex information environments of versions 4, 5 and 6 or whether the rationale of the simple bounds & likelihood heuristic is still able to explain the average forecasting behavior.

In the next section we describe the experimental design and the total sample from the earlier versions of the experiment. In section 3, the bounds & likelihood heuristic is presented. In section 4, we will present the rational expectations hypothesis as a standard economic theory for the behavior of the subjects in our experimental setting. In section 5, both models will be compared. First, we test the models' ability to explain the forecasts of the subjects. Second, as a normative issue, the heuristic's relative efficiency on forecasting the time series itself will be analysed. Third, with the forecasts of about 600 subjects in the seven experimental versions altogether, the effects of providing additional information on the quality and the efficiency of the judgmental forecasts will be explored.

2 The experiment

2.1 The experimental context

In the experiment, subjects make judgmental forecasts of a time series x_t. The time series x_t (hereafter called "base 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 $x_1 = 7, x_2 = 12$, the endogenous variable x_t and the white noise u_t. The discrete random variable u_t is uniformly distributed in the interval [1,6] and was chosen to generate a stationary time series with a regular cyclical pattern. All values are integer due to the integer function *int*. For the forecast of x_t the subjects are given indicators (hereafter named "leading series") which give them a certain insight into the development of the base series. Four leading series were generated for the experiment as follows:

$$y_{k,t} = x_t + c_k + v_{k,t} \qquad \text{for } k = 1, 2, 3, 4. \tag{2}$$

The random variables $v_{1,t}$ and $v_{2,t}$ are triangularly distributed in the interval [-3,3] and the variables $v_{3,t}$ and $v_{4,t}$ are triangularly distributed in the interval [-5,5]. These characteristics of the random variables make the leading series $y_{1,t}$ and $y_{2,t}$ more reliable. The correlations between these strong indicators and the base series (0.85 and 0.88) exceed the correlations between x_t and the weak indicators $y_{3,t}, y_{4,t}$ (0.67 and 0.67). Without the white noise, the leading series would equal the base series plus 12 (c_2), 15 (c_1, c_3) and 17 (c_4) units, respectively. The leading series have a constant lead period of 1. All time series are presented to the subjects graphically. Figure 1 shows all time series applied in the experiment.

Insert Figure 1 about here.

Becker et al. (2000, 2004) reported on four versions (versions 0-3) of the experiment, in which no, one or two indicators were available to the subjects. The study at hand mainly deals with the versions 4, 5 and 6 in which up to two strong and two weak indicators are presented to the participants for their forecast of x_t. The number of weak and strong indicators in each version of the experiment is summarized in Table 1.

Insert Table 1 about here.

Figure 1 is identical to the information set of the subjects at the end of the experiment in version 6.

2.2 The experimental design

The versions 4, 5 and 6 of the experiment were conducted in summer semester 2004. Altogether 90 subjects participated, 30 in each version. They were recruited from undergraduate and graduate courses of business administration, economics and law. The subjects were told in the instructions that the indicators provide a certain insight into the change of the base series but no contextual information was given to the subjects and the time series were unlabelled.

The experiment was carried out with computers. Earlier versions of the experiment reported in Becker et al. (2000, 2004) were almost exclusively conducted with paper and pencil. In order to ensure comparability of the results, the computer interface was designed to reflect the size and structure of the paper chart and the experimental procedure was not altered: After a forecast for period t was entered, the true realization of x_t was graphed first and each value of the indicators for period $t+1$ was graphed in a fixed sequence afterwards. The most recent values of the base series and the leading series were also presented as figures. The time series were graphed in different colors. The subjects were allowed to proceed at their own pace but graphing the time series on the chart was performed slowly, which caused a minimum decision-making time of about 12 seconds per period. The subjects did not see a history of past values at period one of the experiment. Most subjects forecasted the mean of the presented indicators in the first period, i.e. values of about 25, because they were not given any hints on the level of the base series. They noticed quickly that the base series is below this level.

