

MASTER

The collaboration of planners and the forecasting system a study in the confectionery sector

van Oudenhoven, B.

Award date:
2020

[Link to publication](#)

Disclaimer

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain



**EINDHOVEN
UNIVERSITY OF
TECHNOLOGY**

The collaboration of planners and the forecasting system: A study in the confectionery sector

Master thesis

Bas van Oudenhoven

Supervisors

Dr. P.P.F.M. van de Calseyde (TU/e, Assistant Professor)

Dr. J.M.P. Gevers (TU/e, Associate Professor)

Ir. B.B.J.P.J. van der Staak (TU/e, Doctoral Candidate)

Host company supervisor

Eindhoven, December 2019

This page is intentionally left blank.

TU/e School of Industrial Engineering
Series Master Theses Operations Management and Logistics

Keywords: forecasting, judgmental adjustment, adjustment size, forecast accuracy, moving average, decision support system, Blattberg-Hoch model

Abstract

Even though sophisticated computational approaches to sales forecasting have been developed, many organizations still (partially) rely on human judgment to estimate expected demand. Previous research has shown that the application of both statistical- and human forecasting methods can be beneficial. How these two approaches can be optimally combined however, is a topic of research that remains largely unexplored.

The scope of the research lies on the collaboration between the system and the human forecaster. The effect of judgmental adjustments to the system-generated forecasts and their impact on the forecasting accuracy of deseasonalized demand is evaluated.

The research finds that the impact of judgmental adjustments to the statistical forecast is undeniably positive at the host company. Adjusted forecasts are generally made in the right direction, to products that carry a large forecasting error. The adjustments result in a forecast that is more accurate than the initial statistical forecast that the planner adjusted. Large adjustments are more effective in reducing error than small adjustments. Adjustments become increasingly effective the closer they are made to demand realization.

Adjustments to important products seem to be of equal quality in terms of forecasting error compared to less important products. Upward adjustments are made more often than downward adjustments, but there is no detectable inclination to overpredict. Upward and downward adjustments do not differ much in quality, although upward adjustments are found to reduce forecast error to a larger extent. Overall, no optimism bias is found.

Based on the analytical results and the findings from the literature study, a set of alternative collaboration approaches were identified and tested. Simulation models were developed and their outcomes were compared to the current forecasting situation. It was found that preventing small adjustments reduces the accuracy of the system, at the benefit of investing less time in the adjustment process thus increasing the average accuracy improvement per adjustment. A model that allows downward adjustments while re-estimating the forecast after an upward indication by the planner is introduced, but is found to harm accuracy. Using an average of the statistical forecast and the judgmental adjustment was also found to be harmful on average. An extension of this approach was constructed by assigning variable weight to the adjustment size for different categories. This extension was found to significantly increase forecasting accuracy with a minimal investment.

Overall, it can be concluded that The host company effectively applies judgmental adjustments, but can improve the efficiency of the process. The thesis shows that restrictions for small adjustments sizes increase adjustment efficiency. However, an integrative method of applying variable weights to the judgmental adjustment shows potential to reduce forecasting error without restraining the forecaster actively in decision making and should be investigated in more detail as an alternative forecasting procedure.

Preface

After almost seven years of studying in Eindhoven, my life as a full-time student is coming to an end. This thesis is the final step for completing my master Operations Management and Logistics. I have poured many hours into this report, which I feel is an addition to the realm of scholarly literature and a great starting point for improving the forecasting process at the host company.

In a distant past, my grandmother worked in the factory and still holds the company to a very high regard. I was therefore very much delighted to see that my open application had been received so enthusiastically by the HR staff and by my to-be company supervisor. I am very grateful for the opportunity the host company has given me and hope that the fruits of my labor are used wisely. I wish the company all the best for the future and it can be sure that to the extent I was not already a loyal customer, they have gained one now!

Secondly, I want to thank everybody at the TU/e that has contributed to this report and my research. After the first time I met Philippe, I was sure that I wanted to perform a research as he had explained it to me. Combining my knowledge of supply chain operations with my fascination for human behavior and what drives it, was a perfect match for me. Philippe gave me the chance to work on something I really care about, which was a very fulfilling experience. On top of this, it was incredibly motivating to work alongside academics that are just as invested in a subject you are yourself. Especially the time, effort and genuine interest Bregje has invested into my thesis was very helpful. I wish both of you the best of luck with the research project and hope our paths will cross again sometime in the future!

Finally, I want to thank dr. Gevers for the very helpful remarks she made about my thesis. They definitely helped raise the quality of the report and of the research, which is a benefit for myself, The host company and hopefully the collection of academic knowledge.

As closing words, I want to thank my family, friends and girlfriend for their unconditional support in me and my project. I could not have managed without you all, for which I am truly grateful.

Bas van Oudenhoven

Management Summary

Introduction

Even though sophisticated computational approaches to sales forecasting have been developed, many organizations still partially rely on human judgment to estimate expected demand. The quality of the produced demand forecasts is essential, as higher rates of forecasting accuracy result not only in financial benefits from reduced inventory, but can also lead to competitive advantages and can consequently increase customer satisfaction (Moon, Mentzer, & Smith, 2003). Previous research has shown that application of both objective and subjective forecasting methods can be beneficial in decreasing the forecast error. How these two approaches can be combined however, is a topic of research that remains largely unexplored.

Objective forecasting is the approach to forecasting in which only computational algorithms are applied. A forecasting model processes the historical data that is presented and returns the value of demand expected based on the previous observations. Subjective forecasting revolves around humans determining the expected demand of a certain product. It relies on the capability of humans to extrapolate historical data and simultaneously include other externalities that are not directly linked to the demand values. Objective forecasting methods are able to process large quantities of data quickly, while subjective forecasting methods can include externalities in the forecast that are difficult to implement in a computational model. By combining the two approaches, both benefits can be used to optimize the forecasting process. A well-known example of combining the two methods is called judgmental adjusting. In this method, a forecast is created statistically first and can be adjusted by the human forecaster subsequently.

In this thesis, a research is set out at the host company to investigate how effective this combination is. The host company currently applies a simple forecasting procedure for its products, by taking the average values of the last 30 weeks and applying this as the forecast for all future weeks. Unrestricted adjustments can subsequently be made to these forecasts of all products at any point in time for any future point in time. Forecasts are generally “fixed”, which means that an adjustment value for period t in the future is also applied to periods $\{t+1, \dots, t+n\}$.

The host company forecasts the deseasonalized series of the final sales figures. Firstly, promotional sales for a period are deducted from the final sales value, as these have been contractually agreed on with the clients and thus are of a fixed size. Next, a seasonal factor is applied to the remainder to find the underlying, seasonless value for the period. In the host company’s forecasting process, this is called the ‘regular sales value’. The four-week ahead forecast of this value is the most important, as the production department creates a production planning four weeks ahead of time.

The research will focus on the collaboration between the system and the human forecaster. The effect of judgmental adjustments to the system-generated forecasts and their impact on the forecasting accuracy of deseasonalized demand is evaluated. By identifying the characteristics of those adjustments that have the biggest influence on the forecast error, it is possible to capitalize on the strengths or steer away from the weaknesses of planners in order to increase the forecasting accuracy. Existing approaches for improving this process are tested and expansions of these approaches are explored.

Literature review

A literature review is performed to acquaint the reader, but more importantly the researcher, with all relevant concepts and theories associated with judgmentally adjusting forecasts. To this end, the literature review is conducted for statistical forecasting models, subjective forecasting models and models that combine these two approaches.

Fildes et al. (2009) and Franses and Legerstee (2009) find that many statistical forecasts are adjusted, up to 80 % and 89.5 % respectively. Biases can influence the value of the adjustment and thus the accuracy and usefulness of the adjustment as a whole. The primacy effect can make a forecaster weigh more recent observations over older ones (Andreassen & Kraus, 1990) possibly causing them to confuse the signal with noise (O'Connor, Remus, & Griggs, 1993), the illusion of control might make forecasters act when no acting is required (Tversky & Kahneman, 1974), while over-optimism, anchoring and inconsistency can make forecasts unreliable (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Lawrence, Goodwin, O'Connor, & Önköl, 2006). Lawrence et al. (2006) however do conclude that forecasters are able to produce more accurate forecasts than statistics can.

Humans have the ability to identify statistical forecasts that require adjusting. This is demonstrated by Mathews and Diamantopoulos (1990) and supported later by Fildes et al. (2009), who also find the direction of the adjustment is generally set in the right direction. The optimism bias could however play a role when the adjustment size has to be determined (Tyebjee, 1987), resulting in over forecasting and a disproportionate amount of upward adjustments compared to downward ones. This is expected to make adjustments downward more useful than adjustments upward (Syntetos, Nikolopoulos, Boylan, Fildes, & Goodwin, 2009). Additionally, Fildes et al. (2009) and Baecke et al. (2017) find that large adjustments deliver more benefits the forecasting accuracy than small adjustments can.

A model aimed at increasing the efficiency of judgmental adjustments, is the Blattberg-Hoch model. This model applies a weight of 50 % to the statistical forecast and 50 % to the judgmental adjustment (Blattberg & Hoch, 1990). By applying these weights, the model reduces the impact of a forecaster's adjustment, which in turn should increase the accuracy of the forecasting process. Fildes et al. (2009) investigate the effects of removing small adjustments from the set, which should in turn increase forecasting accuracy. This thesis applies this methodology in order to investigate the possible benefits in terms of the efficiency of forecasts.

