IMPROVING FORECAST ACCURACY BY GUIDED MANUAL OVERWRITE IN FORECAST DEBIASING

Research in Progress

Blanc, Sebastian M., Karlsruhe Institute of Technology, Karlsruhe, GER, sebastian.blanc@kit.edu
Setzer, Thomas, Karlsruhe Institute of Technology, Karlsruhe, GER, thomas.setzer@kit.edu

Abstract

We present ongoing work on a model-driven decision support system (DSS) that is aimed at providing guidance on reflecting and adjusting judgmental forecasts. We consider judgmental forecasts of cash flows generated by local experts in numerous subsidiaries of an international corporation. Forecasts are generated in a decentralized, non-standardized fashion, and corporate managers and controllers then aggregate the forecasts to derive consolidated, corporate-wide plans to manage liquidity and foreign exchange risk. However, it is well-known that judgmental predictions are often biased, where then statistical debiasing techniques can be applied to improve forecast accuracy. Even though debiasing can improve average forecast accuracy, many originally appropriate forecasts may be automatically "corrected" in the wrong direction, for instance, in cases where a forecaster might have considered knowledge on future events not derivable statistically from past time series. To prevent high-impact erroneous corrections, we propose to prompt a forecaster for action upon submission of a forecast that is out of the confidence bounds of a benchmark forecast. The benchmark forecast is derived from a statistical debiasing model that considers the past error patterns of a forecaster. Bounds correspond to percentiles of the error distribution of the debiased forecast. We discuss the determination of the confidence bounds and the selection of suspicious judgmental forecasts, types of (statistical) feedback to the forecasters, and the incorporation of the forecaster's reactions (comments, revisions) in future debiasing strategies.

Keywords: Decision Support Systems, Judgmental Forecasting, Forecast Bias Correction, Cash Flow Forecasting

1 Introduction

Accurate forecasting is considered vital by today's enterprises since corporate plans as well as planning and decision making in almost all functional units heavily depend on forecast quality. For instance, forecasts of future cash flows play a pivotal role in corporate financial controlling and planning tasks, where the forecasts are the key inputs for liquidity and foreign exchange risk models. As of today, most forecasting tasks are dominated by human judgment (Klassen and Flores, 2001; McCarthy et al., 2006; Sanders and Manrodt, 2003).

Since judgmental forecasts are often produced by different experts with individual backgrounds, attitudes and forecasting procedures, forecasts are regularly found to be biased, resulting in decreasing accuracy and business performance (Leitner and Leopold-Wildburger, 2011). Several empirical studies provide evidence for the existence of cognitive biases in judgmental forecasting. Hogarth and Makridakis, 1981 and Lawrence et al., 2006 provide thorough overviews of heuristics and biases regularly found in judgmental forecasting.

A forecast decision support system (DSS) supports forecasting tasks by gathering, filtering, and presenting relevant information. For instance, provided information range from statistical model forecasts and values of leading indicators to qualitative information such as recent political or economical news. However, several experiments have shown that providing additional information and statistics does not have an unambiguously positive effect on forecast accuracy. For instance, it has been found that providing statistical forecasts as orientation increases accuracy only in stationary settings (Goodwin and Fildes, 1999), or when asking experts to adjust a statistical forecast only when providing a reason for the adjustment (Goodwin, 2000b). While approaches based on presenting and explaining model forecasts can improve judgmental forecast accuracy, it was also found that forecasters in general use information inefficiently, acquire too little of the available information, and are overconfident in their own expectations even if their own forecasting ability is shown to be significantly inferior to the forecast provided by a software (Leitner and Leopold-Wildburger, 2011). As a consequence, biases can often not be completely removed (Bhandari, H., and Deaves, 2008; Lim and O'Connor, 1996), but adequate DSS can nevertheless significantly mitigate observed biases (George, Duffy, and Ahuja, 2000; Remus and Kottemann, 1995).

As an alternative to mitigating biases using DSS, statistical techniques can be used to detect and correct biases ex post, but before the values are used for planning and decision making. While applying these techniques can improve overall accuracy, as for instance shown by Elgers, Lo, and Murray, 1995; Goodwin, 1996, 2000a, the automatic debiasing of forecasts bears the risk of erroneously "correcting" of several originally appropriate forecasts. For instance, an expert might have considered knowledge on future events not derivable by statistical means from past time series and error histories.

