Mat-2.177 Operations research project work seminar
Helsinki University of Technology

COMBINING JUDGMENTAL AND STATISTICAL FORECASTS

Final report
4/16/2003

Otso Ojanen	**48479M**
Ville Karttunen	**51976N**
Janne Karelahti	**43123P**
Maija Kaukonen	**51925V**

Abstract

This study has been conducted in Helsinki University of Technology and is commissioned by Nokia Mobile Phones. The research was done February - April in 2003. The global mobile phone market is forecasted in a way that combines statistical and judgmental forecasts. The goal of the study was to find out how the forecasts should be combined and to analyze possible improvements to Nokia's forecasting process with an emphasis on finding the best way to combine statistical and judgmental forecasts. We concentrated on the analysis of short-term forecasts. Improvements to the actual statistical model were limited out.

The research consisted of a literature review, interviews with Nokia personnel and a questionnaire together with a data-gathering sheet for the three regional market analysts (RMAs). The literature on judgmental and statistical forecasting can be divided in roughly two different categories. In the first part, two kinds of feedback were identified: mean absolute percentage errors (MAPE), and on the other hand time series data patterns (trend, seasonality or stationary data) and degree of noise (high or low). The research on integrating statistical and judgmental forecasts revealed four different approaches: model building, forecast combination, judgmental adjustment and judgmental decomposition. An Excel tool was built to facilitate comparison of different combining methods: correction of RMA forecasts (Theil), correction of Country Manager forecasts (Theil), correct both forecasts then combine (Theil), and simple weighted combination. The tool could also be used for feedback and forecast monitoring.

Finally, we recommended the use of the Excel tool and some improvements to the forecasting process based on the strengths, weaknesses and improvement that had been identified in the interviews.

Table of contents

1 Introduction

1.1 Background

Nokia is the world leader in mobile communications. Backed by its experience, innovation, user-friendliness and secure solutions, the company has become the leading supplier of mobile phones and mobile, fixed broadband and IP networks. Nokia Mobile Phones division closely follows the global mobile phone market by measuring the market size and making forecasts about future developments in different geographical markets. Both short-term and long-term forecasts are prepared for the management team by Nokia forecasting team. The process is a complex one involving sophisticated statistical models as well as expert judgments from several sources such as analysts, country managers, operators and retailers.

Sales and market forecasts are one of the most important forecasts prepared by companies. The decisions about production, distribution, procurement and investments are more or less based on these forecasts. Therefore, significant resources are reserved to provide these forecasts. In preparing the forecasts, some statistical models are conventionally used, for example SARIMAX – time series models in which seasonal variations are easy to account for. However, statistical models cannot take into account external factors that affect the accuracy of forecasts. There are several systematic, non-systematic and unexpected factors that cannot be considered, either for cost reasons or because of difficulties in measurement. Thus, the statistical forecasts are often adjusted using judgmental forecasts prepared by several sources.

The objective in making forecasts is to build a model with a high rate of explaining the market development, to optimize the process of building and acquiring expert forecasts and combine these methods in an optimal fashion such that the forecasts are as close to the realization in all market conditions as possible. The statistical model should be accurate but still as simple as possible. The forecasts obtained from experts are often already combined consensus estimates of several forecasters. The problem is how to correct or combine these forecasts to obtain the definite forecast for a given period. Usual methods include using weightings or simple averages.

Nokia makes a monthly forecast for the quarterly market volume. The whole market volume is obtained by adding together individual country forecasts. In this process, the statistical forecasts are judgmentally adjusted using information from different sources, such as country manager views and sales information from carriers (operators) and retailers. Global macroeconomic issues must also be considered. Separate forecasts are prepared for long-term forecasting but the forecasting process is for most parts the same as for short-term forecasting.

1.2 Research problem and goals

The research problem was to analyze possible improvements to Nokia's forecasting process with an emphasis on finding the best way to combine statistical and judgmental forecasts. The ultimate goal was to improve Nokia's forecasting process such that market forecasts are improved. Issues to consider in this analysis are the structure of the process, the length of the forecasting period and the effects of market characteristics. The goal was to make a statistical analysis with real data to compare different approaches and to interview the participants in the forecasting process to find improvement proposals to the general process as well.

1.3 Research methods and limitations

The research consisted of a literature review, interviews with Nokia personnel and a questionnaire together with a data gathering sheet for the three Regional Market Analysts (RMAs). The duration of the project was about three months. This limited the scope of the research as well as the possibility for gathering market data.

We concentrated on the analysis of short-term forecasts as the testing of improvement proposals and obtaining data was easier. The definite market forecasts consist of the statistical and judgmental components. Thus, both components as well as the method of combining them could be improved. However, finding improvements to the actual statistical model was out of the scope of this project.

1.4 Structure of the report

This report consists of five sections. Section 2 presents the literature review and the interviews held with the forecasting team at Nokia are discussed. Section 3 presents the actual analysis and application of the project. In section 4 we propose some concrete measures to improve the forecasting process and forecast aggregation. Section 5 concludes and finally, in section 6, a project summary is put together.

1.5 Schedule of the research

The research was conducted during beginning of February and end of April in 2003. The project was divided into four separate stages. In the first phase, the project group sharpened the objectives of the project and learned about the forecasting process at Nokia in general. The project group then proposed some directions of further research and activity for the rest of the project's duration. This first phase fixed the other three stages of the project. In the second phase, an extensive literature study was conducted on statistical forecasting in telecommunications sector, judgmental forecasting and methods to correct or combine statistical and judgmental forecasts as well as the impacts of feedback on forecasting performance. In the third phase, the project group concentrated on the overall forecasting process at Nokia, prepared questions for the forecasting team in Nokia and held interviews in which these issues were discussed. Finally, in the fourth phase, the project group gathered data for comparing combination models and built a tool for follow-up of forecast performance. To conclude, a final report of the project was written.

2 Literature review and interviews

The literature on judgmental and statistical forecasting can be divided in roughly two different categories. First, one can study whether or not different feedback methods improve the accuracy of judgmental forecasts. Second, one can study if the accuracy of the forecast could be improved by integrating statistical and judgmental forecasts in different ways. This literature review presents the most promising methods of feedback and integration that could possibly improve the accuracy of market forecasts in Nokia. In the first part of the review, the role of feedback in judgmental forecasting process is analyzed. The conclusions of the reviewed articles are based mostly on laboratory experiments and hence the applicability of the methods in real world may be questionable. In the second part of the review, different integration methods will be presented. Pros and cons of the presented feedback and integration methods will be summarized in the end of the review.

2.1 The influence of feedback in judgmental forecasting

An obvious strategy to improve forecasting results is the provision of feedback in the hope that it will facilitate learning and thus lead to more accurate results. The degree of improvement in accuracy depends on the type of feedback. According to Bolger and Önkal (Harries, 1999), nature of feedback (public, private, positive or negative) does not seem to have any effect on accuracy.