Data collection, processing and cleaning

All the data used for analyses in the thesis are retrieved from the weekly dump files in the ERP system (SAP) The host company applies for forecasting and planning. These dumps contain the information presented in the SAP system at the moment of the finalization of the forecast values. Two dumps are made per week, spanning two halves of the year. The weekly sets are combined for 2017 and 2018. The initial, total set containing all dumps, spans over 8.3 million rows.

Only complete triples are used in the dataset. This means that only observations that have a valid value for the regular sales, regular sales forecast and a judgmental adjustment remain in the dataset. Incomplete, duplicate or invalid observations are removed from the set. A cleaning procedure removing all observations with error values with a Z-score larger than 3 is applied. This results in a set of 93,228 observations, of which 3,198 have a forecast lag equal to 4.

Results, implications and future research possibilities

Contrary to expectations, system-generated forecasts are more often unadjusted rather than adjusted. These results are to be taken lightly, as it cannot be totally verified due to the data storage structure that The host company applies. Retrieving exact numbers was not possible unfortunately, so the expectation is deemed inconclusive rather than rejected.

The thesis shows that adjustments to statistical forecasts increase the accuracy of the forecasts. The adjusted forecast is on average closer to actual regular sales value than the statistical forecast. Moreover, the adjustments decrease the Mean Absolute Error (MAE) associated with the forecasts. This effect does not increase for more important products in the assortment, but the most important products do benefit most from judgmental adjustments. The adjustments are generally made in the correct direction, with a ratio of 67.9 % over all adjustments. Upward adjustments are slightly more often in the required direction than downward ones, with a ratio of 70 % against 69.1 %. Finally, the results allow for an analysis of error values over a longer period of time. The accuracy of the adjustments increases over time, as the moment of demand actualization nears. So not only are the adjusted forecasts more accurate system-generated ones, they are so increasingly.

The inclination to adjust upwards is found in the dataset as well. Even though the required direction is divided evenly over the two directions, the division between the actual adjustment directions is slightly skewed. 52.3 % of adjustments is in the upward direction, which is higher than sheer chance. The results do however show that this does not make upward adjustments less beneficial than downward adjustments. Downward adjustments do improve the forecast accuracy more often than upward adjustments do, with 60.6 % against 57.8 %. However, the upward adjustments decrease the MAE by 16.3 boxes, whereas downward ones deliver an average decrease of 14.4. this makes the upward adjustments slightly more risky, but not less beneficial.

The phenomenon of over forecasting is not found in the data. Even though upward adjustments have a median forecast error of 2 after adjusting, only 43.2 % is over forecast. Surprisingly, there is a larger tendency to under forecast, which occurs for 46.6 % of downward adjustments. Similarly, only 33.1 % of downward adjustments is over forecast and the expected result is thus not found. The relationship between size and error is found however, with large adjustments providing a larger increase in accuracy than small adjustments do.

Based on the analytical results a set of alternative collaboration approaches is identified and tested. Simulation models are developed and their outcomes are compared to the current forecasting situation. It is found that preventing small adjustments reduces the accuracy of the system minimally, at the benefit of investing less time in the adjustment process. A model that allows downward adjustments and re-estimates the forecast after an upward indication by the planner is introduced, but is found to harm accuracy.

Using the Blattberg-Hoch model is found to be harmful on average. The average improvement per adjustment decreases, but a concave function was discovered. Additional investigating lead to the creation of an extension of the Blattberg-Hoch approach, which is constructed by assigning variable weight to the adjustment size for different categories. This extension is found to increase forecasting accuracy and thus increase the efficiency of the forecasting process with a minimal investment.

The host company effectively applies judgmental adjustments, but can improve the efficiency and effectivity of the process. It is advisable for The host company's management to look deeper into the extension of the Blattberg-Hoch model, which shows potential to increase the efficiency of the system. This however is a slim improvement, but comes at relatively low cost and no transfer of responsibilities from the human forecaster to the system.

Additionally, disallowing small adjustments is an approach worthy of consideration. This approach shows a higher accuracy improvement potential, but is based on forecasting human behavior. The size of the adjustment is unknown up until the time of adjusting, so an expectation of adjustment size is required in order to prevent these observations from being adjusted. Moreover, the approach does impose restrictions on the human forecaster. Careful balancing of the transfer of responsibilities and the increase in accuracy is required when selecting either of the two interventions.

List of abbreviations

DSS = Decision Support System

ERP = Electronic Resource Planning

EOL = End of Life

ES = Exponential Smoothing

MA = Moving Average

MAE = Mean Absolute Error

MAPE = Mean Absolute Percentage Error

POS = Point of Sales

PSR = Paired Signed Rank

RQ = Research Question

SKU = Stock Keeping Unit

Table of contents

ABSTRACT.....	IV
PREFACE	V
MANAGEMENT SUMMARY.....	VI
LIST OF ABBREVIATIONS.....	X
TABLE OF CONTENTS.....	XI
1. COMPANY INTRODUCTION.....	1
1.1. BRIEF COMPANY HISTORY	ERROR! BOOKMARK NOT DEFINED.
1.2. COMPANY CHARACTERISTICS	ERROR! BOOKMARK NOT DEFINED.
2. RESEARCH INTRODUCTION	2
2.1. PROBLEM IDENTIFICATION	2
2.2. RESEARCH QUESTIONS	3
2.3. RESEARCH OBJECTIVES	4
2.3.1. PRAGMATIC OBJECTIVE	4
2.3.2. ACADEMIC OBJECTIVE	4
2.4. RESEARCH SCOPE	5
3. STATISTICAL DEMAND FORECASTING.....	6
3.1. AN INTRODUCTION TO FORECASTING	6
3.2. OBJECTIVE FORECASTING METHODS.....	7
3.2.1. TIME SERIES MODELS.....	7
3.2.2. PREDICTING STATIONARY SERIES.....	8
3.2.3. PREDICTING TREND SERIES.....	8
3.2.4. PREDICTING SEASONAL SERIES.....	9
3.2.5. N STEP AHEAD FORECAST	9
3.3. SHORTCOMINGS OF OBJECTIVE FORECASTING METHODS	10
4. SUBJECTIVE FORECASTING AND JUDGMENTAL ADJUSTMENTS	11
4.1. SUBJECTIVE FORECASTING METHODS.....	11
4.2. INTEGRATING FORECASTING METHODS	12
4.2.1. THE ADJUSTMENT DECISION.....	13
4.2.2. ADJUSTMENT CHARACTERISTICS	15
4.2.3. ADJUSTMENT TIMING	18

5. RESEARCH METHODOLOGY	19
5.1. FORECASTING PROCEDURES	19
5.2. DATA COLLECTION.....	19
5.2.1. INITIAL DATABASE CONSTRUCTION	19
5.2.2. CONDITIONAL SELECTION	20
5.2.3. DATA CLEANING	22
5.3. VARIABLES	23
5.3.1. ORIGINAL VARIABLES	23
5.3.2. ADDITIONAL VARIABLES	24
6. RESULTS.....	26
6.1. THE ADJUSTMENT DECISION.....	26
6.2. ADJUSTMENT CHARACTERISTICS	30
6.3. REDESIGNING THE FORECASTING PROCEDURE.....	39
6.3.1. AVOIDING SMALL ADJUSTMENTS	40
6.3.2. REDUCING THE ERRONEOUS DIRECTION ADJUSTMENTS	40
6.3.3. BLATTBERG-HOCH MODEL.....	40
6.3.4. UPWARD INDICATION MODEL	41
7. EXECUTION AND EVALUATION OF SIMULATIONS	44
7.1. ADJUSTMENT SIZE CUT-OFF VALUE	44
7.2. AVOIDING WRONG-SIDED ADJUSTMENTS	45
7.3. COMBINING STATISTICS AND JUDGMENT DIRECTLY	46
7.3.1. BLATTBERG-HOCH MODEL.....	46
7.3.2. DIFFERENT WEIGHTS FOR DIFFERENT CATEGORIES.....	47
7.3.3. SIMULATING THE FORECASTS	48
7.4. INDICATING DIRECTION AND DOWNWARD ADJUSTMENT SIZE	49
7.5. COMPARISON OF SIMULATION OUTPUT	49
8. DISCUSSION	51
8.1. IMPLICATIONS.....	51
8.1.1. PRAGMATIC IMPLICATIONS	51
8.1.2. ACADEMIC IMPLICATIONS.....	52
8.2. SHORTCOMINGS	53
8.2.1. LIMITATIONS OF THE STUDY	53
8.3. ADVICE FOR FUTURE RESEARCH	54
9. CONCLUSIONS.....	56
BIBLIOGRAPHY.....	ERROR! BOOKMARK NOT DEFINED.
APPENDICES	61

APPENDIX 1 - LITERATURE REVIEW SETUP	61
METHODOLOGY	61
PROBLEM FORMULATION.....	62
RESEARCH QUESTIONS.....	62
INCLUSION AND EXCLUSION CRITERIA	63
DATA COLLECTION.....	63
UNIVERSITY COURSE LITERATURE	63
SCHOLARLY ARTICLES AND LITERATURE	64
DATA EVALUATION.....	65
DATA ANALYSIS AND INTERPRETATION.....	65
APPENDIX 2 – SCREEN CAPTURE OF THE SAP FORECASTING MODULE INTERFACE.....	66
APPENDIX 3 - CLEANING OF THE HISTORICAL REGULAR SALES DATA.....	68
APPENDIX 4 – ADDITIONAL GRAPHS	70
MEAN ABSOLUTE PERCENTAGE ERROR VERSUS FORECAST LAG	70
ABSOLUTE PERCENTUAL ERROR VERSUS ABSOLUTE ADJUSTMENT SIZE	71
ABSOLUTE FORECAST ERROR VERSUS ABSOLUTE ADJUSTMENT SIZE	72
MAE IMPROVEMENT PER ABC CATEGORY	73
APPENDIX 5 - STATISTICAL CONCEPTS AND METHODS FOR HYPOTHESIS TESTING.....	75