The subjects were given a significant financial incentive to forecast the time series accurately. The period-oriented payout function $p_t = 10 \cdot max\{3 - |x_t - f_t|; 0\}$ results in 30 Cents for an exact forecast f_t, a forecast error of one (two) unit(s) is rewarded with 20 (10) Cents. This function was also used in the earlier versions of the experiment. The average payments in version 4/5/6 of the experiment were 6.49/5.53/6.08 Euros at an average duration of about 25/30/35 minutes.

3 The bounds & likelihood heuristic

The bounds & likelihood heuristic (b&l heuristic) by Becker and Leopold-Wildburger (1996) models average forecasts. The calculation of the b&l heuristic is based on the same information that is available to the subjects, namely the time series' gestalt characteristics. One predictor is calculated for every time series and all predictors are aggregated to a combined forecast. In the versions 4, 5 (6) four (five) predictors are calculated, one for the base series and one for each leading series. In principle, the heuristic can be extended to any number of leading series in order to incorporate further information. When only base series information is available, other heuristics to explain how judgmental forecasts are formed have been proposed (e.g. Bolger and Harvey (1993)). We do not know any approaches in literature concerned with the problems of multiple sources of information.

For the calculation of the predictors, the base series and the indicators are distinguished. It is assumed that two characteristics of the base series are essential for the forecasts: the average variation and the turning points. It is assumed that the average absolute variations of the base 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. Therefore the subjects should on average predict changes which are 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 maxima observed so far (N_t) and the number of local maxima $\leq x_t$ (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 $\geq x_t$ (m_t) are considered for the calculation of $l_{t(trough)}$. This is shown in definition (3).

$$\begin{aligned} l_{t(peak)} &= \frac{1 + n_t}{2 + N_t} \\ l_{t(trough)} &= \frac{1 + m_t}{2 + M_t} \end{aligned} \tag{3}$$

If no extrema have occured so far, i.e. $m_t = M_t = 0$ and/or $n_t = N_t = 0$ the likelihoods of an upswing or downswing is 0.5. 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 predictor of the base series, $f^x_{t+1,b\&l}$, is described by definition (4):

$$f^x_{t+1,b\&l} = \begin{cases} x_t + b_t(1 - 2l_{t(peak)}) & \text{for } x_t > x_{t-1} \\ x_t + b_t(l_{t(trough)} - l_{t(peak)}) & \text{for } x_t = x_{t-1} \\ x_t - b_t(1 - 2l_{t(trough)}) & \text{for } x_t < x_{t-1} \end{cases} \tag{4}$$

The calculation of the predictors for the indicators is based on the idea that in most cases the base series and the leading series have different dimensions and average variations. In order to relate smaller (larger) average variations of the base series to an indicator y_k, we define the relation factor r^k_t. With a linear relation and a leading period of one, the predictors for the leading series in version 6 of the experiment can be described as follows:

$$f^{y_k}_{t+1,b\&l} = x_t + r^k_t(y_{k,t+1} - y_{k,t}) \quad \text{with } r^k_t = \frac{\sum_{i=1}^t |\Delta x_i|}{\sum_{i=1}^t |\Delta y_{k,i}|}, \text{ for } k = 1, 2, 3, 4. \tag{5}$$

With equations (4) and (5), five predictors are available. If these values differ from each other, a solution has to be found that gives more weight to the predictor with the lower forecasting error. The predictors are weighted with α_t reciprocally proportional to their absolute prediction error D_t:

$$D^s_t = \sum_{i=2}^{t} |f^s_{i,t,b\&l} - x_i| \qquad \text{for } s = x, y_1, y_2, y_3, y_4 \tag{6}$$

These are calculated on ex post basis which is indicated by the parameter i in $f^{y_k}_{i,t,b\&l}$. The calculation of the weights is demonstrated in equation (7):

$$\alpha^s_t = \frac{\frac{1}{D^s_t}}{\frac{1}{D^x_t} + \frac{1}{D^{y1}_t} + \frac{1}{D^{y2}_t} + \frac{1}{D^{y3}_t} + \frac{1}{D^{y4}_t}} \qquad \text{for } s = x, y_1, y_2, y_3, y_4 \tag{7}$$