Types of feedback can be classified into two categories. Balzer *et al.* distinguish feedback into outcome feedback and cognitive feedback (Balzer *et al.*, 1989). Outcome feedback (OCF) is the simplest form of feedback. It just gives actual outcome and indicates how accurate a given prediction was. Balzer *et al.* divide cognitive feedback into three subcategories: task information (TI), cognitive information (CI) and functional validity information (FVI). Task information consists of information about the relations in the environment. Cognitive information describes relations perceived by the forecaster and functional validity information includes forecaster's perceptions of the environment. Remus *et al.* give examples of task information ("Now the series is flat with random noise.") and cognitive information ("You are overreacting to the random variation in the data") (Remus *et al.*, 1996). Benson and Önkal consider these four types of feedback

(OCF, TI, CI and FVI) to be relevant in judgmental forecasting tasks although very little research is concerned with functional validity information (Benson & Önkal, 1992).

There have been several studies that have focused to find out which type of feedback may improve forecasting accuracy. The studies are partly contradicting but some key trends can be sketched out that will be described in more detail in the following. First, the relevance of outcome feedback will be discussed. Does it really has an affect on forecasting accuracy and what are the main implications? After that the effects of cognitive feedback will be considered and, according to studies, how it should be used to get improvements in the accuracy of forecasts. Finally, we will consider the major weaknesses or uncertainties of results in terms of real world implementation.

2.1.1 Outcome feedback

Outcome feedback is the simplest form of feedback that consists of information about the realization of a forecasted event. Generally outcome feedback is considered inferior to cognitive feedback. Brehmer even argues that people lack cognitive schemata needed for improving forecasting results on the basis of simple outcome feedback (Brehmer, 1980). However, a general trend in outcome feedback studies is that some improvement in accuracy is possible if outcome feedback is capable of offering the right kind of information. Todd and Hammond stated already in the 1960's that learning in multiple-cue probability tasks requires information about cue-criterion relationships. Outcome feedback can, indeed, offer this kind of information (Todd & Hammond, 1965).

Fischer and Harvey made a research where they studied the effects of outcome feedback. In their experiment, a group of subjects was required to combine four individual forecasts with different accuracy levels to a final forecast (Fischer & Harvey, 1999). Fischer and Harvey came into the conclusion that provision of outcome feedback affects forecasting accuracy. Fischer and Harvey state that outcome feedback improves forecasting accuracy rapidly but judgmental forecasts can still not beat simple average as Figure 1 indicates. Actually, Fischer and Harvey show that improvements in performance reach asymptotically similar levels to simple average.

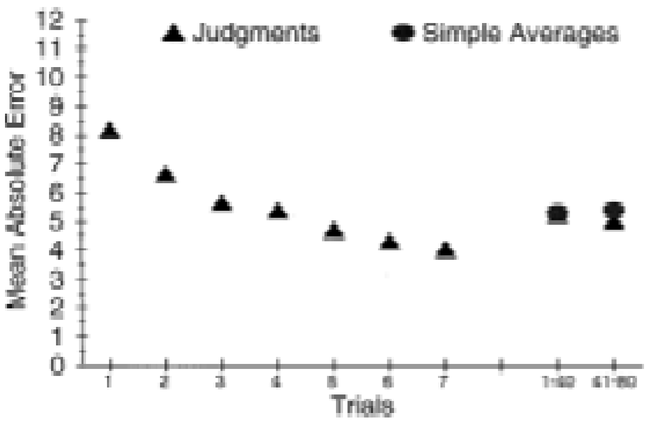

Figure 1: Accuracy improvement in outcome feedback experiment

According to Fischer and Harvey, noteworthy in their results is that even though there was no difference between the accuracy of simple average and judgmental forecasts with outcome feedback the methods were inaccurate for different reasons (Fischer & Harvey, 1999). The averages are erroneous because of an inappropriate weighting policy that is consistently applied in every trial. Therefore, no random noise is incorporated into them. In contrast, subjects have used more appropriate weighting (i.e. they have tried to give more weight on more accurate initial forecasts) in the judgmental forecasts but the inaccuracy rises from inconsistently applied weighting policy. That is why judgmental forecasts incorporate also a random element.

Further improvement in judgmental forecasts would be available, if random element is eliminated or reduced. Fischer and Harvey suggest that this can be done by using a regression model of the judge's behavior. Coefficients of the model could be isolated from the error component and then used to weight the separate forecasts. This technique is also known as bootstrapping and will be described later in more detail.

2.1.2 Cognitive feedback

Cognitive feedback is more highly processed information than simple outcome feedback. General opinion seems to be that cognitive feedback is also more effective than outcome feedback (Balzer *et al.* 1994; Remus *et* al, 1996) and therefore different types of cognitive feedback (especially cognitive information and task information) have been under intensive research.

The most significant form of cognitive feedback seems to be task information. Both Balzer *et al.* and Remus *et al.* have concluded that task information itself can improve accuracy of forecasts but adding cognitive information did not give any additional improvement. (Balzer *et al*, 1989; Remus *et al.*, 1996). In the research of Remus *et al.*, the task information consisted only of information on time series patterns (flat, sloping upward n units per period, sloping downward n units per period). Correspondingly, the cognitive information was expressed in the following manner: "You are overreacting to the random variation in the data", "now your forecasts are too low; you are failing to keep up with the upward sloping series" etc. Remus *et al.* suggest as a practical implication that task properties information should be provided to forecasters. This is possible with help of statistical programs or spreadsheet macros.

Fischer and Harvey have also studied impact of cognitive information combined with functional validity information and stated that provision of this combination of feedback enables judges to outperform the simple average (Fischer & Harvey, 1999). In the Fischer and Harvey's experiment subjects were asked to combine final forecast from four initial forecasts with different accuracy levels. They noticed that providing information on mean absolute percentage errors (MAPE) of initial forecasts calculated over all trials up to the current one outperformed mechanical simple average. The method was also more efficient than simple outcome feedback or information on errors produced by four initial forecasters.

Providing information on signed errors in the four sources in the Fischer and Harvey's experiment resulted to inferior results, indeed. This may be interpreted that not all information is for good in judgmental forecast tasks. Fischer and Harvey suggest that subjects may have devoted their attention mostly to the signs instead of the sizes of errors. The failure to integrate information about the sizes of errors prevented learning and thus lead to poor forecasts.

The implications of Fischer and Harvey's research are the following. Judges can combine individual forecasts in a way that outperforms simple averaging if information about the past performance of those individual forecasters is explicitly available (recommended form is according to Fischer and Harvey MAPE). Fischer and Harvey emphasize especially explicit form of this information because their studies clearly show that memory of a person cannot serve for the same purpose.

Sanders concentrated only on task properties information and showed that providing information on time series data patterns (trend, seasonality or stationary data) and degree of noise (high or low) to subjects leads to significant improvement in forecasting accuracy (Sanders, 1997). Sanders also states that improvements in forecasting accuracy has more to do with ability of judges to discount random noise than to predict specific time series patterns. Series noise level seems to be critical factor and information on patterns only an enhancing factor.

The practical implication according to Sanders's research is that all judgmental forecasting processes should be performed with help of information on time series data patterns and level of noise. This should be easy to implement because majority of statistical computer programs can be used to periodically analyze required data from past time series.