List of Figures

Figure 1: Unavailable figure due to confidentiality	1
Figure 2: Schematic depiction of a four week ahead forecast	9
Figure 3: Forecasting type in relation to demand history and forecast horizon (from (Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016))	11
Figure 4: Schematic representation of voluntary integration (from (Lawrence, Goodwin, O'Connor, & Önkal, 2006)).....	12
Figure 5: Schematic representation of mechanical integration (adapted from (Lawrence, Goodwin, O'Connor, & Önkal, 2006)).....	13
Figure 6: Mean Absolute Error per forecast lag for statistical and adjusted forecasts for the corrected full dataset	28
Figure 7: Forecast error for adjustments plotted against the size of the adjustment	31
Figure 8: Effect of adjustment size on the improvement in MAE.....	37
Figure 9: Mean MAE values for all forecast lags with linear regression.....	38
Figure 10: Mean MAE improvement value for all forecast lags with a loess regression.....	39
Figure 11: Improvement values plotted by the corresponding weight assigned to the judgmental adjustment.....	47
Figure 12: Schematic representation of literature review process	61

List of Equations

Equation 1: Blattberg-Hoch formula as described by Fildes et al. (2009)	41
Equation 2: Formula applied for calculating the forecast value in the direction indication model	42
Equation 3: General form of Blattberg-Hoch model formula for introducing different weights	46

List of Tables

Table 1: Number of rows per initial dataset	20
Table 2: Sizes of different datasets utilized for analysis	22
Table 3: Variables utilized by the SAP system for producing a demand forecast	24
Table 4: Additional variables created for the dataset	25
Table 5: Size characteristics of ABC product categories for the full dataset of adjustments	26
Table 6: Forecasting error measure for the statistical and adjusted forecasts for the full datasets....	28
Table 7: Hypothesis tests for the accuracy of adjusted versus unadjusted forecasts	29
Table 8: Hypothesis tests for best forecast statistics per ABC category.....	29
Table 9: Hypothesis tests for MAE statistics per ABC category	29
Table 10: Hypothesis tests for comparison of MAE improvement between categories	30
Table 11: Hypothesis test for correctness of adjustment decision direction	32
Table 12: Hypothesis tests for required adjustment direction versus actual adjustment direction	32
Table 13: Hypothesis tests for adjustment direction propensity	33
Table 14: Hypothesis tests for the best forecast per adjustment direction	34
Table 15: Effect of correctness adjustment direction on MAE (boxes)	34
Table 16: Bias characteristics for both adjustment directions	35
Table 17: Over- and under forecasting characteristics for all adjustments in the dataset	36
Table 18: Over- and under forecasting characteristics for correct adjustments.....	36
Table 19: Percentage of correctly selected adjustment direction per size quartile	38
Table 20: Summary by focus, advantages and disadvantages of the proposed adjustment simulations for the research.....	43
Table 21: Error measurement values for disallowing small adjustments to statistical forecasts	44
Table 22: Error measurement values when 50% of wrong-sided adjustments is removed	45
Table 23: Error measurement values for BBH model	46
Table 24: Forecast improvement in units of MAE as compared to the current forecasting procedure for variable weight model and upward indication model	50
Table 25: Research questions for the literature review	62
Table 26: Criteria applied for data collection	63
Table 27: Search queries used for data collection with number of hits per database indicated	64

1. Company Introduction

Chapter 1 is the introductory chapter of the thesis. Upon request by the host company, it is not included in the public version of the thesis. Company information can be distributed on need-to-know basis after request at and consideration by the author.



Figure 1: Unavailable figure due to confidentiality

2. Research introduction

This chapter will be devoted to formulating the specific problem The host company has observed, which will be tackled in the current thesis. Subsequently, the objectives of the thesis from the perspective of the company are presented, as well as the academic objectives and how the thesis aims to contribute to academic literature. The presentation of the acquired academic knowledge resulting from the literature review is divided into a chapter on statistical forecasting (chapter 3) and a chapter on judgmental adjustments of statistical forecasting (chapter 4). The latter will also incorporate the formulation of several hypotheses that will shape the execution of the research. Preparation and planning for the literature review can be found in Appendix 1. Chapter 5 will show the methodology applied for the research and thus give insight into how the research is performed.

2.1. Problem identification

The host company produces and resells approximately 180 products, resulting in a total of €90 million in sales. This order of sales implicates a large and complex supply chain. To ensure operational excellence across the entire supply chain, The host company applies statistical forecasting methods to estimate the expected amount of sales that will occur at a certain period of time in the future. These forecasts are created for weekly intervals and updated every week, but checked and adjusted on a monthly basis by a human forecaster. The quality of the produced forecasts is essential for The host company, as higher rates of forecasting accuracy not only result in financial benefits from reduced inventory, but can also lead to competitive advantages and can consequently increase customer satisfaction (Moon, Mentzer, & Smith, 2003).

The forecast values are made definitive four weeks ahead of demand realization, so the production department has the possibility to prepare its production planning. Until this point, forecasters and account managers are allowed to change the statistical forecast values. This is done to correct for information that has not been incorporated in the statistical calculations, like logistical problems or production issues. These swings in the demand pattern that have not been captured by the system can cause a disparity between the statistical forecast and the realized demand. The planners at The host company attempt to reduce this discrepancy, by adjusting the forecast to a value that they believe to more accurately represent expected demand.

An adjustment can be made for periods as far as a year ahead, by “fixing” the expected demand value after a certain point in time. This means that the expected demand for a certain period can be applied to several succeeding, consecutive weeks. In other words, an adjustment of the statistical forecast is generally made for multiple weeks in the future simultaneously. These values can be adjusted every time a check of the product is performed.

In a final step, the logistics manager investigates the final decision created by the forecasting team. This planner examines the adjustments and corrects them for errors and mistakes if he feels there are any. These final values are then forwarded to the production department, where the forecasts are used in the production planning process. This final evaluation of forecasts serves as the “handbrake”-moment to correct for mistakes that the forecaster feels the production department should be made aware of. It is performed right before demand actualization, as a one-week ahead forecast. It is only performed in case of drastic, last-minute changes in the expected demand.

Based on this process description, it can be said that The host company applies judgmental adjustments in its forecasting process. They are applied restrictively, as the adjustments made by the forecaster are applied as a restriction on the system-generated forecasts and not incorporated in the statistical forecasting calculations (Baecke, De Baets, & Vanderheyden, 2017). Currently however, The host company monitors neither the behavior of the forecasters, nor the results of their adjustments. Within the Supply and Demand management team, this has created some concern. They suspect a negative influence of forecasters' adjustments on the accuracy of the original, statistical forecasts. This is due to a recent development, where The host company was forced to incur losses in inventory, as demand forecasts were too high. Management expects this to have occurred as a result of the adjustments, but is unsure of the validity of this claim.

This problem boils down to a lack of understanding of the changes in forecast accuracy resulting from judgmental adjustments to statistical forecasts. This means that The host company cannot tell if the human forecaster is decreasing the discrepancy between sales forecast and actual sales with the judgmental adjustments made or not. The statistical forecasts of expected demand are known, as well as the final, adjusted forecast. However, no deeper insight is currently being produced with this information. This implies that an erroneous forecast value might be the result of a bad forecasting system output that is worsened by a human forecaster, a bad forecasting system output that is dampened by a good human forecaster, or a good forecasting system output that is worsened by a human forecaster.

Based on this information, the first problem statement is:

Problem statement 1: The host company currently has no insight into the impact that different types of judgmental adjustments have on the quality of the demand forecasts.

In the light of The host company's future improvement plans, the absence of these insights becomes problematic. The company is on the verge of acquiring a new ERP software package, the costs of which will go over multiple millions of Euros. Optimally implementing the new ERP system is difficult, since no information on the advantages and disadvantages of adjustments is currently present. Consequently, no presumptions can be made about how to capitalize on the benefits of these adjustments and simultaneously to avoid the potential harm they might cause. Currently, the forecasters are allowed to adjust statistical forecasts if they feel there is a need to do so. However, this approach might not be the most effective solution for adjusting forecasts. This leads to the second problem statement:

Problem statement 2: Given that different types of adjustments may have different effects on the quality of the demand forecasts, The host company wants to know the most efficient and effective approach to adjusting statistical demand forecasts.

2.2. Research questions

In order to generate a solution for the problem statements as identified in 2.1, a set of research questions is defined in this paragraph. These will give guidance to the thesis, functioning as starting points for the literature review and shaping the hypotheses formed in the literature section. In order

to be able to provide a solid solution to the problem, the research questions will rely heavily on the problem statements. The questions are as follows:

Research question 1: What relationships exist between the characteristics of judgmental adjustments and the accuracy of the demand forecast when adjusting a statistical forecast?

Research question 2: Given these relationships, how can adjustments to the statistical forecasts be made most effectively and efficiently?

These research questions will be answered by analyzing the process of adjusting statistical forecasts. To this purpose, the following elements of a product's forecast are considered: the initial statistical forecast for a period, the corresponding judgmental adjustment by the planner, and the actual sales value. All analyses will be performed on data retrieved from The host company. The hypotheses will guide the research towards answering the research questions. In turn, these hypotheses are based on an extensive literature review.

2.3. Research objectives

A master thesis should produce results that are both interesting and useful. Not simply for the host company to be applicable in day-to-day conduction of business, but also for the scientific community as a whole. This implies the thesis should simultaneously contain considerable business related outcomes (optimization, cost-savings, etc.) and contribute to existing scientific research on judgmental forecasting. In this section, these two goals will be formulated for the current research in order to be able to reflect on the success of the thesis afterwards.