In our work we address the research questions, how to integrate statistical debiasing results into a model-driven DSS in order to differentiate erroneous from beneficial correction. We address the issue of erroneous corrections by augmenting debiasing techniques with a model-driven DSS aimed at mitigating high-impact biases. When a novel forecast has been submitted, a debiased benchmark forecastis calculated concurrrently together with its confidence bounds. The forecaster is prompted for action if – and only if - his forecast is out of these bounds and can either justify or revise his expectation. This procedure has several beneficial properties. Most importantly, we avoid erroneously altering highly accurate forecasts by allowing the forecaster to justify his forecast. In doing so, forecasters can be expected to be focused on queried forecasts since queries require strong statistical motivation. When prompted, the (debiased) benchmark forecast is based on biases and errors observed in past forecasts of the expert himself, which can be explained statistically and verbally.

The remainder of the paper is structured as follows. First, in Section 2 we briefly review forecast debiasing techniques. We then present a case study with cash flow forecasts of a large international corporation in Section 3.Finally, in Section 4 we discuss the required mechanism, in particular regarding confidence corridors, selection of suspicious forecasts, and "'human–machine' interactions with the forecasters.

2 Statistical Debiasing Methods

The most common approach to statistical forecast debiasing is Theil's method (Theil, 1966). For a forecast F_t for an actual item A_t at time t, the parameters for the debiasing of forecasts can be estimated by regressing the actuals on the forecasts: $A_t = \alpha + \beta F_t + \varepsilon_t$, with residuals ε_t. α and β are estimated as a and b; the debiased forecast can then be calculated as $C_t = a + bF_t$. Theil's method uses ordinary least squares (OLS) for parameter estimation; it has been evaluated in numerous studies and for various applications, for instance in a laboratory setting as well as on empirical earnings and sales forecasts (Elgers, Lo, and Murray, 1995; Goodwin, 1996, 2000a).

As an extension of Theil's method, Goodwin, 1997 proposes using discounted weights of errors in the estimation to give more weight to more recent observations. The weighting of past observation, in contrasts to equal weights in Theil's method, is motivated by potentially time-varying biases, for instance because of improved forecasts or structural changes. Technically, the errors are weighted geometrically in

this model; descending weights γ^t with discount factor $\gamma \geq 1$ are used in a weighted least squares (WLS) regression. Since OLS corresponds to WLS with $\gamma = 1$, i.e. equal weights, we treat Theil's method as a special case of WLS-based debiasing. Goodwin, 1997 have shown that weighted estimation performs better than Theil's method on various types of time series.

Both Theil's method and the weighted approach are well established and we will use both methods later in our use case. As mentioned above, debiasing can only be expected to improve *average* forecast accuracy . Accuracy increases for many forecasts, while decreases for some others. This is especially true when one-time events occur that lead to a strong deviation of time series from past developments and patterns.

3 Case Study: Debiasing of Cash Flow Forecasts

Cash flow forecasts play a pivotal role in corporate financial management tasks. For instance, the forecasts are used in liquidity management to ensure solvency and in foreign-exchange risk management to identify and hedge exposures resulting from foreign business activities. Inaccurate forecast are an unreliable basis for corporation-wide financial plans and can lead to liquidity shortages, uncovered currency risks or increased hedging costs. In large multinational companies, cash flow forecasts are prepared for individual subsidiaries from different countries and distinct business divisions and corporation-wide financial plans are derived by consolidating the delivered forecasts.

Unfortunately, while in general there is awareness of the importance of accurate financial forecasts for corporate planning and control (Graham and Harvey, 2001; Kim, Mauer, and Sherman, 1998), there is practically no research available that empirically analyzes corporate cash flow forecasts. Hence, corporate financial controllers have little guidance on how to improve the quality of their cash flow forecasts. In this study we examine to what extend forecast accuracy can be improved by removing biases in cash flow forecasts. Our work is motivated by many studies reporting successful debiasing of judgmental forecasts with statistical techniques. In particular, we study how often automatical forecast debiasing leads to increases or decreases of accuracy, before we then introduce or model-based DSS.