2.1.3 External validity of feedback studies

As noted earlier, a vast majority of studies related to judgmental forecasting are made as laboratory experiments. These special conditions should be taken into account when reviewing how well results correspond to real world. Goodwin and Wright summarize several problems that may occur in experiments (Goodwin & Wright, 1993). Only the most significant sources of error are discussed here. First, cues in experiments are pre-selected by a researcher, which may imply to subjects that they should use those cues. On the other hand, subjects may not be able to take full advantage of the contextual information, which is offered to them. The key implication of this is, as Sanders noted, that practitioners using the information should be educated enough to understand the information provided (Sanders, 1997).

Goodwin and Wright also note that the information in laboratory experiments is, in general, perfectly reliable (Goodwin & Wright, 1993). In laboratory studies it is almost impossible to assess the impact of the organizational and political factors that may affect judgmental decisions. Also, the forecaster is not likely to have any preference for certain outcome before making the forecast. Besides, the forecasting may take place in a dynamic system where forecaster itself affects the environment and where relative importance of separate cues may vary with time. In such a dynamic environment underlying rules of the system may change during the learning process and thus hinder actual learning. All these problems are at least partly recognized also in Nokia forecasting process.

Another practical source of error may be anchoring and adjustment heuristics. Tversky and Kahneman showed in their pioneering study that predictions are subject to anchoring and adjustment heuristics (Tversky & Kahneman, 1974). Harries claims that anchor effects are not eliminated by financial incentives and even highly professional experts are biased toward arbitrarily anchored values (Harries, 1999). This should also be taken into account when planning efficient feedback system. Lim and O'Connor and Fischer and Harvey also state that after people have made their initial predictions they are unwilling to take advice and alter their predictions. Feedback cannot affect this unwillingness (Lim & O’Connor, 1995; Fischer & Harvey, 1999).

2.2 Integrating statistical and judgmental forecasts

At least four different approaches for integrating statistical and judgmental forecasts exist: model building, forecast combination, judgmental adjustment and judgmental decomposition (Webby & O'Connor, 1996). The approaches are not mutually exclusive but interact with each other. For example, model building must be done before forecast combination and judgmental adjustment. The following subsections briefly review the literature in each category. However, the second category will be reviewed more deeply since it appears to be the most promising approach in order to achieve the objectives of this project. The four different approaches for integrating the forecasts are presented schematically in Figure 2.

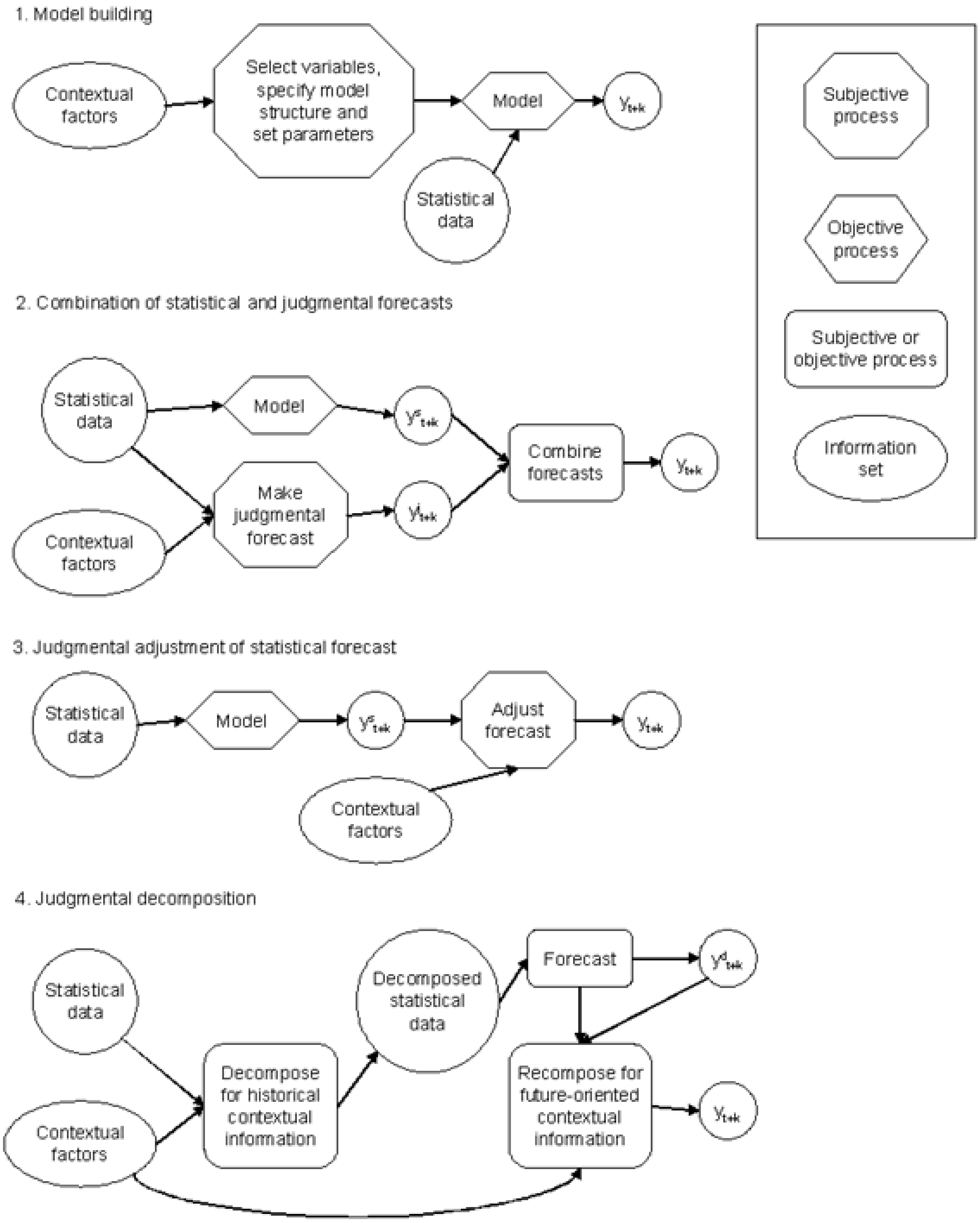

Figure 2: Integrating judgmental and statistical forecasts (Webby & O'Connor, 1996)

This section considers the aspects in which judgment is important in model building. In model building, the judgment is used at least in variable selection, model specification, parameter estimation and data analysis (Bunn & Wright, 1991). Where model specification and parameter estimation can be done using techniques such as bootstrapping and multiple regression, judgment is needed in identifying the variables. If the relationship between the variables is linear, bootstrapping appears to be successful.

2.2.1 Combination of statistical and judgmental forecasts

It is widely recognized that combining two or more independent forecasts result in improvements in accuracy (Clemen, 1989; Goodwin, 2000). Especially if one chooses to combine statistical and judgmental forecasts, the resulting forecast will likely be more accurate than the original forecasts. What comes to forecasting difficulty, combination seems to be more effective for the series with low mean average percentage error (MAPE), that is, for 'easy' series. If either of the forecasts to be combined is presumably more accurate than the other, the forecasts should not be combined. It is evident that if one combines a very good forecast with a poor one, the resulting forecast will be less accurate than the original good forecast (Sander & Ritzman, 1990).

Seasonality does not seem to improve the benefits gained from combination (Lawrence et al, 1986). It is also worthwhile to notice that combination appears to improve accuracy especially in the short run. That is, combination should be used primarily when the forecast horizon is relatively short (Conroy & Harris, 1987).