2.3.1. Pragmatic objective

The first objective to be formulated is the pragmatic objective, which considers the host company. For The host company, it is essential for the thesis to produce results that generate some kind of business-related insight. Therefore, the thesis aims to provide The host company with new information about their forecasting process and in particular about the interaction their forecasters have with the forecasting system. Additionally, the thesis will present an improved situation in which forecasters and forecasting system are aligned in better fashion, in order to achieve higher levels of performance. This final section of the pragmatic goals is coherent with research question 2 (RQ2), and the achievement of this goal will thus rely on the ability of the thesis to answer this question.

2.3.2. Academic objective

Apart from the business objective, the thesis will also consider the academic impact of the research it aims to conduct. This goal is best reflected by RQ1, which aims to find the apparent effects of human intervention on statistical forecasts within The host company. This produces two academically relevant results when answered.

Research on the topic RQ1 considers has been done before, so the thesis does not claim to do greenfield research and explore uncharted territory. However, it can potentially support findings made in literature and thus confirm existing theories. Moreover, research in this area is partly reliant on laboratory studies and studies in controlled environments with unexperienced forecasters. Arvan (2018) found that an upward trend exists for case studies and real-data studies. Additionally, studies

of integrating methods have increased in popularity recently. Still, room exists for empirical research aimed at the confectionary sector. It allows establishing under which pretenses judgmental adjustments prove beneficial in a currently unexplored forecasting environment (Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016).

Additionally, the forecasting method that will be examined is different from existing literature. The host company applies a rolling forecast to its data, which gives forecasters the opportunity to adjust the statistical forecast for a long period of time before the final decision has to be made. These adjustments will have a different level of forecasting accuracy and provide a differing quality. The relation of accuracy with the adjustment's period in time before demand realization will be examined in this research. Currently, no other literary source has been identified to include this factor in the calculation of the optimal forecasting setting. Thus, the thesis will contribute to literature by researching outside the known bounds of previous research.

By examining the answers to RQ1, answers to RQ2 can be formulated. Through identification of the characteristics of adjustments that have the biggest influence on the reduction of forecast error, it is possible to capitalize on the strengths and steer away from the weaknesses of adjustments in order to increase the forecasting accuracy. Existing improvement approaches are applied and expansions on these approaches are explored. The thesis therefore contributes to literature by validating previously made claims and by putting forward claims about new improvement approaches.

2.4. Research scope

In order to keep the thesis and thus the graduation project as a whole manageable, and to assist attaining the research objectives and answering the research questions, a scope is set for the thesis. This scope will limit the amount of possible directions the research can venture into, thus helping to keep the research focused. In this thesis, the scope will be set on the statistical demand forecasting process within The host company, along with interventions performed by human forecasters. Processes that directly influence the forecasting process will be considered, all others are not included in the research. This means that purchasing and production processes will be ignored, but the information they produce regarding expected sales can be included if it is picked up in the forecasting process.

3. Statistical Demand Forecasting

Chapter 3 contains an introduction to the domain of forecasting in general, with statistical demand forecasting in particular. It introduces the most important concepts and theories in this field, as well as its benefits and shortcomings.

3.1. An introduction to forecasting

Companies operate in the realm of uncertainty, with ambiguous externalities having serious implications on their affairs and supply chain. This uncertainty arises from the period of time that spans between the production decision and the corresponding demand realization for a product (Ghiani, Laporte, & Musmanno, 2013). To reduce the degree of uncertainty a company is confronted with, it can adopt a forecasting system as a part of its supply chain and resource planning.

Ghiani, Laporte and Musmanno (2013) define forecasting as “an attempt to determine in advance the most likely outcome of an uncertain variable”. This means that by applying forecasting, the company aims to gain knowledge about the value of a certain variable in the future to assist the decision making process in the present. In the case of demand forecasting, the future demand value is forecasted as a reinforcement for current day strategic decision making. Since the bulk of strategic decisions (capacity & location decisions) and operational decisions (inventory control and production planning) are driven by demand, statistical forecasts are able to provide great value to the company (Nahmias, 2013; Arvan, Fahimnia, Reisi, & Siemsen, 2018).

Nahmias (2013) identifies a set of five characteristics that model makers should consider if they are to add value to the strategic decision making process with their model:

- First of all, it is important to realise that a forecast is highly unlikely to exactly forecast the value that is ultimately realized in real life. As Nahmias (2013) put it more consisely; “forecasts are usually wrong”, a property of forecasting that is frequently disregarded. Variability, from the market or the environment, can cause unexpected changes in the forecasted phenomenon, to which systems should be able to respond properly (Cachon & Terwiesch, 2012).
- Secondly, forecasts should always include some type of forecasting error, to elaborate the information produced by the forecast. The type of forecasting error that is applied can influence which forecasting model is utilised.
- The third characteristic concerns the level of aggregation the model applies. A forecast made for the entire Stock Keeping Unit (SKU) will have a lower variance of forecasting error than a combination of forecasts made on individual Point Of Sales (POS) level of the same SKU.
- Fourthly, the accuracy of a forecasting model will be reduced if the time horizon it has to span is increased. This comes from the variability mentioned in the first point which is exarcabated the longer it has to manifest itself. Bell (1984) refers to this as “unreasonable long-run behavior”, as the forecast will start to deviate at some point in the future.

- Finally, forecasts should not be applied to exclude available information. Potential accuracy increases are lost if the forecasting model is inefficient; if it is unable to optimally integrate all known information into the forecast (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009).

Another characteristic of forecasting models that model makers should consider is that no forecasting model can be considered to be the absolute best. All models are approximations of specific settings and should be customized to the properties of the product or company it is used for (Ghiani, Laporte, & Musmanno, 2013).

Forecasts can be generated by a vast amount of different methods, divisible into two categories; quantitative- and qualitative methods (Ghiani, Laporte, & Musmanno, 2013), or objective and subjective respectively (Nahmias, 2013). Quantitative methods are based on historical data, which are historic values of relevant variables. This method is applicable if substantial amounts of data are available (Ghiani, Laporte, & Musmanno, 2013). Qualitative methods are based on human judgment and other non-historical, non-quantifiable data (Nahmias, 2013). Statistical forecasting models apply a historical dataset to arrive at a forecast and are therefore synonymous to quantitative methods of forecasting.

3.2. Objective forecasting methods

As mentioned in the previous paragraph, quantitative (or objective) models are based on the use of quantifiable information. These statistical forecasting models can be divided into two different categories; time series models and causal models (Nahmias, 2013).

- *Time series models* utilise only a set of historical data to distil an expected value of the variable under investigation. This property has earned time series models the name 'naïve' models, as they disregard all other types of information about the forecasted phenomenon.
- *Causal models* can also use historical data, but exclude the values of the variable that it aims to predict. These models adopt several variables that show some connection to the variable to be forecast in linear or non-linear fashion to arrive at a prediction for its future value. The key difference between the two model types is that time series forecasting applies a recursive calculation, while causal models calculate the forecast directly.

3.2.1. Time series models

Time series models apply a form of extrapolation, as they use historical data to make assumptions about the behaviour of this data in the future. Nahmias (2013) identifies four different patterns that appear regularly in time series trend, seasonality, cycles and randomness.

- *Trend* is the tendency of a time series to have an increasing or decreasing nature over a period of time.
- *Seasonality* refers to a pattern in the time series that occurs at fixed periods in time.
- *Cycles* are patterns that repeat themselves, but the intensity and duration at which they occur differs at every occurrence.

- *Randomness* typifies time series for which no pattern can be found in the data.

These patterns can be found in combination with one another (i.e., a seasonal pattern can have an upwards trend).

Ghiani, Laporte and Musmanno (2013) use a similar explanation for patterns in time series. However, they use the term 'irregular' for a time series that is completely random, that is already explained by the randomness term, as shown by Nahmias (2013). Therefore, the terminology of Nahmias (2013) is applied throughout the remainder of the thesis.

A set of commonly applied statistical forecasting methods can be identified from Nahmias (2013). These methods are applied to the time series' dataset, in order to arrive at a prediction. In the following sections, these methods are introduced shortly, in order to get acquainted with them.

3.2.2. Predicting stationary series

The methods in 3.2.2 are most suitable for forecasting stationary time series. This type of time series is constructed from a stationary demand process, which entails that the parameters of demand arrival (mean, variance, etc.) remain constant over a period of time. In other words, there is a constant arrival factor, with addition or subtraction of a random error term. Nahmias (2013) identifies two methods for forecasting this type of series.

- *Moving Average (MA)*: this method is very straightforward, as the expected demand for the next period is simply the average demand of N preceding periods.
- *Exponential Smoothing (ES)*: in the case of exponential smoothing, the forecast produced for a period is the weighted average of the forecast produced for the preceding period and the corresponding realized demand with a smoothing factor alpha. This is similar to the MA method, but requires less storage of data, as only the most recent observation has to be stored (Chatfield, 2005).

Both methods are applied quite easily, as only one parameter has to be determined for them to work. How these values should be determined, is not specified and is a case of trial and error. A downside for these methods is they lag behind a potential trend, making them useful for stationary series.

3.2.3. Predicting trend series

Series that contain a trend have a tendency to either continuously increase or decrease. For predicting a trend, Nahmias identifies two useful methods; regression and analysis and double exponential smoothing.