3.1 Sample Company and Available Data

The data used in our analysis is based on a unique dataset of real-world cash flow forecasts and corresponding realizations provided by our sample company, a large multinational corporation. The corporation is a diversified group with three large and relatively independent business divisions that are very different in terms of products and markets. We name the divisions "agricultural products" (AP), "health and pharmaceutics" (HP) and "industrial materials" (IM). The company has a legal structure with over 300 separate legal entities. Financial management is however centralized with local financial managers at the subsidiaries reporting to the company's central finance department. Cash flow forecasts are generated worldwide by the partner company's subsidiaries and are delivered to the finance department.

The forecasts cover monthly intervals with forecast horizons of up to at least 12 months and are available for invoices issued and invoices received. Our dataset comprises actual cash flow volumes from January 2008 to December 2013, the corresponding forecasts were delivered from November 2006 to September 2013 by the 34 largest subsidiaries on a quarterly basis and for a total of 25 currencies. Overall, the raw dataset consists of 10,656 monthly cash flow volumes with 12 associated forecasts (with different horizons). Individual actual time series contain 72 values (one per month in the period of available data).

For a cross-validation of debiasing methods, parameters are estimated and evaluated per time series in a rolling evaluation. For this purpose, we use the last 24 months (01/2012 – 12/2013) as evaluation period for which forecasts are debiased out-of-sample. The parameters for debiasing are always estimated using the whole history of available forecasts and actuals prior to the month a (debiased) forecast is generated, since debiasing can be assumed to heavily rely on the incorporation of current biases.

3.2 Evaluation of Forecast Debiasing

In our experiments, we apply statistical debiasing using WLS with discount factors of 1 (corresponding to Theil's method) and 1.1 (which performed best in a case study by Blanc and Setzer, 2015) to our empirical judgmental forecasts and quantify and compare their performance. We apply both methods to different subsamples generated by splitting the dataset per business division and invoice type (resulting in six subsets plus one sample containing all datasets).

As evaluation criterion, we use the absolute percentage error (*APE*). With actual volume A and corresponding forecast F, *APE* is defined as $|(A-F)/A|$. *APE* is a common forecast accuracy measure for forecasts with widely differing volumes (and widely differing error volumes, respectively). We calculate the error of the original forecast (F) as well as of the debiased forecast (C) and then calculate the relative *APE* improvement: $\Delta APE = (APE_F - APE_C)/APE_F$. Based on the recommendation by Armstrong and Collopy, 1992 for sets of errors, we use the median ΔAPE (instead of the mean value) as it is less prone to extreme values. For hypothesis testing, we use a transformed version defined as $\Delta_T APE = -\ln(1-\Delta APE)$ with an approximately symmetric distribution.

Aggregated results are presented in Table 1. Each row of the table corresponds to one subset of the data (combination of business division and invoice type). For each subset, median ΔAPE for $\gamma = 1$ and $\gamma = 1.1$ is presented.

Business Division	Invoice Type	Median ΔAPE ($\gamma = 1$)	Median ΔAPE ($\gamma = 1.1$)
All	All	- 0.3 %	8.0 % ***
AP	II	1.1 %	2.6 %
AP	IR	11.2 % ***	16.3 % ***
HP	II	- 4.4 %	4.0 % ***
HP	IR	6.5 % ***	15.3 % ***
IM	II	0.4 %	8.2 % ***
IM	IR	- 12.9 %	10.3 % ***

Table 1. Accuracy improvements of debiasing methods with weight 1 *and* 1.1*; stars indicate results of one-sided Wilcoxon signed-rank tests for improvements ($\Delta_T APE > 0$)*

Taking median ΔAPE across all datasets (first row of the table) as a criterion, debiasing leads to a highly significant accuracy improvement ($p < 0.001$) of 8.0% when a discount factor is used. In contrast, using Theil's method does not lead to a significant improvement of forecast accuracy. However, the performance of both debiasing methods strongly depends on the specific business division as well as on invoice type. For $\gamma = 1.1$, accuracy is increased significantly for invoice received and issued for business divisions HP and IM. Improvements are considerably higher for invoices received for HP than for invoices issued. A possible explanation is that business is overall rather stable in HP making invoices the companies issue much easier to forecast (and control) than invoices received from partners, which for instance can be unexpectedly shifted from one month to the next. For division IM, which is strongly influenced by macro-economic uncertainty, improvements differ only slightly between invoices issued and received. In business division AP, with a high seasonality of cash flow time series, only the accuracy for invoices received is improved - forecasts for invoices issued seem to be largely unbiased.