It seems that the greater the number of forecasts to be combined, the better the accuracy of the resulting forecast will be. Also, mechanical combination of the forecasts appears to produce better forecasts than the subjective approach (Lawrence et al, 1986). It is however difficult to determine what weights should be attached for each forecast. It has been demonstrated that judges give too great weights to their own forecasts and underestimate the importance of statistical forecasts despite the better accuracy of the latter ones (Goodwin, 1996). It seems that taking a simple average produces often the best outcome. Especially if the past data is scarce, it is effectively impossible to estimate the optimal weights. Even if the past data is available, the forecasts should be unbiased and forecast errors should form a stationary series over time. The situation is very unlikely to occur in the real world (Goodwin, 2001).

Goodwin also points out that the greater the correlation between the forecast errors, the smaller the value of the combination procedure will be because the second forecast brings little new information to the combination. Ideally, the correlation between the forecast errors should be strongly negative. However, in practice this seldom happens (Goodwin, 2002).

2.2.2 Bootstrapping

One way to improve the accuracy of the judgmental forecasts is to simply use multiple linear regression to build a model of a judge's forecasts. The generic bootstrapping model can be stated as

$$F_t = a + b_1 x_{1,t} + b_2 x_{2,t} + \ldots + b_n x_{n,t}, \qquad (1)$$

where F_t is the judge's forecast for period t, $x_{i,t}$ are the values of the explaining variables or cues available to the judge at t, b_i is the weight attached to cue i and a is a constant. In order to be able to use this approach, one has to of course know the cues the judge is using when constructing her forecast. For example, if it is known that a judge is using advertising expenditure and competitor's price to forecast sales, the weights that the judge attaches to the variables can be estimated by a multiple linear regression. After the weights have been estimated, they can be used to compute a corrected judgmental forecast for the particular period (Goodwin, 2002).

In general, the bootstrapping technique has performed fairly well in many areas such as bankruptcy prediction and the prediction of student performance. The models appear to outperform the judges because they average-out the inconsistencies of human judgment. However, it is not sure if the technique is suitable for judgmental time-series forecasting. For example, the large number of cues available for the judge makes it difficult to identify the cues for the model. Also, one possibly important source of information for the judge - contextual information - cannot be included in the model (Goodwin, 2002).

2.2.3 Correction for bias

If it is likely that judgmental forecasts are systematically biased or make inefficient use of available information, they can be corrected by at least two different methods. According to Theil, the mean squared error (MSE) of a set of forecasts can be decomposed as

$$MSE = (\overline{Y} - \overline{F})^2 + (S_F - rS_Y)^2 + (1 - r^2)S_Y^2, \qquad (2)$$

where $\overline{Y}$ and $\overline{F}$ are the means of the outcomes and judgmental point forecasts, S_Y and S_F are the standard deviations of the outcomes and judgmental forecasts and r is the correlation between the judgmental point forecasts and outcomes (Theil, 1971).

The first term of the decomposition represents mean bias. This is the tendency of the forecasts to be too high or too low. The second term of the decomposition represents regression bias. The term expresses how well the forecasts track the actual observations. For example, the forecasts may be too high when the actual observations are low and vice versa. The third term of the decomposition represents random errors in the forecasts.

By using an optimal linear correction proposed by Theil, the mean and regression bias can be eliminated from a set of forecasts for which the outcomes are known. The model can be written as follows

$$Y_t = a + bF_t + i_t, \tag{3}$$

where Y_t is the actual observation, F_t is the original judgmental point forecast and i_t is the residual for period t. The estimated parameters $\hat{a}$ and $\hat{b}$ obtained from the regression can then be used to correct the original judgmental point forecast as follows

$$P_t = \hat{a} + \hat{b}F_t, \tag{4}$$

where P_t is the corrected judgmental point forecast for period t.

Another correction method proposed by Fildes can be applied, if the information on the cues used by the forecaster are available (Fildes, 1991). In this model, the forecaster's errors are regressed on to the cues as follows.

$$e_t = a + b_1 x_{1,t} + b_2 x_{2,t} + \ldots + b_n x_{n,t} + i_t, \tag{5}$$

where e_t is the judge's forecast error for period t, $x_{i,t}$ are the values of the cues for period t, i_t is the residual for period t and a is a constant. In this method, the forthcoming judgmental forecasts are corrected by the predicted error. Fildes's method reduces essentially the effect of the third term in the decomposed MSE. Goodwin points out, that in this light it might be worthwhile to combine the two different correction approaches into a united model of the form

$$e_t = a + b_0 F_t + b_1 x_{1,t} + b_2 x_{2,t} + \ldots + b_n x_{n,t} + i_t. \qquad (6)$$

Naturally, the models presented above work ideally only if the residuals for the subsequent periods are uncorrelated. For example, the forecasts accuracy may improve over time resulting in a non-constant bias. If the biases change gradually, one can use discounted weighted regression in combination with Theil's method proposed by Goodwin (Goodwin, 1996). This method effectively puts more weight on the performance of the recent forecasts when estimating the parameters. However, if the biases change sporadically, combined Theil's and Fildes' method may be the most appropriate tool (Goodwin, 2002).

2.2.4 Correcting and combining

Since the judgmental forecasts tend to suffer from biases, it may be reasonable to correct them before combining with the statistical forecasts, the so called correct-then-combine strategy (Goodwin, 2000b). Still, Goodwin points out, that correcting the judgmental forecasts by Theil's method and then combining them with statistical forecasts might not be as good an idea as it may sound at first. Firstly, correction might be so successful that combination will not improve the corrected forecast at all. On the contrary, it may even worsen the accuracy of the forecast. Secondly, potential benefits of the combination may be reduced by the smoothing effect of Theil's method on the judgmental forecasts. The effect increases the correlation of the judgmental forecasts errors with those of the statistical ones.

The results obtained from the laboratory experiment and in the two real world cases (a European textile company and a UK-based engineering company) by Goodwin and Fildes suggest that the correction method is the most accurate from the set of strategies presented above. However, the results can be hardly generalized to apply in all the industries. Also, the sample sizes in their experiments were quite small which may have underestimated the effectiveness of the correct-then-combine strategy (Goodwin, 2000b). It is therefore important to test the accuracy of the alternative approaches in the particular industry before choosing the final model to be used.

2.2.5 Judgmental adjustment of statistical forecasts

In judgmental adjustment of statistical forecasts, also known as voluntary integration by Goodwin, the statistical forecast is given for the judge who can then decide how to use this information in composing the final forecast. Hence, it is possible for the judge to completely or partially ignore or accept the statistical forecast. Voluntary integration often means only judgmentally adjusting statistical forecasts, but it could also mean modifying prior judgmental forecasts based on newly arrived statistical forecasts (Goodwin, 2002).

When only time-series information is provided for the judgmental forecaster and the statistical model, judgmental adjustments of the statistical forecasts tend to reduce the accuracy due to anchoring and adjustment heuristics. Also, the initial accuracy of the statistical forecast affects to the effectiveness of adjustment. On the other hand, if the judgmental forecaster can exploit contextual information (e.g. company's products will be taxed more heavily in the future), judgmental adjustment of statistical forecasts can improve the accuracy of statistical forecasts.