- *Regression analysis*: over the values of the historic dataset, a linear relation is assumed. A line is then created that has the smallest deviation from the set of data points and extrapolated for predicting future demand.
- *Double exponential smoothing (Holt's method)*: a method based on the exponential smoothing technique. In this variation, a second smoothing constant is introduced. This results in a linear component on top of the exponential smoothing forecast, which is suitable for predicting trend series.

3.2.4. Predicting seasonal series

For seasonal series, the forecasting techniques become more complicated and involve more calculations. Nahmias (2013) identifies three techniques specifically applicable for a seasonal series:

- *Seasonal factors*: applicable for seasonal series with no trend. For every period of data observed in the time series, a seasonal factor is calculated. This factor is an estimated value for a specific period which is to be multiplied with the mean of the series in order to arrive at the forecast.
- *Seasonal decomposition*: useful when the seasonal series contains a trend as well. Apart from calculating a seasonal factor, a MA is applied in the calculation in order to keep track of the increasing or decreasing tendency of the demand process.
- *Winter's method*: a complex forecasting technique, based on the concept of triple exponential smoothing. Apart from smoothing the forecasting error and a trend, it now incorporates seasonal factors as well.

The host company applies seasonal factors in its forecasting procedure. Firstly, promotional sales for a period are deducted from the final sales value, as these have been contractually agreed on with the clients and thus are of a fixed size. Next, a seasonal factor is applied to the remainder to find the underlying, seasonless value for the period. In the host company's forecasting process, this is called the 'regular sales value'. This deseasonalized time series is forecasted by the system with a moving average.

3.2.5. N step ahead forecast

Forecasting methods can produce a prediction of the demand value for a certain period in the future. To the model, it does not matter which period of time this is. It will produce a forecast with the historical information that is given to it for a certain number of periods ahead of that period. So in period t , a forecast is calculated for n periods ahead; period $t + n$. This forecast is noted as F_{t+n} . In the case of The host company, the forecast is produced for four periods ahead. This implies $n = 4$, meaning that in the current week F_{t+4} is forecasted. This is illustrated by the red block in Figure 2.

The host company applies a moving average model of 20 weeks, implying the average demand value of the most recent 20 observations is used as a forecast. This is depicted in Figure 2. The average is calculated by summing all periods from now, D_t , until 20 periods back, D_{t-19} , and dividing the sum by 20. This then becomes the forecast.

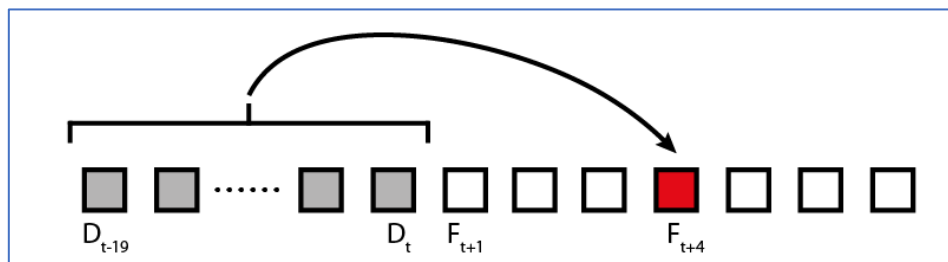


Figure 2: Schematic depiction of a four week ahead forecast

The system The host company applies is not strictly a four step ahead forecast, as the calculations are also applied to the preceding and succeeding periods. The calculated average can be applied for a period of up to 30 weeks ahead, but are used very little at that point. Four weeks before demand realization, the most important resource acquisition- and planning decisions are made. Therefore, the forecasting method is four-step ahead forecast in practice.

3.3. Shortcomings of objective forecasting methods

The first paragraph of section 3.2 introduced a flaw in time series models that is crucial in their application. Time series models can only interpret historical data and disregard all other available information. This means that crucial information that is present within the company or forecasting department is lost when a time series model is blindly applied. Information such as promotional sales, weather conditions and other production or purchasing information can be vital in generating a useful forecast. A causal model could take away this this problem, but incorporating a lot of information into a forecasting algorithm can make it complex, which can prove to be counterproductive (Lawrence, Goodwin, O'Connor, & Önköl, 2006).

Moreover, these automated forecasting methods distance the forecaster from the data that a forecast has to be produced for. Although the results the methods often produce useful and reliable results, the forecaster will be unable to learn from the data (Bell, 1984).

Besides the distance that objective methods create between the human forecaster and the forecast, these methods can also create trust problems. This distrust could be caused by algorithm aversion, the propensity of people to distrust or ignore computational algorithms and favour human expertise (Arvan, Fahimnia, Reisi, & Siemsen, 2018; Dietvorst, Simmons, & Massey, 2016). This bias is costly, as it can lead to suboptimal forecasts (Dietvorst, Simmons, & Massey, 2016).

A potential solution to these flaws would be to apply subjective forecasting methods. To obtain a further grip on this subject and introduce the hypotheses for the research, subjective models will be presented in more detail in chapter 4.

4. Subjective forecasting and judgmental adjustments

As briefly touched upon in section Shortcomings of objective forecasting methods 3.3, qualitative models can offer a solution to the shortcomings of quantitative models. This chapter introduces the most widely applied qualitative models and discusses their advantages and disadvantages. Moreover, it introduces the concept of judgmental adjustments. The literature found regarding this subject is linked to the research questions and problem statement, in order to formulate hypotheses.

4.1. Subjective forecasting methods

As opposed to objective forecasting methods that are based purely on the analysis of underlying data, subjective forecasting methods rely on human judgment (Nahmias, 2013). Nahmias (2013) distinguishes four types of subjective forecasting methods, which are as follows:

- *Sales force composites*: The company's sales team is asked for their opinion of the expected demand which is combined into the forecast for the period under review.
- *Customer surveys*: Customers are asked to fill in surveys in an attempt to gather sufficient purchasing information that can serve as the forecast for a certain period of time in the future. The population that is surveyed should be representative for the entire customer base and the amount of replies should be plentiful.
- *Jury of executive opinion*: If no historic sales information exists about a certain product, as is the case with a new product's launch, a group of experts comes together to discuss the product. Together they arrive at a forecast for the new product's sales pattern.
- *The Delphi method*: The Delphi method is similar to jury of executive opinion. However, the Delphi method applies a round-based system that provides different experts with different information in every round. Implementing the round system prevents a bandwagon effect, in which people tend to agree with the most commonly available opinion (Ghiani, Laporte, & Musmanno, 2013).

These subjective forecasts function best when they are applied to a time horizon of medium- or long length, thus for strategical decision making problems (Ghiani, Laporte, & Musmanno, 2013). This is opposed to the application of statistical methods, which are more suitable for shorter time horizons as well as in the case of large amounts of decisions to be made in a short span of time (Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016).

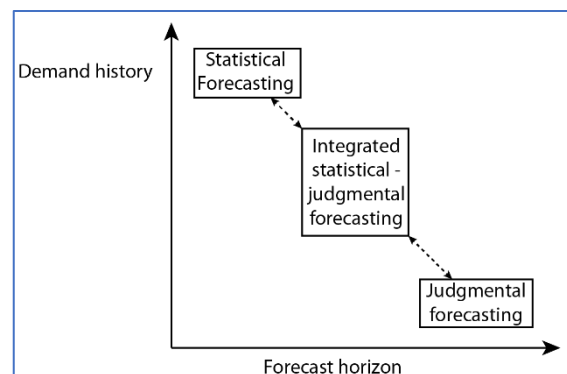


Figure 3: Forecasting type in relation to demand history and forecast horizon (from (Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016))

Moreover, subjective forecasting can work with very little to no historical data present. This is depicted in Figure 3, where the optimal usefulness of the forecasting technique is placed along two axes: the amount of historic information available and the length of the forecast horizon.

The main advantage of applying subjective methods over objective methods is the possibility to capture contextual knowledge in the forecast. This knowledge reflects the information learned by human forecasters while working in a certain environment, where the forecaster has become more sensitive to certain cause and effect relationships, as well as environmental cues (Sanders & Ritzman, 1995). The contextual knowledge contains a large part of the unmodeled information not incorporated in time series models, which is information that time series models struggle with (Goodwin & Fildes, 1999). Therefore, it can be concluded that subjective forecasting can offer a solution an intrinsic shortcomings of objective forecasting.

Combinations of objective and subjective forecasting methods are called combination forecasts (Lawrence, Goodwin, O'Connor, & Önkal, 2006) or integration forecasts (Goodwin, 2002). These forecasting methods can use the speed of statistical forecasting and produce a large set of forecasts quickly based on historical data and human input. Alternatively the statistical forecasts can be adjusted by human forecasters as to incorporate the contextual component. In theory this sounds like a perfect solution, but just as the objective forecasting methods, subjective forecasting methods have some undeniable shortcomings. In the remainder of this chapter, combination forecasts will be explained more thoroughly and will be linked to the thesis' research.

4.2. Integrating forecasting methods

Forecasting by means of combination forecasts can be roughly divided into two categories:

- *Voluntary integration*: A statistical method is applied to arrive at an initial forecast, which can be adjusted by a human forecaster afterwards. This type of intervention in a statistical forecasting method is called a 'judgmental adjustment' (Goodwin, 2002). It is a restrictive judgment model, as the forecaster's prediction is used as a restriction on the statistical forecasting model's outcome (Baecke, De Baets, & Vanderheyden, 2017). This type of integration is depicted schematically in Figure 4.
- *Mechanical integration*: In this type of combination, a statistical model is applied after a forecaster has released an initial forecast, based on non-historical data and possibly historical data too (Goodwin, 2002). The forecaster's prediction is used as an input variable in a statistical forecasting model, creating an 'integrative judgment' model (Baecke, De Baets, & Vanderheyden, 2017). This forecasting method is illustrated in Figure 5.