Using these terms the combined forecast, $f_{t+1,b\&l}$, of the b&l heuristic is generated for version 6. The forecast values of the next period consist of a linear combination of all predictors:

$$f_{t+1,b\&l} = \alpha^x_t \cdot f^x_{t+1,b\&l} + \sum_{k=1}^{4} \alpha^{y_k}_t \cdot f^{y_k}_{t+1,b\&l} \tag{8}$$

In analogy, everything is computed for version 4 and version 5, respectively.

4 The Rational Expectations Hypothesis

The rational expectations hypothesis (REH) suggests that economic 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. For our experiment this implies that the subjects know the true models that generated the time series. 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. This information set allows the calculation of the rational expectations values.

In version 0 the REH values can simply be calculated from period 3 on ($x_1 = 7, x_2 = 12$) by replacing u_t in (1) with its expected value 3.5: $f^0_{t,REH} = x_{t-1} - int(\frac{1}{2} \cdot x_{t-2}) + 3.5$. These values represent the optimal forecasts when the true model is known. For our experiments with indicators (version 1-6), the calculation is more laborious. Auxiliary variables (g_t and h_t) need to be introduced. In order to avoid excessive notation the procedure is only described for version 6 of the experiment. It is assumed that the distributions of the discrete random variables u, v_1, v_2, v_3 and v_4 are known to the forecaster. The stochastic independence of the random variables allows the calculation of their simultaneous distribution $f(u, v_1, v_2, v_3, v_4) = f_u(u) \cdot f_{v_1}(v_1) \cdot f_{v_2}(v_2) \cdot f_{v_3}(v_3) \cdot f_{v_4}(v_4)$. Furthermore, equations (1) and (2) for k = 1,2,3,4 are assumed to be known to the forecaster in version 6. After reordering these equations we obtain:

$$u_t + v_{k,t} = y_{k,t} - c_k - x_{t-1} + int(\frac{1}{2} \cdot x_{t-2}) \qquad (9)$$

We introduce the Kronecker delta δ which has the value 1 if equation (9) is fullfilled simultaneously for all k = 1,2,3, 4 and has the value 0 otherwise. The auxiliary variables g_t and h_t can be enumerated by summing up all discrete values of the five random variables.

$$g_t = \sum_{u=1}^{6} \sum_{v_1=-3}^{3} \sum_{v_2=-3}^{3} \sum_{v_3=-5}^{5} \sum_{v_4=-5}^{5} u \cdot f_u(u) \cdot f_{v_1}(v_1) \cdot f_{v_2}(v_2) \cdot f_{v_3}(v_3) \cdot f_{v_4}(v_4) \cdot \delta \qquad (10)$$

$$h_t = \sum_{u=1}^{6} \sum_{v_1=-3}^{3} \sum_{v_2=-3}^{3} \sum_{v_3=-5}^{5} \sum_{v_4=-5}^{5} f_u(u) \cdot f_{v_1}(v_1) \cdot f_{v_2}(v_2) \cdot f_{v_3}(v_3) \cdot f_{v_4}(v_4) \cdot \delta \qquad (11)$$

With these auxiliary variables, the values of the REH for version 6 of the experiment are enumerated as follows:

$$f^6_{t,REH} = x_{t-1} - int(\frac{1}{2} \cdot x_{t-2}) + \frac{g_t}{h_t} \qquad (12)$$

The REH values for versions 4 and 5, $f^4_{t,REH}$ and $f^5_{t,REH}$, are enumerated analogously. With these values it can be tested whether the rational expectations hypothesis gives a valid explanation of the subjects' forecasting behavior.