Still, there appears to be a number of drawbacks in this integration approach. First, the danger of double counting bias arises when the forecast is produced by a regression model from which a variable has been omitted. If there is collinearity between the omitted variable and a variable in a model, the latter variable acts like a proxy for the omitted variable. Hence, the effects of the omitted variable will be taken into account by the proxy and judgmental adjustments may actually double-count some of the effects.

Second, when the statistical forecast is reliable for the part of the time-series pattern, judges may ignore the statistical forecast completely. As an example, a statistical forecast may perform very well for the normal patterns in a time-series excluding the effects of special events. Rather than adjusting the statistical forecast for the effects of special event, judges tend to produce the forecast completely by judgment.

Third, the opportunity to learn and improve the judgmental adjustment process diminishes because the adjustments are often made on an ad hoc basis without proper documentation and rationale (Goodwin, 2002).

According to Goodwin, the accuracy of the adjustment process can be improved by making it more difficult for the judge to alter the statistical forecast. In a laboratory experiment Goodwin noticed that when the judges had to explicitly make a request to change the statistical forecast, the number of harmful adjustments reduced without reducing the tendency to make adjustments when appropriate. Moreover, if the judges were required to indicate a reason for adjusting the statistical forecast, the number of harmful adjustments were reduced further (Goodwin, 2000a). It is of course difficult to say how well the results obtained in a laboratory environment apply in the real world.

2.2.6 Judgmental decomposition

This approach is a three-level process in which the judgmental and statistical methods may operate at any of these levels. First, the time-series is decomposed for any historical data after which the forecast is generated from the decomposed time-series. Finally, the forecast is recomposed with any future-oriented information. The difference between judgmental adjustment and judgmental decomposition is that while the first one tries to model misspecifications in the forecast after the statistical forecast is made, the latter one tries to remove the effects of past contextual factors from the series history before accounting for future-oriented factors in the forecasts (Webby & O'Connor, 1996).

As a clarifying example, consider the case where the initial statistical model is built according to some specific fundaments. In the first phase, these fundaments are sorted out and analyzed individually. Now, the judge may have an opinion that the weight of a certain fundament in the outcome has changed or some fundament has become obsolete. Consequently, the new model will be generated according to the modified fundaments. The final forecast can now be produced using the updated model.

Decomposition has appeared to be effective when judge’s knowledge about the topic is relatively scarce. On the other hand, Goodwin and Wright point out that decomposition would be ineffective if the decomposition is mechanical and the judge is skeptical about the decomposition technique being applied. Also, the approach is ineffective if the judge is unfamiliar with the technique used for decomposition or the duration of the task is too long (Goodwin & Wright, 1993).

2.3 Interviews

We conducted three interviews. The persons interviewed were the three regional market analysts of the Nokia market analysis team. The following points arose from the interviews.

The role of the regional market analyst (RMA) has many sides. In the short run they are responsible for coordinating the top-down modeling process and gathering the country inputs. This consists of sending the initial forecasts to the countries and gathering and analyzing their results including also communication with the countries. Sometimes RMAs also update the statistical model. In the long run, they make their own assumptions and manage modeling.

The statistical forecast is sent to the country responsible more then a week earlier and any comments on that are requested. Normally 1-3 persons form the country adjustments to the forecast. The country may want to change the initial statistical forecast e.g. if there has been some change in regulations. The RMA's duty is then to consider whether the adjustments are realistic. If there are irrational changes to previous year's realization, seasonal variation and growth percentages are checked.

The statistical and judgmental forecasts are combined according to RMA's best knowledge using statistical adjustment. They may be weighted differently depending on how reliable the statistical forecast is seen. In general, there is no research on how to weight the available forecasts. A quartile stage comparison should be carried out in order to make sure the trend is reasonable and the forecasts are broadly equal. After the combination the result is presented to the regional management board and afterwards sent to the global team.

In the long term the assumptions, the parameters in the statistical model, can be changed if it is deemed necessary. The model updating process as a whole takes a lot of time and in the short term the assumptions are not easily changed.

The most important factors taken into account when combining the forecast are:

- Reasonability of the suggested judgmental forecast
- Profile of the forecaster

- History trend
- Bottom-up information concerning events on the market
- Long term trends
- Additional information received

There are several factors and situations, which make the forecasting hard. The main ones are:

- Uncertainty about the macro-economy
- Other macro level issues e.g. war or natural disasters
- Changes in regulations
- Factors difficult or impossible to model
- Challenges of modeling replacement

Even though the forecasting success is not followed at a personal level, some errors and biases could possibly be identified. Some forecasters may seem to be optimistic; some on the other hand are conservative. It is often assumed that in the short term people usually overestimate and in the long term underestimate changes. RMAs' errors can be caused by for example lack of visibility through the markets or sticking too eagerly in some trend even though it cannot be observed properly. The model also comprises parameters that are not validated often which may produce systematic errors.

The RMAs receive feedback from the global team in which the forecast, the realizations and the corresponding error are indicated. This is sent also to those responsible for country forecasts.

According to the RMAs, the *strengths* of the process include:

- Competency of the model
- Usefulness of the process and use of results
- Commitment of the management
- Networking of the people related to the process

- Combining the statistical forecasts and the views from the markets and regional management
- Storage of history data

The process has also its *weaknesses*. The main ones are:

- Complexity of the replacement modeling
- Regionally non-standardized process
- Lack of clear dead-lines on regional level
- No regional differences
- Too generic model
- Model's inability to react to changing situations
- Conversion to new type of markets.
- Difficulties in observing correlations

When conducting the interviews the following *improvement suggestions* were identified:

- Increase of visibility in the process
- Standardization across the regions, a checklist for the process steps
- Clarifications about the model parameters
- Back up the accuracy of the measures with actual data
- Continuously communicate the value added by the forecasting process to the company
- Provide enough resources (time, personnel) to every area to ensure learning through feedback and quality of forecast reviews
- Improving communications to the countries
- Clear dead-lines
- Model working in different markets
- One-to-one meetings in which assumptions and numbers are openly discussed

The practical meaning of these improvement suggestions should be analyzed more thoroughly when deciding which improvements to the process are actually implemented.

3 Analysis and application of results

The analysis of obtained results consists basically of two distinct parts. First, we analyze the feedback in Nokia market forecasting. Second, we present a tool for comparing the presented methods for combining statistical and judgmental forecasts and managing forecasts and forecast feedback. The tool is built in Excel which is a natural environment for this kind of analysis tool.

3.1 Feedback in Nokia market forecasting

At the moment in Nokia, the global team delivers outcome feedback. It consists of the forecast, the realization and the error. Yet, the feedback can be totally ignored and the progress of the forecasts or forecasters is not followed consistently. While most of the experiments concerning feedback's influence to forecast accuracy are performed in laboratory, nothing can be said for sure about their effectiveness. At least Nokia's dynamic business environment will hinder learning process for sure. Yet, we would recommend the following.

Task information should be provided to the forecasters. One possibility is to provide only task information on time series data patterns (trend, seasonality or stationary data) and degree of noise (high or low). As mentioned earlier in the study, these could be produced with statistical computer programs or Excel. Another way is to provide cognitive information e.g. mean absolute percentage errors (MAPE) of initial forecasts. The signs of the errors are not important, on contrary it has been stated that attention should be paid to the sign instead of the size of the error. The information about the past performance needs to be explicitly available, that is, either in tabular or graphical form. A recommended measure would be mean absolute percentage errors (MAPE).