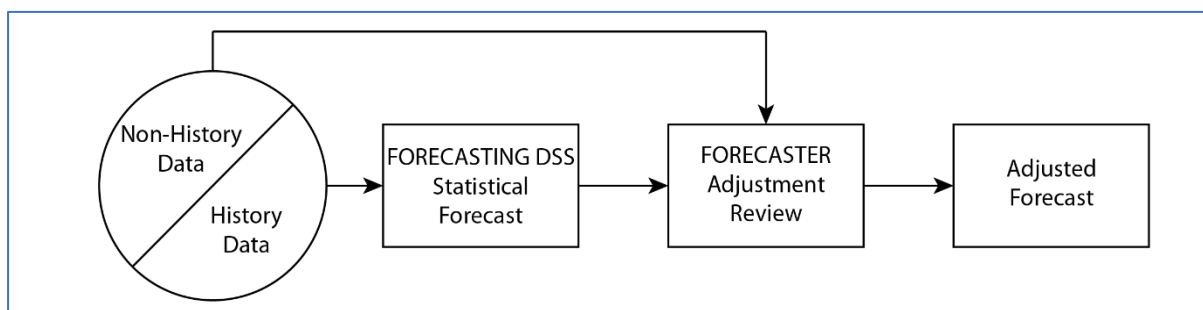


Figure 4: Schematic representation of voluntary integration (from (Lawrence, Goodwin, O'Connor, & Önkal, 2006))

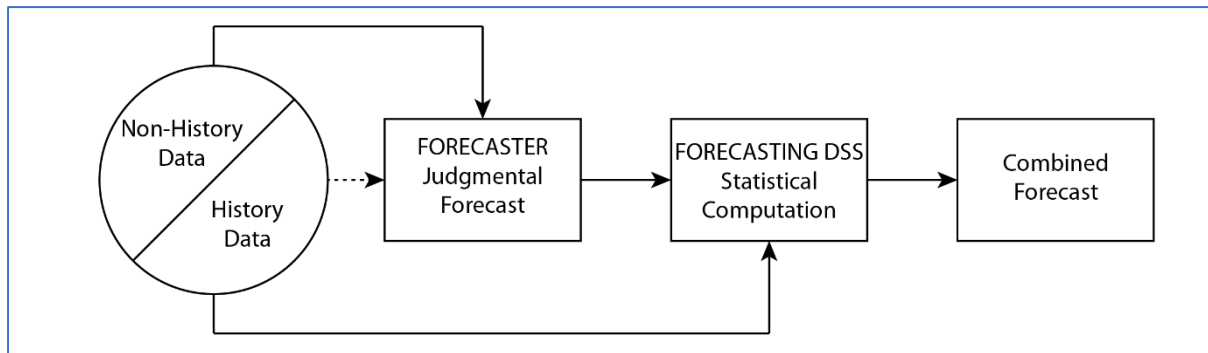


Figure 5: Schematic representation of mechanical integration (adapted from (Lawrence, Goodwin, O'Connor, & Önkal, 2006))

A very well-known example of mechanical integration is the 50/50 heuristic. It was introduced by Blattberg and Hoch in 1990 as an attempt to combine a simple database model with human intuition (Blattberg & Hoch, 1990). The heuristic simply takes the statistical forecast and the human forecast and averages the two. The approach is tested in the fashion sector, by estimating sales for certain items at one company, and among three different companies that forecasted their coupon codes' redemption rates. The results indicate that managers can correct for almost 25% of the variance not explained by statistical models and improved the forecasting accuracy in all settings, proving their complementary value (Blattberg & Hoch, 1990).

For forecasting its future demand and sales, The host company applies a voluntary integration method of forecasting. Statistical forecasts are produced by the Decision Support System (DSS) and thereafter evaluated by the human forecaster. The statistical forecasts are adjusted if deemed necessary. In this process of judgmental adjustments of quantitatively produced forecasts, two stages are defined (Arvan, Fahimnia, Reisi, & Siemsen, 2018).

- *Adjustment decision*: The human forecaster decides whether or not the system-generated forecast is in need of an adjustment.
- *Adjustment characteristics decision*: If an adjustment is considered to be necessary, the human forecaster makes a decision on the size and the direction of the adjustment.

Both stages possess certain properties that can potentially benefit or hamper the forecasting process and its accuracy, due to the subjective nature of the judgmental adjustments. The next two sections will cover these properties and link them to the research questions and objectives via hypotheses. Finally, a new adjustment stage is introduced to be reviewed in this research.

4.2.1. The adjustment decision

As mentioned previously, the biggest advantage of judgmental forecasting is the possibility of including contextual information in the forecast. Lawrence et al. (2006) regard contextual information as a requirement for human forecasters to be more accurate than statistical models. Webby et al. (2005) find that the absence of this information in objective models is the reason why forecasters prefer judgmental forecasting for important products in their portfolio. Thus, the presence of contextual information in the decision making process can increase the perceived necessity for an adjustment.

Research also found that confidence in the forecast increases if the forecaster adjusted initial statistical forecast (Kotteman, Davis, & Remus, 1994). This is the result of the so-called illusion of control, a bias that inflates people's confidence in the outcome of decisions when they have a say in shaping the decision (Langer, 1975). It implies that a forecaster is more likely to ignore a system-generated forecast and feels more confident about his/her own forecast, even if there is little evidence to justify a higher probability of outcome of the forecast.

Finally, forecasters possess a need to show they have handled and reviewed data (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). As a means of demonstrating their occupation with the forecasting task, they intervene in the statistical forecast by adjusting the system-generated forecasts. This is so-called 'tinkering with data' effect (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009).

A slightly different explanation comes from Patt and Zeckhauser (2000), who use the term "action bias". They put forward the idea that people have an intrinsic- or routinely-based desire to take action, which "clouds decision making". This desire is carried over to situations where action is not required, resulting in a negative utility of action.

Fildes et al. (2009) find that up to 80% of statistical forecasts is adjusted. In a similar vein, a different study shows that in 89.5% of cases, forecasters made an adjustment to statistical forecasts (Franses & Legerstee, 2009). These results are based on one-step ahead forecasts for products in the pharmaceutical industry. The data analyzed in this study came from a total of 37 countries, so it accounts for local effects that might occur. Even though the sector under investigation is vastly different, the first hypothesis is drawn up as follows:

Hypothesis 1: Forecasters are more likely to judgmentally adjust the statistical sales forecasts, as compared to accepting them.

Human biases in forecasting are an insurmountable problem for various researchers and practitioners. For example, the primacy effect can make a forecaster weigh more recent observations over older ones (Andreassen & Kraus, 1990) possibly causing them to confuse the signal with noise (O'Connor, Remus, & Griggs, 1993), while over-optimism, anchoring and inconsistency can make forecasts unreliable (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Lawrence, Goodwin, O'Connor, & Önkal, 2006). Given these biases, some researchers therefore argue that forecasting methods should only implement a judgmental component during promotional periods and around special events (Goodwin, 2000), as only periods of high demand variability could benefit from incorporating non-modelled information (Lawrence, Goodwin, O'Connor, & Önkal, 2006).

However, one should consider that most studies on forecasting methods are set in a laboratory setting and less than 17% of the articles used data and information from practice (Arvan, Fahimnia, Reisi, & Siemsen, 2018), to which Goodwin (2000) is no exception. An issue arises when these artificial series are applied, as the unique combination of historical knowledge and contextual knowledge is not utilized (Lawrence, Goodwin, O'Connor, & Önkal, 2006). This sets the approach of adjusting forecasts at a disadvantage, as they are not used to their full potential.

Lawrence et al. (1985) compared 111 series from the M1 forecasting competition's total of 1001 series. This forecasting competition compared forecasting methods between themselves using real-life sales data (Makridakis, et al., 1982). They concluded that judgmental forecasts were at least equally accurate as statistical forecasts, but did so with lower standard deviation of error. Additionally,

they discovered that statistical forecasting methods were more correlated with one another, than they were correlated with judgmental forecasts. Consequently, a combination of judgmental- and statistical forecasts could in fact provide benefits to the forecasting process (Lawrence, Goodwin, O'Connor, & Önkal, 2006). Moreover, Willemain (1991) conducted an experiment with 10 volunteers, who adjusted a naïve and automated forecast of 24 series made from M1 data. The presentation of these sets rotated among the volunteers, who adjusted the forecasts in two separate sessions. He concluded that humans can correct statistical methods effectively once these methods have gone too far off track. Fildes et al. (2009) gathered data from four supply-chain companies. Three of these companies were manufacturing companies that applied monthly forecasts, while the fourth company was a retailer that forecasted on a weekly basis. They found that forecasters are able to alter statistically forecasted values when necessary.

Lawrence et al. (2006) concluded through a literature review that forecasters are able to produce forecasts more accurate than system-generated forecasts. In understanding this, they provided the two following reasons. Firstly, human forecasters possess unmodeled information that the system does not have. Additionally, the system works with information that is no more recent than the last observation. The forecaster can pick up and apply information before the system can and thus has an edge of the statistical forecast. In line with this reasoning, the hypothesis is as follows:

Hypothesis 2a: Judgmentally adjusted forecasts will have a higher level of accuracy as compared to the initial, statistically produced forecasts.

Moreover, Fildes et al. (2009) drew attention to fields of possible future research, and indicate that more 'important' categories could experience higher accuracy values as a result of judgmental adjustments. This sentiment is supported by the occurrence product familiarity. Product familiarity delivers a significant forecasting accuracy improvement, due to forecasters being exposed to the product on a more frequent basis (Edmundson, Lawrence, & O'Connor, 1988). This research thus used the frequency of exposure as a proxy for product familiarity, which has been confirmed to increase accuracy in a recent study by Baecke et al. (2017).