5 Experimental results

It is of predominant interest whether the bounds & likelihood heuristic can explain the forecasts even in the environment of multiple leading series. It can be assumed that subjects who are confronted with a large amount of information apply selection strategies and do not combine all available predictors linearly as is done by the b&l heuristic. In this case the performance of the heuristic as a descriptor of average behavior would worsen in the versions 4, 5 and 6. The performance will be compared to the REH.

The second focus of the analysis is the forecast of the base series. The bounds & likelihood heuristic forecasts the time series with the same information that is available to the subjects, namely the observed realisations of the time series. We call this a visual forecasting technique. We are interested in the efficiency of these visual forecasts compared to the optimal forecasts of the REH. Furthermore, the forecasts of the subjects will be analyzed for their accuracy in all seven experimental versions. The information set which results in the highest forecasting accuracy will be discovered as well as the efficiency of judgmental forecasts when much additional information is provided. The efficiency of the subjects' forecasts will be investigated by a comparison to a statistical model.

The first six periods serve as a phase for familiarization and practice. Hence, periods 1 to 6 are not taken into account within the analysis. We only consider periods 7-42.

5.1 Modeling the average forecasts of the subjects

The variances of average forecasts vary between the periods. In some periods, the forecasts of the subjects accumulate around three or four values, while in others they diverge strongly. One reason for these differences can be found in the directions of change of the four (five) time series. When the leading series indicate different directions for period $t+1$ than the base series in period t, the subjects are confronted with divergent signals. They have to decide whether to follow the trend of the base series or to follow the trend of the indicators. Three categories of divergent signals are defined as index subsets $D_s, D_m, D_w \subseteq \{7, 8, ..., 42\}$:
In strongly divergent periods D_s the signs[1] of all indicated directions are equal and differ from the base series: $(sig_{y_1} = sig_{y_2} = sig_{y_3} = sig_{y_4}) \wedge (sig_x \neq sig_{y_k})$ for k = 1,2,3,4. In medium divergent periods D_m all indicated directions differ from the base series but the signs are not required to be equal: $sig_x \neq sig_{y_k}$ for k = 1,2,3,4. If only one of the leading series indicates a different direction than the base series, these periods D_w are denoted as weakly divergent: $sig_x \neq sig_{y_1} \vee sig_x \neq sig_{y_2} \vee sig_x \neq sig_{y_3} \vee sig_x \neq sig_{y_4}$. This definition implies $D_s \subseteq D_m \subseteq D_w$. The remaining periods are defined as convergent, whereas corresponding definitions exist and $C_s \subseteq C_m \subseteq C_w$. With exception of the weakly divergent periods in version 6,

[1] The signs are defined as $sig_x = sig(x_{t-1} - x_{t-2})$ and $sig_{y_k} = sig(y_{k,t} - y_{k,t-1})$ for k = 1,2,3,4.

the average variances of the forecasts are substantially higher in all divergent periods than in the corresponding convergent periods. The average variances are highest in version 5 and lowest in version 4, whereby these differences between the versions are highly significant. The presentation of indicators with high variance results in a high variance of average forecasts.

With this first impression of the subjects' forecasts, we now want to analyze the degree to which the average forecasts can be explained by the b&l heuristic and the REH. For this purpose, we estimate a simple linear regression with the average forecast $f_{t,avg} = \frac{1}{N}\sum_{i=1}^{N} f_t^i$ (with the subject i and N participants) in each version as a predicted variable and the two models as predictors:

$$f_{t,avg} = \alpha + \beta f_{t,\theta} \qquad \text{with } \theta = b\&l, REH. \tag{13}$$

The total sample is included in this regression but, by including all 36 periods, potential learning processes or the time invariance of the results cannot be observed. If the subjects learn to behave rationally or learn to behave in the sense of the b&l heuristic, then the coefficients of determination of regression (13) should increase over time. In order to test for learning behavior, the forecasting horizon is divided into three subsets, periods 7-18, 19-30 and 31-42. The acid test for the stability of the performance is the analysis of periods which are more difficult to predict according to our definitions of divergent and convergent periods. The total dataset of 36 periods is divided into these subsets.