The global team should produce the information in the same way as the general feedback currently. The feedback should be pointed out to RMAs and country managers so that everyone would be able to follow the development of their own forecasts. The timeframe followed could be e.g. eight quartiles. The Excel tool developed in this project could be used as a starting point for forecast feedback analysis. The Excel tool containing sheets of the relevant countries could be provided to forecast process participants. To the other direction, from forecasters to the market analysis team, motivations and explications about forecasted values should be provided. This could also be included in the Excel tool as discussed previously.

If the resources are sufficient, Nokia could try to provide information both about time series patterns and MAPE in different market areas and compare the results and experiences. It is also possible to choose only other of the methods or to provide the forecasters with both. The decision should be based on such issues as need for resources, suitability and effect on forecasting accuracy.

Whatever information is provided, it is highly important to ensure that people understand the information and offer enough education, as it is needed. To minimize errors or biases caused by heuristics and other external factors mentioned in the literature review, the sources of should be taken into account, their effects should be explained to the forecasters and any means to reduce strength of the factors should be discussed together. Also forecasters should be offered incentives to strive for correct predictions, for instance, a bonus salary system.

3.2 Choosing an appropriate forecast combination method for Nokia

We presented four different model concepts in the literature review. The process in use at Nokia uses the judgmental adjustment of statistical forecasts approach. The most important flaws of this method were presented in the review. This being the starting point we concentrated first on finding alternative methods to integrate the different forecasts. The literature on combining forecasts showed that, as expected, there is no single method that would perform exceptionally well in combining forecasts.

Based on the literature review it seems that simple averaging is hard to beat in combination. If data about the forecasts were available, this assumption could have been tested. However, now this analysis remains the task of Nokia forecasting team. We believe that the analysis is worthwhile despite the discouraging results in the literature.

The combination methods to try out at Nokia were chosen based on the literature review. We chose to implement test for the following methods:

1. Correction of RMA forecasts (Theil)
2. Correction of Country Manager forecasts (Theil)

3. Correct both forecasts then combine (Theil)
4. Simple weighted combination

3.3 Applying the model

As the risk of not obtaining enough data to validate and test different models was realized, we decided to implement a tool for Nokia to follow the performance of these methods as more data is gathered, instead of a data analysis. Furthermore, the tool could be used for performance monitoring and feedback purposes. It is for example easy to analyze relative accuracies across countries, country managers, statistical models and RMAs. Furthermore, for historical reference the tool can be used to save comments about decisions made regarding forecasts. In each Excel cell, it is possible to save a comment. Possibly some sort of color coding could also be used in cell background, to highlight forecasts that have some particular characteristics, such as larger uncertainty, missing data, or unusual market or country conditions. Also, standard Excel reporting and visualization functionality is available for graphical illustration or reporting to other systems. A screen-shot of the tool is presented in Figure 3.

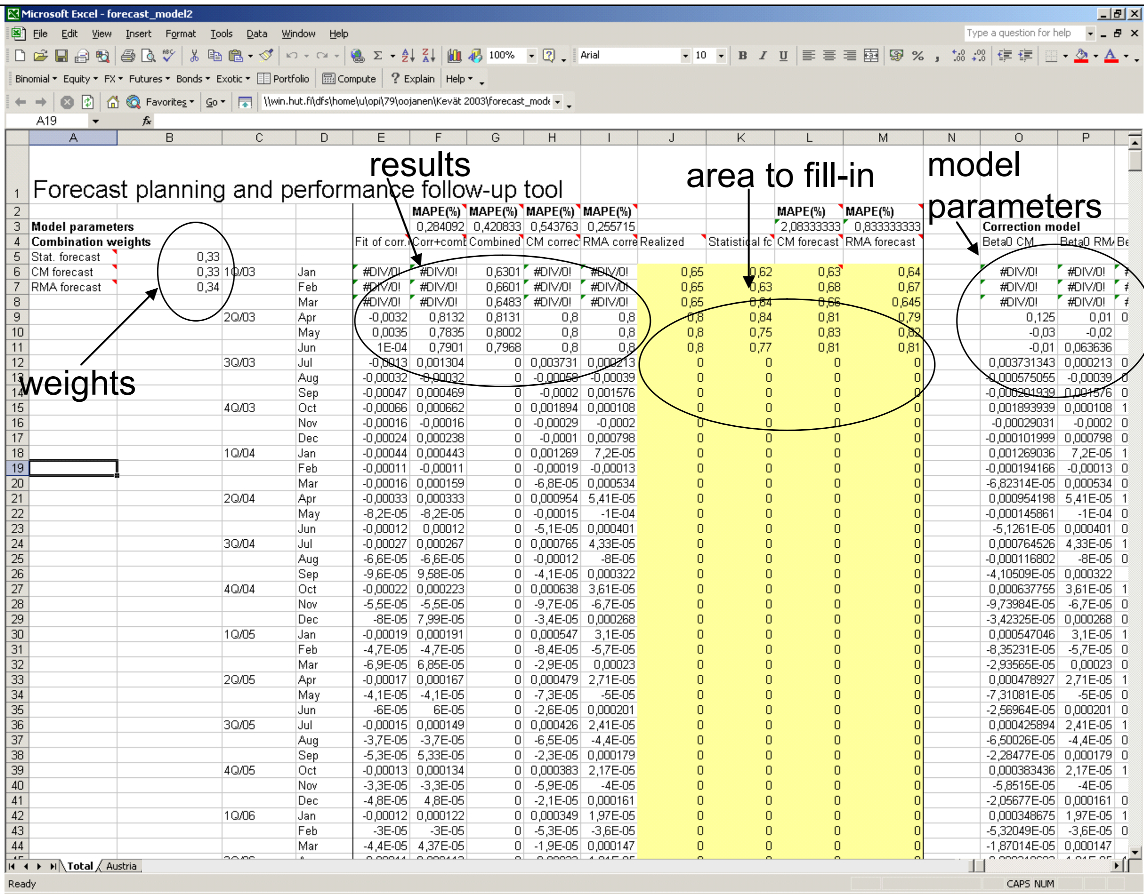

Figure 3: Forecast planning and performance follow-up tool in Excel

The basic functioning of the tool is such that there is one sheet for each country because there must be a separate model for each country due to their different characteristics, including forecast variance. RMA fills in statistical and own forecasts for each country and country manager forecasts are filled for the corresponding countries. The cells to fill-in are marked with yellow color. The realizations should be updated as they are obtained. Currently there is possible uncertainty in the value of the realization is not included in the model. The realizations are assumed perfectly accurate. The tool offers four forecasts using different methods for each round. The tool includes all previous information to compile the latest forecast. This means that in Theil's 'correction' and 'correct then combine' methods, the regression model for correction is calculated using all previous data. However, the previous regression parameters are preserved for later review. Also, a forecast error for the correct then combine method is calculated. It is easy to include similar columns for other forecasting methods as well.