The category of important products as coined by Fildes et al. (2009) is defined in the current thesis as 'products that contribute most to the company's sales volume'. These products are generally referred to as ABC categories, which include products based on the share they have in realizing revenue. Products in the A category are responsible for the largest set of a company's revenue, products in the B and C contribute increasingly less. As such, one would expect that forecasters are more frequently exposed to important products, on the account of them being sold to a larger set of customers. This in turn increases the familiarity of a forecaster with these products and their demand pattern behavior. The hypothesis becomes as follows:

Hypothesis 2b: Judgmentally adjusted forecasts will be more effective for products higher in the ABC hierarchy, compared to adjustments made to lower ranked products.

4.2.2. Adjustment characteristics

After a forecaster has decided on the necessity of an adjustment, the next set of decisions arises regarding the adjustment characteristics. Firstly, the direction of the adjustment is to be determined. The direction (or sign) of the adjustment means whether the forecaster will change the forecast value

upwards, thus increasing the forecast value, or downwards, as to decrease the initial forecast value. A study by Mathews and Diamantopoulos (1990) showed that managers are well able to select the most outlying forecast values and correct them in the appropriate direction. This is later supported by Fildes et al. (2009) who found after analyzing more than 60,000 forecasts and outcomes from real-life companies that, on average, the direction of the adjustment is set in the right direction. Based on this, the following pattern is expected to emerge at Pijnenburg:

Hypothesis 3: Judgmental adjustments to statistical forecasts are more likely to be made in the correct direction.

The management team has indicated that The host company currently over forecasts for the majority of its SKUs. They expect that forecasters could be setting the forecasting values too high. The forecasts on which the production schedules are based result in excess inventory of SKUs and thus in unnecessary losses after demand realization. However, as the management has no insight on the performance of the forecaster (see paragraph 2.1), they cannot substantiate their claims. However, previous literature does provide an explanation for this phenomenon.

In paragraph 4.2.1 the term over-optimism has briefly been mentioned, as well as confidence in human adjustments that was elaborated on further. Over-optimism, or the optimism bias, is a bias that arises from a combination of wishful thinking and the illusion of control (Tyebee, 1987), which increases the likelihood estimation for positive events and decreases the likelihood estimation for negatives ones (Tversky & Kahneman, 1974). The illusion of control makes the forecasters overly confident of their capabilities. Combined with wishful thinking and their inclination to expect a positive outcome, will lead to forecasters willingly adjusting forecasts upwards. Fildes et al. (2009) in their empirical study found this pattern, with 55% of judgmental adjustments being in the upward direction. Additionally, people are known to prefer their own opinion over a forecasting system (Webby, O'Connor, & Edmundson, 2005) and might draw less attention to a downward forecasting advice due to the optimism bias. Therefore, a new hypothesis can be established:

Hypothesis 4a: Forecasters will be more inclined to adjust the initial statistical forecasts in an upward direction.

Fildes et al. (2009) also reviewed the effectiveness of the adjustment direction on the accuracy of the forecasting system. By comparing the initial statistical forecast to the judgmentally adjusted forecasting values, they discovered that upward adjustments increase the Mean Absolute Percentage Error (MAPE) by 10.2 %. The opposite is true for downward adjustments, which resulted in improvement of almost 23 % for the MAPE. A different study showed similar results specifically for a restrictive judgment setting. Here, upward adjustments increased the MAPE by 19.9 %, while downward adjustments decreased the error measure by 18.2 % (Baecke, De Baets, & Vanderheyden, 2017). Finally, Syntetos et al. (2009) show in their research “quite conclusively” that downward adjustments outperform upward ones. Their results are based on a lumpy demand pattern however, which is not true for The host company’s demand.

A theoretical explanation for this phenomenon comes from the bounded nature of the downward adjustments. An upward adjustment is expected to be much larger due to its potentially unbounded size. A downward adjustment on the contrary is bounded by zero, as negative forecasts are non-sensible.

Mistakes in the selection of size for downward adjustments are thus mitigated by its boundedness, which is not the case for upward adjustments. Practical explanations are the aforementioned wishful thinking and the optimism bias. This leads to an additional hypothesis:

Hypothesis 4b: Downward adjustments are more beneficial for the forecasting accuracy than upward adjustments.

In addition to upward adjustments being more present in the total set of adjustments, another implication can be drawn up from Fildes et al. (2009). The adjusted forecast tended to be overly optimistic, even if the adjustment was made in the downward direction. Upward adjustments were adjusted too far upward in almost three out of four cases. This could be explained by the phenomenon of 'overpredicting', in which forecasters consistently assign too high values to the expected demand (Lawrence, Edmundson, & O'Connor, 1985). This phenomenon was however observed in a different forecasting setting than their study and could thus be related only to that specific setting.

For downward adjustments, over-optimism was found almost half of all cases (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). This means that when forecasters feel the statistical forecast is too high and adjust the final forecast downward, the realized demand will still be lower than the adjustment value almost 50% of the time. This result could arise from the propensity to adjust upwards; leaving all statistical forecasts that needed a downward adjustment over forecasted regardless of the direction. The downward adjustments seem to be more accurate, as they near a more fifty-fifty division of over- and under forecasting.

This would mean that on average, adjustments would be higher than the expected demand value. Additionally, if the adjustment was made in the correct direction, an inclination to adjust too high is probably present as well due to reasons mentioned earlier. This leads to the following hypothesis:

Hypothesis 5: Adjusted forecasts are expected to be consistently higher than the realized demand value.

The second adjustment characteristic the forecaster has to consider is the size with which the forecasts will be adjusted. Previous research has found a relationship between adjustment accuracy and adjustment size. Larger adjustments will generally produce a larger accuracy improvement than adjustments that are relatively small (Diamantopoulos & Mathews, 1989). These small adjustment can even produce results that are less accurate than the unadjusted forecast (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009).

Fildes et al. (2009) examined this claim in their research and found it to be valid still. Baecke et al. (2017) upheld the claim that small adjustments are harmful, but added that the relationship of size with improvement is curvilinear. This implies that both small and very large adjustments will be detrimental to the forecasting accuracy, following the relationship's inverted U-shape. However, Baecke et al. (2017) obtained this result in a single company, implying a possibly local phenomenon. The results obtained by Fildes et al. (2009) were found over multiple companies and the relationship they found thus is more robust against local effects.

Fildes et al. (2009) expect this is due to differing levels of available or used information for large versus small adjustments. Large adjustments are bound to be based on considerable evidence that is expected to have a large impact on the sales pattern. The information on which these large adjustments are based is generally more reliable and will lead to more reliable results.

For small adjustments, this is generally not the case. The information is probably less reliable than the information on which a large adjustment is based, which can cause the forecaster to hedge the adjustment by keeping it relatively small as to avoid a potentially large error. Moreover, small adjustments could also be a result of the ‘tinkering with data’-effect introduced in 4.2.1, where forecasters simply adjust statistical forecasts to show they are actively handling the data. This results in adjustments that are based on very little, if any, information and will prove harmful for the forecasting accuracy. Thus; the following can be hypothesized:

Hypothesis 6: Large adjustments are more effective for decreasing forecasting error than small adjustments.

4.2.3. Adjustment timing

Finally, a new adjustment stage will be considered in this thesis; the decision regarding when the planner is able to add significant forecasting accuracy to the system. In this research, it will be referred to as “adjustment timing” and will express the moment in time before demand realization when the forecaster will make the adjustment decision. A hypothesis is to be formulated for the investigation this research will do on the timeframe of adjusting and what its influence on the forecasting accuracy will be. As currently no empirical data can be identified on the adjustment timing decision, the evidence provided is purely based on the characteristics of subjective and objective forecasting.

Syntetos et al. (2016) provide Figure 3 in their article, specifying the most appropriate time window for forecasting a time series. This distinction can be found in other sources as well (Ghani, Laporte, & Musmanno, 2013; Nahmias, 2013). This division however is based on the expected practicality of implementing these methods, not on the comparative accuracy of them. However, the division does make sense, as subjective forecasting can incorporate many factors that cannot be modeled and objective forecasting can handle large data quantities easier. The information that can be incorporated into the model by forecasters increases over time, as more information about the demand realization becomes available. Additionally, forecasters will get more acquainted with the products that they monitor extensively for a longer period of time due to the large exposure time (Edmundson, Lawrence, & O'Connor, 1988). The expected improvement in forecast accuracy resulting from a judgmental adjustment (as hypothesized in H2a) will be higher if more information is available as a more substantiated decision can be made. The expected pattern to be found therefore is as follows:

Hypothesis 7: Judgmental adjustments will be increasingly beneficial for the forecasting accuracy, the closer they are made to the moment of demand realization.

5. Research Methodology

Chapter 5 introduces the research design of the study that will be executed at the host company. The process of data collection is discussed and the variables considered are put forward. Finally, the method and process of analysis of the data are explained.

5.1. Forecasting procedures

The host company applies a four step ahead, 20-week moving average to forecast the expected demand for its products. In order to estimate the demand for the forthcoming periods, the Electronic Resource Planning (ERP) system calculates the mean of the past 20 weeks and extrapolates this over a period of 35 weeks in the future. These forecasted values are not cast in stone, as The host company applies a rolling forecast principle. This means that the forecast is not made for 35 weeks ahead at a single period in time, but changes after every period by adding the most recent observation to the equation. The values of the moving average are corrected by a seasonal, after which the effects of promotions are added.