Time series regressions often suffer from serial autocorrelation of the residuals. This phenomenon is diagnosed by the Durbin-Watson (DW) statistics. The low degrees of freedom in the sample subsets widen the inconclusive region of the DW statistics. Therefore, and in order to test for autocorrelation of higher time lags, we also apply the Beusch-Godfrey test. In Table 2, the DW statistics are reported. The results of the regressions are only reported for version 6 due to lack of space. The results of versions 4 and 5 are completely in line´with these results, thus no important information is lost by not reporting the figures. Only five of the 60 OLS regressions altogether suffer from significant serial autocorrelation. In these cases we estimate standard errors with a procedure as suggested by Newey/West (1987). The estimated coefficients are presented in Table 2 with their t-values in parentheses, whereby the null hypotheses are $H_0 : \alpha = 0$ and $H_0 : \beta = 1$.

In version 4/5/6 the heuristic explains 97.5%/95.2%/97.9% of the variance in the forecasts, while the REH explains only 77.1%/72.1%/76.2%. With the exception of the b&l heuristic in version 4, the intercepts of all estimates are not significantly different from zero and the slope coefficients are not significantly different from one. In the regressions of the subsets, the heuristic explains more than 90% of the variance of the average forecasts in all subsections. The REH performs worst in the last twelve periods with coefficients of determination of only 53.7%/51.2%/47.2%. Thus, the subjects do not learn to behave more rationally, quite the contrary. The b&l heuristic performs very well from the beginning of the considered periods to the end. The definition of divergent and convergent periods reveals two interesting phenomena:

First, independent of the definition of the divergent periods both models explain more of the variance of average forecasts in convergent periods. Second, the b&l heuristic explains more of the variance of average forecasts in almost all cases with a minimum coefficient of determination of 0.857. These results demonstrate that the b&l heuristic is clearly superior to the REH.

Insert Table 2 about here.

At this point it is interesting to analyze the performance of the heuristic and the REH in the four other versions of the experiment. For this purpose (13) is estimated for the average forecasts, the b&l heuristic and REH of versions 0-3. From earlier results[2] we know that the b&l heuristic performs better than the REH. The intercepts are significantly different from zero in many cases but this is true for both models. The coefficients of determination are almost equal in version 0 but differ clearly in the other versions. In versions 2 and 3, the coefficients of determination of the REH are only 0.639 and 0.651 while they are above 0.9 in the case of b&l.

It can be concluded that the b&l heuristic is a consistently better model than the REH. It explains average forecasting behavior surprisingly well, even in the complex information environments of versions 4, 5 and 6. The comparisons with experimental versions including less information sources reveal that the heuristic's performance is not only stable over the considered 36 periods within each version but also between the seven versions. The model explains 97.9% of the variance of average forecasts in version 6 in which four different leading series are presented to the subjects. The subjects obviously do not apply selection strategies on average as we hypothesized. The linear combination of all available predictors is a valid basis for the explanation of the average forecasting behavior.

5.2 The efficiency of the heuristic

The b&l heuristic represents a forecasting technique which is only based on visually perceived gestalts characteristics of the time series whereas the REH assumes perfect knowledge of the model. We are interested in the relative efficiency of our visual forecasting technique. In this context it must be considered that the experimental dataset of x_t, $y_{1,t}$, $y_{2,t}$, $y_{3,t}$ and $y_{4,t}$ generated by equations (1) and (2) represents only one of many results drawn from the independent distributions of u_t, $v_{1,t}$, $v_{2,t}$, $v_{3,t}$ and $v_{4,t}$. We want to demonstrate that our experimental results may not only be applied to this single case but also hold for other realizations. To face this crucial question we not only analyze the experimental dataset but also perform a Monte Carlo simulation independently of the experiment in order to generalize our results for other equally probable realizations of the time series. The vectors of X and Y are drawn according to (1) and (2) over 42 periods. This is replicated over 100,000

[2]For details see Becker et al. (2004).

times. The relative efficiency of the b&l heuristic is demonstrated by the measure

$$\eta = \frac{\sum_{t \in T} |f_{t,REH} - x_t|}{\sum_{t \in T} |f_{t,b\&l} - x_t|} \tag{14}$$

with $T = 7, ..., 42$. This simulation is performed for all seven experimental versions. The measure η of the experiment and the simulation (of the recent versions) are presented in Table 3.