The forecasting process is based on quarterly forecasts that are prepared monthly. This presents a challenge for the modeling process because the forecast updating period differs from the actual forecasting period. We solved the problem in this Excel tool so that separate models are prepared for the first, second and third months of each quarter. This is well justified because the forecasts tend to sharpen during the on-going quarter. However, this makes the gathering of data even slower as only one observation is obtained each quarter. Thus, probably at least two years of history data is needed before the correction method provides even moderate estimates about possible need for correction.

The tool provides the mean absolute percentage error values for all forecasts and combination or correction methods. This value is updated each period and uses all information available up to that period. The model could be also be improved to calculate and store this information for each period. The MAPE information could be used to follow the development of forecasting accuracy in time.

The tool is designed so that previous model results are not updated as more data is gathered. This allows an *ex post* analysis of the obtained model results and also facilitates for example experimentation of different weights in combining the forecasts. Also, the model parameters for the Theil correction method are saved for each round which makes it easy to analyze the need for correction and the development of the correction parameters afterwards. For example, it is easy to find out whether there are systematic biases in country or RMA forecasts.

The tool is such that it is relatively easy to use once the terminology and model background in the tool have been explained. Thus, the tool could be used also as a basis for providing feedback about the success of individual forecast makers. The tool could also be further developed, for example to allow for more forecast makers and possibly new combination or correction methods.

4 Recommended improvements and implementation proposals

4.1 Proposed improvements to combining statistical and judgmental forecasts

To sum up the previous sections we present here the concrete improvement proposals in the forecasting process. However, it must be noted that the time limit on the project was tight to obtain a holistic view of the process and its true bottlenecks. The following items seemed to be the most important issues to pay attention:

1. Gather data about forecast errors and test whether statistical or judgmental forecasts are more accurate on average. This knowledge could be applied in defining weights for the forecast combination procedure instead of simple averaging.

2. Test the optimal combination weights using an *ex post* analysis

3. The variables that RMAs and country managers use in the forming their forecasts process should be identified. This would permit the use of bootstrapping methods or combined Theil and bootstrapping. As noted in the literature review, bootstrapping has performed well in real world applications.

4. The judgmental adjustment process should be made more controlled. This could be introduced by making it compulsory to supply explicit reasons and motivation if the statistical forecast is modified by country management. The research has proved this to improve forecasting accuracy significantly. These motivations could be integrated to the Excel model and modified forecasts could be marked with color codes to follow modifications and possibly even their size (color codes for different sizes of relative changes in forecasts relative to the statistical forecast).

5. In literature, the integration has been shown to be useful only in short-term forecasting. Therefore we recommend a careful analysis before using combination for making forecasts for more than six to nine months.

6. Model revision
   The statistical model in itself received some critique and it should possibly be revised.

The above-mentioned improvement areas can be implemented on the current process and model. Yet, the statistical model itself should be further improved. This would however necessitate a new project including e.g. workshops with regional managers and country managers. The model parameters should also be clarified. Especially the replacement modeling should be improved by modeling more than one lifecycle type. This is particulary important in maturing markets. The model should also be able to take regional differences and different types of markets better into account.

4.2 Improvements to the general forecasting process

According to the interviews, it can be stated that the process in general works well and available results are benefited from. We see no need to propose changes to the structure of the process. Yet, some weaknesses and improvement ideas have been detected. Our suggestions are presented below.

1. Standardization of the process

Standardization should be carried out across regions. The ideal process needs to be documented and discussed with countries. For example, scheduled checklists can be used to make sure every necessary step is taken on time.

2. Clear deadlines

Would make it easier to people to work and especially to form the final forecasts.

3. Increase of process visibility

The team members should be brought together every now and then to discuss the important issues and to help the communications inside the process. Especially the communications to countries should be made explicit. This should also include communication and clarification on what is the role of the global team.

4. The feedback

Should be included in to the process. This would allow backing up the accuracy of the measures with actual data and learning from previous forecasting mistakes. This process should involve all participants that prepare forecasts. As mentioned in the feedback partition, the feedback should be delivered in a form of patterns and noise or/and for example MAPE. Errors or biases caused by heuristics and other external factors should be tried to minimize in the ways presented earlier. One-to-one meeting between RMA and a country manager can also be seen as a recommended form of a feedback. This way the previous forecasts could be reviewed and the functioning of the model could be discussed. This way a important feedback link from the countries to the region management could be formed.

5. Check-up of resources

To reconsider whether the resources are sufficient and they are placed efficiently. E.g. implementing the feedback system would most likely require some additional workforce. Time should also be reserved for the analysis and review of forecasting feedback.

6. Incentives to support accurate forecasts

If possible, a bonus system should be constructed in a way that forecasters on country level would be motivated to forecast as accurately as possible. The implementation of this improvement suggestion is however totally dependent on Nokia's reward systems and prioritization.

5 Conclusions

The Nokia forecasting process is relatively complex and despite the complexity it is well designed and functional. One of the biggest questions was how to combine statistical and expert forecasts. Data for a proper analysis of proposed methods was not available so we settled for providing a tool that enables the forecasting team to benchmark the proposed and possibly other methods for itself. This is a more practical approach since the building of a sufficient database may take several years. However, we believe it should be possible to identify some general trends in the correction and combination methods already after about one year.

In addition to improving the combination of forecasts, we included an analysis of the general forecasting project to the project scope. As a result of this analysis some general improvement proposals were made. In addition, some proposals came up already in the interviews of the Regional Market Analysts. According to the interviews, the forecasting process in general works well. The following seven improvements were suggested: 1. Standardization of the process, 2. clear deadlines, 3. increase of process visibility, 4. feedback, 5. check-up of resources, 6. incentives to support accurate forecasts, and 7. model revision.

In the literature review four feasible alternative methods were identified to integrate judgmental adjustments with statistical model. It was shown in the review that none of the three methods had been demonstrated to be clearly superior to others. It was also recognized during the project that it would not be possible to obtain enough data to validate different models. Therefore, it was suggested that Nokia would continue this project internally and collect required data to test if there is any statistical difference between different methods.

The chosen methods were 1. Correction of RMA forecasts, 2. Correction of CM forecasts and, finally, 3. Simply combining forecasts 4. Correcting both forecasts then combining them. In every model the correction was recommended to be implemented using Theil's method presented in the literature review.

An easy-to-use Excel model was also implemented to carry out data gathering and testing phases. There is a separate sheet for each country in the tool because each country has to be analyzed as an individual system. The tool indicates forecast errors and mean average percentage error (MAPE) for each integration method. The tool is designed so that previous results are saved and thus a later analysis is possible. Experimentation of different weightings or breakdown of model parameters of Theil's method to identify systematic biases is convenient. The Excel tool is based on quarterly forecasts that are prepared monthly. Alas, probably at least two years of history data is desirable before the correction method is able to provide even moderate predictions about possible need for correction.

The literature review dealt also the importance of feedback as a straightforward way to enhance learning and hence improve judges' forecasting accuracy. Although several drags on judges' learning process were identified, the importance of this kind of out-of-box thinking in terms of improving integration of judgmental and statistical forecasts should be further emphasized.