In contrast, using Theil's method with equal weights for all observation only leads to significant accuracy improvements for two data subsets: invoices received for business divisions AP and HP. Accuracy for all other subsamples is not, or even negatively, influenced by Theil's method. Using a weighted method therefore seems to be crucial for applying forecast debiasing methods to real-world corporate time series data. This is reasonable and fits the original motivation for the introduction of descending weights in statistical prediction and regression: unweighted models never "forget" over-aged values (as well as outliers), leading to decreased robustness and accuracy.

Overall, we observe a significant error reduction using debiasing techniques with a moderate discount factor of 1.1. However, since debiasing is applied to all forecasts regardless of the particular confidence in the original forecast (or the corrected one), it is likely that originally appropriate forecasts are also corrected (in the wrong direction), leading to higher errors than necessary.

3.3 Good versus Bad Corrections

Figure 1 shows the conditional distribution of the error of the original forecast depending on the position of the forecast in the error distribution around the corresponding debiased forecast. Results for forecasts of cash flows in 2013 are displayed. The error distributions of the corrected forecasts are pre-generated using data from 2011 and 2012 and $\gamma = 1.1$. Technically, the errors of the debiased forecast for actual items in 2011 and 2012 are used for estimating the variance of errors, s^2. In line with common forecasting literature, we assume that errors follow a normal distribution with mean 0. For the data for 2013, we then calculate the p-value of the quantile of F_E in $N(F_C, s^2)$ for each original forecast F_E and corresponding corrected forecast F_C.

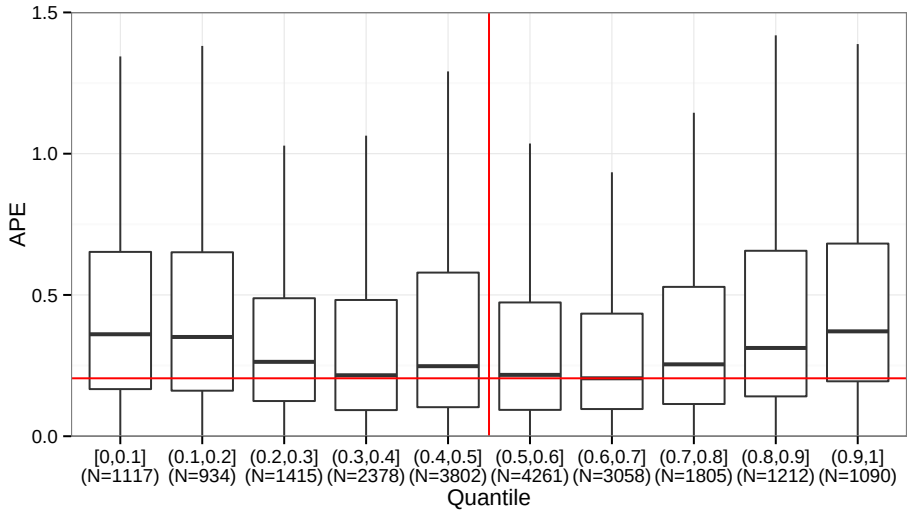

Figure 1. Forecast error of the original forecast (measured by APE) depending on the position in the confidence interval around the corrected forecast (generated with $\gamma = 1.1$)

The plot shows that the median *APE* of an original forecast increases with its distance to the value of the debiased forecast (the center of the distribution, red vertical line). However, it can also be seen that over 25% percent of the original forecasts in the outer-percentiles of the distribution have a very low error, even below the median error of the best quantile bin (20% in our case, red horizontal line)). This group of forecasts may then be adjusted by the debiasing-model in a way that the *APE* is increased. This especially impacts overall accuracy when actuals (and therefore errors) are large in volume.