Insert Table 3 about here.

Comparing the values of η of the experiment and the simulation, only small differences are found. With the exception of version 2, the values diverge by about five percentage points. These results prove that the experimental data are representative for the stochastic processes considered. The results also indicate that the efficiency of the b&l heuristic decreases from 90% to 60% between version 0 and version 6. This is not surprising because the REH generates perfect forecasts at a theoretically infinite number of leading series. The relative efficiency of the heuristic converges to zero in this case.

The b&l heuristic can be applied to an arbitrary number of leading series. However, from our results it can be concluded that the efficiency of this visual forecasting technique will consequently decline with an increasing number of indicators.

5.3 The accuracy of judgmental forecasts

Two aspects of the forecasting accuracy of the subjects are discussed: First, the seven versions of the experiment result in seven different information sets. It will be tested by a comparison of the forecasting errors which information set allows subjects to achieve the highest forecasting accuracy. Second, these forecasting errors are not only calculated for the average forecasts $f_{t,avg}$ but also on the individual level. These figures give insight into the effects of combining forecasts. It is a well known fact that the combination of separate forecasts yields to a lower forecasting error (see Makridakis/Winkler (1983)). From a comparison of the error measures the relative advantage achieved by the averaging of the forecasts will be identified.

The mean forecasting errors (ME), the mean absolute forecasting errors (MAE) and the mean squared forecasting errors (MSE) of the average forecasts and the average individuals in versions 4, 5 and 6 are reported in Table 4. We define the mean absolute forecasting error (MAE) of the average forecasts as $\text{MAE}_{avg} = \frac{1}{36}\sum_{t=7}^{42} |f_{t,avg} - x_t|$ and the average mean forecasting error of N individuals i as $\text{MAE}_{ind} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{36}\sum_{t=7}^{42} |f_t^i - x_t|$. The ME and MSE are defined correspondingly, whereby this definition implies $\text{ME}_{avg} = \text{ME}_{ind}$ (see Table 4).
The most accurate average forecasts are made by the subjects in version 1 who are given one reliable indicator (y_1). The presentation of the indicator y_3 in version 2

results in even lower forecasting accuracy than in the experiment without any indicators. Adding strong indicators improves the forecasting accuracy, adding weak indicators reduces it. This finding is consistent over the error measures in the experimental versions. The ranking of the versions changes only slightly when the average individual performance is considered. Subjects in version 1 achieve the best results and subjects in version 0 the worst. The differences between MAE_{avg} of the seven versions are significant according to a Kruskal-Wallis-H test (Chi-square=18.915, sig.=0.004). The same is true for the differences of MAE_{ind} (Chi-square=343.617, sig.=0.000). Hence, the performance of the groups and the individuals are comparable between the versions.

In Table 4, the MAE-relations $\text{MAE}_{avg}/\text{MAE}_{ind}$ of version 4,5 and 6 are reported. These relations range from 69% in version 0 to 83.4% in version 3. This means that the MAE^0_{avg} of the 264 subjects who participated in version 0 is 31% lower than MAE^0_{ind}. Combining the forecasts by a simple procedure such as the arithmetic mean increases the forecasting accuracy tremendously. Even in version 3, where this advantage is smallest, the combination of the forecasts results in a reduction of the mean absolute error of 16.6%.

Insert Table 4 about here.