It was shown that simple outcome feedback was not enough to achieve above average improvements in forecasting process. Nonetheless, outcome feedback with help of statistical correction, for example bootstrapping, can offer significant improvements in the forecasting precision. In addition to outcome feedback, the provision of more elaborated feedback – information on time series properties and level of noise – is recognized to be an efficient way to enhance forecasting accuracy. Almost any statistical computer program could be used to calculate and provide this kind of information. As indicated by the literature review, also the explicit form of feedback is necessary. Data and past performance of judges should be available either in graphical or tabular form. The Excel tool can also be used for this purpose.

5.1 Reliability and validity

The proposals are based on interviews with a small number of people and their opinions. The project group had only a few months to conduct the project without prior experience about the forecasting process in Nokia. Therefore, it is possible that some parts of the process are misunderstood or misinterpreted, which may have led to inconsistent conclusions. However, this risk has been minimized by regular discussions with Nokia's forecasting team and reviewing of project documents, including this report.

Although the forecasting performance tool has been developed and tested by the group, the tool should be independently verified by Nokia to ensure the correct operation and validity of the results it provides.

If the Theil's correction method is used, it must be tested, that the variance of errors remains approximately constant. If for example accuracy is significantly ameliorated by forecasting process improvements, the regression may no more be valid and a generalized least-squares method should be used instead. In combining the forecasts, some attention should be paid to the combination weights. A good initial guess is to use uniform weights or weighting judgmental forecasts somewhat more heavily but the weights may be experimented with once sufficient data for a proper analysis has been collected.

The tool was developed and tested by the project group but as no real data was available the testing was only preliminary. If Nokia is to try out the tool, the validity of its functioning should be verified.

5.2 Future activities

The Excel tool is only a first version of a tool for monitoring and benchmarking forecasts and obtaining and providing feedback. If the tool is helpful some new features could be developed, such as calculating MAPE values for separately for each round and allowing using a user-defined relative timeframe for the models and MAPE and other result values, for example using only the forecasts of two previous years in calculations.

As we conducted the interviews a clear need for improvements in the model were identified. Especially the replacement modeling was seen complex but also clarification of parameters and model's ability to react to changing situations were missed. Based on these observations we recommend that the modeling would be further researched taking into account RMAs' experiences, expectations and expertise.

6 Project summary

In the beginning of the project, the objective was only vaguely defined. The first task of the project group was to find out and decide the actual objective of the study. This was done in co-operation with Nokia's Finnish forecasting team. The objective sharpened from benchmarking methods for combining judgmental and statistical forecasts to a more general study of the forecasting process at Nokia and possible ameliorations to it. In the beginning the objective was to analyze some data obtained on the forecasts and compare the errors in between statistical, pure judgmental and judgmentally adjusted statistical forecasts. In the conception phase of the project the objective was refocused.

The project group had regular gatherings to discuss the issues that came up. The responsibilities were divided such that the whole group first studied the general forecasting process and discussed possible improvements to it and then assisted in gathering references for the literature review. The actual literature review was a responsibility of two group members while the other two members focused on the interviews and questionnaires. In the regular meetings the group exchanged information of matters that came up in the work. The group discussed the results of the first phases of the project and the actual implementation of the model was given to two group members. Finally, the report structure was sketched such that group members were given the task to write the sections on their responsibilities in the final report.

The actual amount of work was relatively well in line with the planned work load. 120 hours was reserved for the project in total and the estimated amount of work per group member was about 100 hours including course meetings, presentations, interviews, meetings, research, model building and report writing.

The main contribution of this project was to point out the weaknesses and strengths of the Nokia forecasting process and propose some concrete measures to be taken to further streamline it. The main objectives were reached although time restrictions prevented a thorough data analysis of forecasts. However, the implemented monitoring tool is easy to take into use in the process if desired. With more time, a more in-depth analysis would have been possible.

The initial risks in the project were that the data would not be obtained or would be too superficial, the research objective is not consistent with the principal's expectations and that the project exceeds its resource limits. The data risk realized during the project, some data was not obtained in time and some data's quality was not acceptable. However, this risk was seen as probable already in the beginning so it did not disturb the progression of the project. The same applied to the objective risk, serious effort was put in the beginning to clarify the goals of the project. The clarity of goals offset the fact that there was little follow-up and control of the project from the side of Nokia's forecasting team. Also, the resources were allocated such that the group worked efficiently and reached the objectives in the reserved time limit.

7 Sources

Balzer, W. K., Doherty, M.E. & O'Conner, R. 1989. Effects of Cognitive Feedback on Performance. *Psychological Bulletin*, 106: 410-433.

Balzer, W. K., Hammer, L. B., Summer, K. E., Birchenough, T. R., Martens, S. P. & Raymark, P. H. 1994. Effects of cognitive feedback components, display format, and elaboration on performance. *Organizational Behavior and Human Decision Processes*, 58: 369-385

Benson, P. G., Önkal D. 1992. The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, 8: 559-573.

Brehmer, B. 1980. In one world: not from experience. *Acta Psychologica*, 45: 223-241.

Bunn, D., & Wright, G. 1991. Interaction of judgmental and statistical forecasting methods: issues and analysis. *Management Science*, 37: 501-518.

Conroy, R., & Harris, R. 1987. Consensus forecasts of corporate earnings: analysts' forecasts and time series methods. *Management Science*, 33: 725-738.

Fildes, R. 1991. Efficient use of information in the formation of subjective industry forecasts. *Journal of Forecasting*, 10: 597-617.

Fischer, I. & Harvey, N. 1999. Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, 15: 227-246.

Goodwin, P., & Wright, G. 1993. Improving judgmental time-series forecasting: a review of the guidance provided by research. *International Journal of Forecasting*, 9: 147-161.

Goodwin, P. 1996. Statistical correction of judgmental point forecasts and decisions. *Omega, The International Journal of Management Science,* 24: 551-559.

Goodwin, P. 2000a. Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, 16: 85-99.

Goodwin, P. 2000b. Correct or combine? Mechanically integrating judgmental forecasts with statistical ones. *International Journal of Forecasting,* 16: 261-275.

Goodwin, P. 2002. Integrating management judgment and statistical methods to improve short-term forecasts. *Omega, The International Journal of Management Science*, 30: 127-135

Harries, C. 1999. Judgmental Inputs to the Forecasting Process: Research and Practice. *Technological Forecasting and Social Change*, 61: 75-82

Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. 1986. The accuracy of combining judgmental and statistical forecasts. *Management Science*, 32: 1521-1532.

Lim, J. S. & O'Connor, M. 1995. Judgmental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making*, 8: 149-168.

Remus, W., O'Connor, M. & Griggs, K. 1996. Does Feedback Improve the Accuracy of Recurrent Judgmental Forecasts? *Organizational Behavior and Human Decision Processes*, 66: 22-30.

Sanders, N. R. 1997. The impact of Task Properties Feedback on Time Series Judgmental Forecasting Tasks. *Omega, The International Journal of Management Science*, 25: 135-144

Sanders, N. R., & Ritzman, L. P. 1990. Improving short-term forecasts. *Omega, The International Journal of Management Science*, 18: 365-373.

Theil, H. 1971. *Applied economic forecasting*, Amsterdam: North-Holland Publishing Company.

Todd, F. J. & Hammond, K. R. 1965. Differential effects of feedback in two multi-cue probability learning tasks. *Behavioral Science*, 10: 429-435.

Tversky, A. & Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. *Science*, 185: 1124-1131.

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