This issue is further illustrated in Figure 2, where the *APE* difference between debiased forecasts and the original forecasts are depicted. Difference per quantile bin are approximately symmetric, with positive median improvements. However, the larger the difference between original and debiased forecast (outer quantile bins) the higher the variance of error differences. Very large improvements or declines in accuracy can in many cases be noted for the most deviating bins. This is quite intuitive as strongly differing errors between original and debiased forecasts can only occur in case of large difference between the two forecasts.

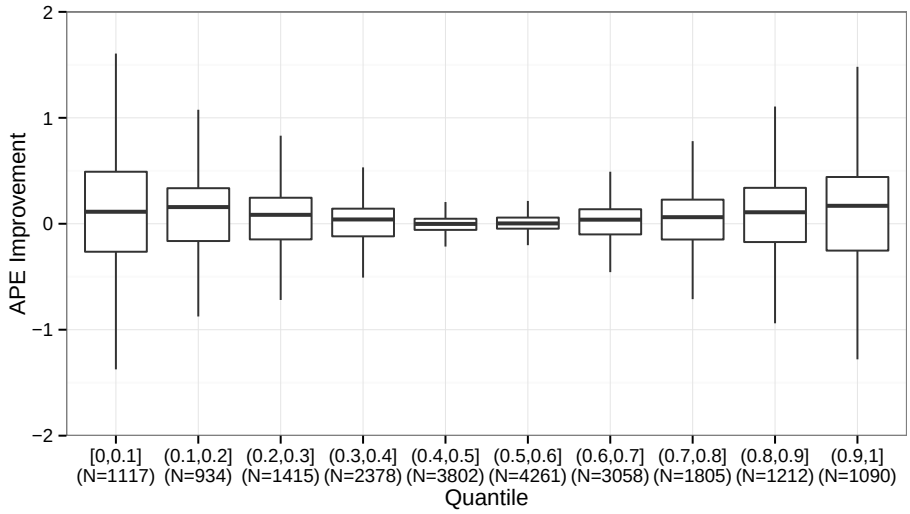

Figure 2. Improvement of the forecast error (measures by APE) depending on the position in the confidence interval around the debiased forecast (generated with $\gamma = 1.1$)

4 Guided Manual Overwrite Mechanism

The idea of the DSS we propose in this paper is to differentiate the critical cases pointed out in the case study – which lead to strong accuracy improvements or declines –, by 'asking' the forecasters directly. A key challenge is the determine a relatively small set of suspicious forecasts with a large impact on overall accuracy to ensure that the queries are handled with adequate care.

In order to allow for an effective application of this idea, a very cautiously designed mechanism is required, addressing the following aspects regarding the identification of suspicious forecasts as well as feedback and interaction design. First, we need to identify suspicious forecasts. Therefore, we compute and apply the error distribution of the debiased forecast in order to determine the position (quantile) of the original forecast as described in Section 3.3. Based on a predefined confidence level α, we check if the p-value lies out of the confidence interval ($\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$). For this step, a reasonable confidence level α has to be identified, further referred to as $CONFIDENCE\_LEVEL$, which controls the number of queries and should limit them to a manageable number. In addition, we need to focus on high-volume (and therefore high-impact) forecasts. This is necessary to avoid forecasters being queried too often or with too many low-impact items, leading to reluctance and carelessness when handling future queries. We consequently have to define a threshold volume, $VOLUME\_LEVEL$ by setting a lower bound for the difference (in volume) between debiased and original forecast.

Second, we need to determine a feedback and interaction design. Feedback design is related to the question, which information and in which form are presented to the forecaster when he is prompted for action. The options range from simply presenting the debiased forecast without further information ($FORECAST\_ONLY$) or presenting additional statistics like the position of the original forecast in the debiased forecasts's distribution ($FORECAST\_DISTR$) to providing additional verbal information on the type and interpretations of the assumed biases ($FORECAST\_BIAS$). Over time, additional statistics on recent actions (revise, keep and comment) and their ex post results can be presented ($ACTION\_ANALYSIS$).