5.4 The efficiency of average forecasts

In the experiment, 596 different subjects participated altogether. In seven versions they were provided with a systematically varied number and quality of additional information. As it was shown in the preceding section, their forecasting accuracy increases by adding information in terms of (reliable) indicators. However, the more information is provided to the subjects, the less efficient the forecasts will be even if the accuracy increases. We will test the efficiency of the subjects' forecasts by comparing them to the performance of a statistical model.

The collective contribution of 596 subjects to the forecast of the base series can be measured by at least two methods: The most simple way of combining all forecasts is their arithmetic mean value. With a VAR model, we will test whether a weighted combination of the forecasts will achieve better results. In the latter case, the base series x_t is explained by the seven average forecasts:

$$x_t = \alpha + \beta_1 f^0_{t,avg} + \beta_2 f^1_{t,avg} + \beta_3 f^2_{t,avg} + \beta_4 f^3_{t,avg} + \beta_5 f^4_{t,avg} + \beta_6 f^5_{t,avg} + \beta_7 f^6_{t,avg} \quad (15)$$

In the former case a simple linear regression with the average forecast of all subjects $\frac{1}{596}\sum_{i=1}^{596} f^i_t$ as a predictor for the base series is performed.
As one possible statistical procedure in the forecasting task, a VAR model is estimated with two time lags and $y_{1,t}, y_{2,t}, y_{3,t}$ and $y_{4,t}$ are used as predictors for the base series x_t in order to generate a comparable statistical model for the subjects' forecasts:

$$x_t = \alpha + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 y_{1,t} + \beta_4 y_{2,t} + \beta_5 y_{3,t} + \beta_6 y_{4,t} \quad (16)$$

In analogy to the efficiency measure η the relations of mean absolute errors of the regressions are calculated. When the periods 3-42 are considered the regression including the average of all 596 individual forecasts achieves a relative efficiency of only 33.6% while model (16) achieves 45.7%. These relations are fairly constant when the database of considered periods is reduced, for instance to periods 3-20 or 3-30. The efficiency of the forecasts are low even at the beginning of the experiment when only few observations are available. However, we found that the efficiency of the subjects can be improved by applying alternative combination procedures.

6 Conclusion

In this study we report on seven versions of a forecasting experiment. The bounds & likelihood heuristic explains average forecasting behavior very well and consistently better than the rational expectations hypothesis, especially when a large number of leading series is provided. The representativeness of the experimental results for the considered stochastic processes was proved with a simulation study. We analysed the forecasting accuracy of the subjects and found reliable information to improve the performance. We also demonstrated that the efficiency of the forecasts declines with an increasing number of indicators.

It is a surprising finding that the heuristic performs best in version 6, when the information set includes four leading series for the forecast of the base series. We hypothesized that in versions 4, 5 and 6 the rationale of the heuristic, namely the linear combination of up to five predictors, will not explain subjects' average behavior properly. The opposite is true. However, it cannot be expected that this result holds for an arbitrary number of leading series. The information processing capacity of the subjects is limited by the visual perception of the time series. If they are overloaded with information they may select single leading series or they may not be able to determine the reliablity of information and pay attention exclusively to the base series. This will be analyzed in future research, but consisdering our results on the Rational Expectations Hypothesis and statistical analyses it has to be considered that the efficiency of the subjects and the bounds & likelihood heuristic will decline with a growing number of indicators.

7 References

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.

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

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

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

Bolger, F., Harvey, N., 1993. Context-sensitive Heuristics in Statistical Reasoning. The Quarterly Journal of Experimental Psychology, 46A, 779-811.

Brennscheidt, G., 1993. Predictive 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.

Garner, A.C., 1982. Experimental Evidence on the Rationality of Intuitive Forecasters, In: V.L. Smith (Eds.), Research in Experimental Economics, Greenwich. Conn: JAI Press, pp. 113-128.

Goodwin, P., Fildes, R., 1999. Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy? Journal of Behavioral Decision Making, 12, 37-53.