The option $FORECAST\_DISTR$ is illustrated in Figure 3. Among the submitted forecasts for different currencies and months (shown in the table), suspicious and large-impact forecasts are marked (forecast for EUR in 03/2015 in the figure) and an action by the forecaster is required. Based on the position of the original forecast in the probability distribution around the debiased forecast, which is displayed to the forecaster, a revision of the forecast or ignoring the query (plus a comment) are potential actions.

	01 / 2015	02 / 2015	03 / 2015	04 / 2015	05 / 2015
AUD (m)	1.02	3.64	14.24	6.28	16.91
EUR (m)	7.98	12.47	**10.31**	4.59	5.11
KRW (bn)	12.27	12.01	11.60	5.68	17.40
USD (m)	5.33	3.81	5.44	8.80	7.65

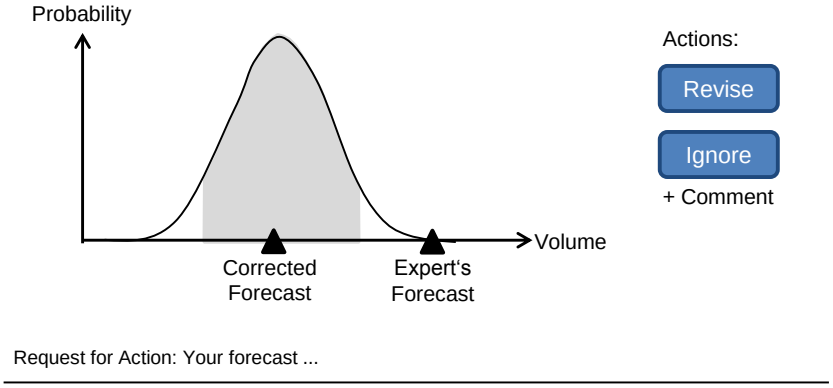

Figure 3. Illustration of Feedback-option FORECAST_DISTR. A forecaster is prompted for action as one of the delivered forecasts (cash flows in EUR in 03 / 2015, as marked in the table) falls out of a confidence band computed on the distribution of past errors of the debiased forecasts. The forecaster can either revise his forecast towards the proposed statistical estimate or keep his forecast unchanged and provide an explanation.

The *FORECAST_BIAS* option includes a description of biases found in past forecasts. For example, basis biases that can be identified using linear techniques are the so called mean and regression biases. The mean bias corresponds to a systematic under- or overestimation of the actual value. The regression bias covers a component of the forecasts which is (linearly) correlated to the actuals but nevertheless does not scale perfectly with actual values. For instance, a regression bias exists if small values are slightly and high values are substantially underestimated (or overestimated). Both biases (if present) can be quantified and might be presented to the forecaster in this modus. The option *ACTION_ANALYSIS* is of particular interest to handle overconficence, i.e., if some forecasters repeatedly decline revising forecasts that are more accurate in most (or even all) cases.

5 Summary and Outlook

In this paper, we studied debiasing of judgmental forecasts using statistical techniques in a case study based on a unique dataset of real-world cash flow forecasts and corresponding realizations provided by a large multinational corporation. We showed that overall the accuracy of forecasts can be improved significantly when using a weighted, linear debiasing approach. However, the case study also revealed that numerous cases of initially highly accurate original forecasts are automatically 'corrected' in the wrong direction, leading to decreased accuracy in many cases. We therefore proposed an approach to differentiate cases leading to strong accuracy improvements or declines by querying forecasters directly via supporting statistics and an interaction with a forecast DSS.

This procedure has several beneficial properties. The system prompts forecasters for action only if a submitted forecast is out of confidence bounds and of high impact on overall accuracy. Therefore, it is more likely that a forecaster is more focused on forecasts declared as suspicious. In these cases, the benchmark forecast is based on his own biases observed in past forecasts and can therefore be explained statistically and verbally. Furthermore, based on a previous study on forecast debiasing in the domain of

corporate financial forecasting, we found that debiasing judgmental forecasts often outperformed purely model-based forecasts and we expect to provide better feedback using debiased forecasts. Hence, we expect debiased forecasts to give better orientation than purely model-based forecasts.