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

Lim, J.S., O'Connor, M., 1996. Judgmental forecasting with time series and causal information. International Journal of Forecasting, 12, 139-153.

Newey, W. K., West, K. D., 1987. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55, 703-708.

Markidakis, S., Winkler, R.L., 1983. Averages of forecasts: some emprical results. Management Science, 29, 987-996.

Sanders, N.R., 1997. The impact of task properties feedback on time series judgmental forecasting tasks. Omega: International Journal of Management Science, 25, 135-144.

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

Figure 1: The base series and all indicators used in the experiment

Version	0	1	2	3	4	5	6
Strong Indicators	0	1	0	1	2	1	2
Weak Indicators	0	0	1	1	1	2	2

Table 1: Number of strong and weak indicators in the experiment

model	periods	α	β	R^2	DW	df
b&l	7-42	0.246 (1.265)	0.996 (-0.100)	0.979	1.997	34
	7-18	0.224 (0.818)	1.010 (0.256)	0.985	2.150	10
	19-30	0.278 (0.879)	0.979 (-0.553)	0.985	1.894	10
	31-42	-0.182 (-0.323)	1.055 (0.764)	0.956	1.802	10
	C_s	0.208 (0.919)	1.001 (0.034)	0.979	2.456	26
	D_s	0.503 (1.210)	0.958 (-0.737)	0.979	3.056	6
	C_m	0.133 (0.540)	1.007 (0.226)	0.980	2.533	21
	D_m	0.601 (1.768)	0.953 (-1.022)	0.975	2.524	11
	C_w	0.401 (0.902)	0.984 (-0.281)	0.984	1.069	5
	D_w	0.185 (0.809)	1.001 (0.033)	0.977	2.001	27
REH	7-42	0.734 (1.074)	0.901 (-1.151)	0.762	1.777	34
	7-18	0.711 (0.785)	0.869 (-2.025)	0.839	0.949+	10
	19-30	-0.639 (-0.639)	1.096 (0.7803)	0.889	1.923	10
	31-42	3.175 (1.955)	0.601 (-1.985)	0.472	1.919	10
	C_s	0.832 (1.072)	0.898 (-1.052)	0.766	1.557	26
	D_s	0.339 (0.207)	0.915 (-0.401)	0.756	1.364	6
	C_m	0.721 (0.896)	0.896 (-1.051)	0.798	1.499	21
	D_m	0.512 (0.298)	0.941 (-0.257)	0.604	2.360	11
	C_w	-0.967 (-1.432)	1.005 (0.067)	0.972	1.545	5
	D_w	0.966 (1.150)	0.896 (-0.945)	0.712	1.841	27

Note: t-values of $H_0 : \alpha = 0$ and $H_0 : \beta = 1$ in parentheses
+: significant serial autocorrelation of lag 1 or higher; in these cases standard errors are enumerated with the method by Newey/West (1987).

Table 2: Regression results in version 6

Dataset	Measure	Ver. 4	Ver. 5	Ver. 6
Simulation	$\|f_{t,b\&l} - x_t\|$	0.993	1.090	0.969
	$\|f_{t,REH} - x_t\|$	0.630	0.710	0.584
	η	63.5%	65.1%	60.3%
Experiment	$\|f_{t,b\&l} - x_t\|$	1.203	1.405	1.224
	$\|f_{t,REH} - x_t\|$	0.726	0.883	0.698
	η	60.3%	62.8%	57.0%

Table 3: Relative performance of the b&l heuristic in experimental and simulated datasets

Version		Average Forecasts			Individual Forecasts			MAE-
	ME	MAE	MSE	Theil's U	MAE	MSE	Theil's U	relation
4	0.163	1.157	1.949	0.519	1.417	3.359	0.672	81.7%
5	0.314	1.445	3.276	0.672	1.806	5.664	0.870	80.0%
6	0.269	1.213	2.311	0.565	1.593	4.441	0.765	76.0%

Table 4: Forecasting accuracy of the subjects in versions 4,5 and 6 of the experiment