Design options for confidence corridors and the determination of suspicious judgmental forecasts were introduced and discussed together with different feedback mechanism and procedures for long-term accuracy improvement. In cooperation with our research partner we just started a long-term project on this type of manual overwrite. We will conduct A/B-testing in order to evaluate which treatments and variants are most effective in practice, and how we can keep the attention of the forecasters continuously high. A further challenge will be the inclusion of the action chosen by the forecasters as well as their provided comments and justifications. A particularly relevant question is whether forecasters choose the (ex post) right action and, if not, whether the *ACTION_ANALYSIS* option will provide value for reflection and longer-term accuracy improvements.

References

Armstrong, J. S. and F. Collopy (1992). "Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons." *International Journal of Forecasting* 8, 69–80.

Bhandari, G., K. H., and R. Deaves (2008). "Debiasing Investors with Decision Support Systems: An Experimental Investigation." *Decision Support Systems* 46 (1), 399–410.

Blanc, S. M. and T. Setzer (2015). "Analytical Debiasing of Corporate Cash Flow Forecasts." *European Journal of Operational Research* 243 (3), 1004–1015.

Elgers, P. T., M. H. Lo, and D. Murray (1995). "Note on Adjustments to Analysts' Earnings Forecasts Based Upon Systematic Crosssectional Components of Prior-period Errors." *Management Science* 41 (8), 1392–1396.

George, J. F., K. Duffy, and M. Ahuja (2000). "Countering the Anchoring and Adjustment Bias with Decision Support Systems." *Decisions Support Systems* 29 (2), 195–206.

Goodwin, P. (1996). "Statistical Correction of Judgmental Point Forecasts and Decisions." *Omega* 24 (5), 551–559 (5).

— (1997). "Adjusting Judgemental Extrapolations using Theil's Method and Discounted Weighted Regression." *Journal of Forecasting* 16, 37–46.

— (2000a). "Correct or Combine? Mechanically integrating judgmental forecasts with statistical methods." *International Journal of Forecasting* 16, 261–275.

— (2000b). "Improving the Voluntary Integration of Statistical Forecasts and Judgment." *International Journal of Forecasting* 16 (1), 85–99.

Goodwin, P. and R. Fildes (1999). "Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy?" *Journal of Behavioral Decision Making* 12 (1), 37–53.

Graham, J. R. and C. R. Harvey (2001). "The Theory and Practice of Corporate Finance: Evidence from the Field." *Journal of Financial Economics* 60 (2), 187–243.

Hogarth, R. M. and S. Makridakis (1981). "Forecasting and Planning: An Evaluation." *Management Science* 27 (2), 115–138.

Kim, C., D. Mauer, and A. Sherman (1998). "The Determinants of Corporate Liquidity: Theory and Evidence." *Journal of Financial and Quantitative Analysis* 33 (3), 335–359.

Klassen, R. and B. Flores (2001). "Forecasting Practices of Canadian Firms: Survey Results and Comparisons." *International Journal of Production Economics* 70 (2), 163–174.

Lawrence, M., P. Goodwin, M. O'Connor, and D. Oenkal (2006). "Judgmental forecasting: A review of progress over the last 25years." *International Journal of Forecasting* 22, 493–618.

Leitner, J. and U. Leopold-Wildburger (2011). "Experiments on Forecasting Behavior with Several Sources of Information - A Review of the Literature." *European Journal of Operational Research* 213 (3), 459–469.

Lim, J. S. and M. O'Connor (1996). "Judgmental Forecasting with Interactive Forecasting Support Systems." *Decision Support Systems* 16 (4), 339–357.

McCarthy, T., D. Davis, S. Golicic, and J. Mentzer (2006). "The Evolution of Sales Forecasting Management: A 20-year Longitudinal Study of Forecasting Practices." *Journal of Forecasting* 25 (5), 303–324.

Remus, W. and J. Kottemann (1995). "Anchor-and-adjustment behaviour in a dynamic decision environment." *Decision Support Systems* 15 (1), 63–74.

Sanders, N. and K. Manrodt (2003). "The Efficacy of Using Judgmental Versus Quantitative Forecasting Methods in Practice." *Omega* 31 (6), 511–522.

Theil, H. (1966). *Applied Economic Forecasting*. North Holland Publishing Company.