The forecasting process within The host company is a rolling forecasting process. At every point in this rolling forecast, an adjustment can be made to the expected regular sales. These are the baseline sales that would be expected to occur if no promotional event would intervene. The adjustments made by the demand planner for a certain product are made at the client level, and involve filling in the total expected sales volume. This means that the demand for each product is forecasted by the SAP system for each customer, which subsequently can be judgmentally adjusted. This adjustment does not ask for an increment or decrease in the forecast, but for an entirely new forecast value. The majority of forecast checks are on a monthly basis, during which point the forecast can be adjusted for any point in time in the future. The user interface of the adjustment tool is given in Appendix 2.

The final forecast produced by the system, which also contains (delayed) promotional sales, other contractual sales and a seasonal factor, is the second value that can be adjusted in the forecasting process. This forecast is adjustable until the planners' meeting, which is on Thursday until 16:00. In this meeting, the forecasts for the coming weeks are checked and finalized. Moreover, during this meeting the handbrake correction (as explained in section 2.1) can be applied.

If the final forecast is expected to miss actual demand according to the logistics manager, a final alteration is made by making an adjustment this final time. This correction usually is very accurate, as it is applied to correct for a systematic error in the promotional sales calculations. Unsold parts of promotional contracts are added to the final forecast in a backorder type fashion. However, The host company wants to apply a lost sales assumption in the forecasting procedure for these unsold promotional products. Therefore, this row is also used as a systematic correction for systematical erroneously forecasted unsold promotional products.

5.2. Data collection

5.2.1. Initial database construction

All the data used for analyses in the thesis are retrieved from The host company's internal storage facilities, the weekly dump files and the ERP system (SAP) The host company applies for forecasting and planning. These dumps contain the information present in the SAP system at the moment of the finalization of the forecast values, thus at Thursday 16:00. Two dumps are made per week divided over two halves of the year. These halves span from week 1 to 26 and from 27 to 52.

One dump is made for the current half-year and a second one for the subsequent half-year. They store the data recorded in the SAP system columns at the time of dumping (Appendix 2), not the changes made between over time. This means that every dump file contains the sales history and lastly made adjustment for past weeks, as well as statistical forecasts and judgmental adjustments for future weeks. As mentioned in 5.1, adjustments are made at the client level, thus combinations of product information and customer are stored for every week in the half-year that the dump was made for. These data points are stored as rows in the dumps, with each row representing a unique combination in that particular dump.

The analyses for the thesis as well as the database construction process are executed in the statistical computing language R (version 3.6.0). Two databases will be constructed initially in order to create the master set. Firstly, the historical sales set is created. This file contains the information about The host company's actual sales per week per client. Only four dump files are required for this, being week 26 and 51 from both years. As the final dump of a half-year contains historic information only and therefore sufficient for gathering all demand actualization information that occurred in 2017 and 2018. Week 51 has to be used instead of week 52, since unfortunately no dumps for week 27 to 52 are made in week 52. This totals the number of data points to 181,622.

Secondly, the forecasting dataset is constructed. This file is a grand summary of the client-oriented dumps for a two-year period of time, spanning from week 1 to 52 of 2017 and week 1 to week 51 of 2018. Every dump is loaded into R separately, combined with the other half from its week and provided with additional variables to represent the calendar week for the forecast period and the dump, calendar month for the dump and the sequential entry number of the dump into the total set. The weekly sets are combined into two sets for 2017 and 2018, after which a variable is added reflecting the year of creation of the dump. The total set combines the two yearly files, resulting in an initial starting point of over 8.3 million rows.

	Total set 2017	Total set 2018	Total set
Number of rows	4,872,049	3,485,720	8,357,769

Table 1: Number of rows per initial dataset

5.2.2. Conditional selection

The files initiated in the previous section require extensive cleaning, as they contain data points that are either unnecessary for this research or will damage the accuracy of the results from the analyses. The dump files are a rough representation of reality, due to the presence of products in the dataset whose information is unreliable, missing or incomplete.

Due to the construction method of the initial datasets, some weeks contain incomplete information. Since the sales information is derived from half-year files, the historical set is created from four half-year dumps. The final dump for the first half-year is created in week 26 of the running year, the final dump for the second half in the week 51 of the running year. However, the week in which the dump is made is ongoing. This means that not all sales information can be included in the dump. Definitive information only becomes available in the next week, in which no dump is made of the previous half-year. As a consequence, the historical regular sales values should be removed for the weeks that the half-year files are constructed in. This means that sales for week 26 and week 51 of both years should be removed.

Also, week 52 sales will not be part of the historical sales set on account of it coming after week 51 and no final sales information about week 52 will be known in week 51. The sales values are dropped from the historical set and will not find their way into the total set. This means that all rows that contain a forecast for one of these weeks will be removed, as there is no final sales value to compare them to. Forecasts that were made in these weeks are still useful, considering the dumps do contain definitive forecasting values. The periods for which the forecasts are made in week 26 and 51 do contain the correct historical sales value and thus are included in the dataset.

In order to complete the total set, it is linked to the historical sales dataset. This entails the completion of every line of data in the forecasting set with the eventual sales actualization, so forecasting performance can be examined efficiently. This means that even a row containing a forecast made in week 2 for week 12 will be completed with the sales value of week 12. As a result, the forecasting performance can be measured for every row in the dataset.

The first cleaning measure is to remove duplicates from the total set. A dump made of the current half-year will contain both forecasts and historical values associated with that half-year. In order to create a set that contains only forecasts, all data points that contain information about a period in the past are removed. As the eventual realized demand values have been added to the forecasting rows through the historical sales file, the historical information is not lost.

Secondly, the products that are not treated like the bulk of products are removed. Either no forecast is made for these products, or no adjustment to the forecasts is stored, so inclusion of them will not be useful for the analysis.

Due to inclusion of sales data, it is possible to remove data points from the dataset that belong to the five categories mentioned above. The sales for these products are not stored within the regular sales variable, but in the other variables available in the SAP system. Zero regular sales could thus imply that the product was sold as one of the five exceptions rather than through regular sales, or is an end-of-life (EOL) product of which the client no longer orders any amount. Furthermore, the zero sales periods could also be part of an intermittent demand pattern. In line with the reasoning of Fildes et al. (2009) and Syntetos et al. (2016), intermittent demand patterns are removed from the total dataset as they should not receive the same attention as regular products.

In addition, products with a statistical forecast value of zero are deleted from the dataset. A forecast of zero is an erroneous forecast value, as The host company applies a MA approach. If a sale has occurred in the past, this value can never be equal to zero as the mean of all observations is calculated. The forecast value of zero will therefore be the result of some exceptional event or, possible, of a system error. The data point is therefore removed from the set.

Furthermore, periods with negative sales are removed from the set as they indicate a return or recall of products. Since the forecasting process does not take returns or recalls into consideration, these periods are not representative of forecasting performance and will be deleted from the set.

In case of a negative forecast, a large recall has likely clouded the calculation of the mean for a low volume product. This results in a forecast lower than zero, which is an unrealistic value and rows carrying this value will be deleted.

Judgmental adjustments that have negative 1 as their assigned value are removed from the dataset. This is a specific input in the SAP system, which means that the product does not have to be forecasted any longer. It has been dropped from the assortment and will be deleted from the system in the future. This would generate false results since the judgmental adjustment is not actually negative 1, but the forecaster is telling the system to ignore a certain product. Therefore, all rows that have a negative 1 for the judgmentally adjusted value are dropped from the set.

After all these corrections, the total set has decreased to 1,201,046 rows of data. This is a reduction of 85.6%, meaning that over 7.1 million rows have been deleted. In the remainder of this chapter, the structure of the dataset is explained and the methods of analysis are explained on the basis of the variables that make up the structure.

5.2.3. Data cleaning

After constructing the dataset by selecting relevant data points from the initial set, a final cleansing is required for the data to be useful for analysis. Firstly, only the data points that have been adjusted are extracted for the analysis. This implies only the rows of data that contain a complete triple (statistical forecast, adjusted forecast, actual sales) are useful for analysis. Additionally, rows in which the statistical forecast exactly equals the judgmental adjustment are not included. The host company's SAP system can be told to accept the statistical forecast in an implicit and explicit way. The implicit method involves entering a zero in the column for the judgmental adjustment, the explicit method is as described above and requires filling in the value of the statistical forecast. Unfortunately, the SAP system automatically generates a zeroes in the judgmental adjustment column if no value is entered manually. This implies that no knowledge can be generated about the acceptance of statistical forecasts. Due to the lack of information on the part of this type of "adjustment" (acceptance), it has been decided to not evaluate the explicit acceptance in this thesis, as it would convey a message about incomplete information.

When creating a new set consisting of all the complete and useful triples, this set spans 96.196 rows. This set represents all the adjustments made to statistical forecasts in 2017 and 2018.

	All data points	Forecast lag equal to 4
Initial extracted data points	96,196	3,312
Z-score corrected	93,228	3,198

Table 2: Sizes of different datasets utilized for analysis

In order to obtain information about the most crucial forecasting moment, a second set is created that contains only triples made four weeks ahead of sales realization. These forecasts comprise about 3.4% of the total set of adjustments (see Table 2). This is double the mean, resulting from the amount of adjustments increasing over time due to fixing adjustment values for longer periods.

Finally, a cleaning procedure is performed based on the z-scores of the errors of the observation. All triples that have a z-score larger than 3 for their statistical and adjusted error are removed from the set. This roughly translates to removing all observations that are more than three standard deviations removed from the center of the normalized values. This results in a final set of 3,198 rows that are useable for analysis. Note that the subset of data points with a lag equal to four is made first, after which a z-score correction is performed. This is done to ensure outliers are removed based on values for the specific period and not the full set.

5.3. Variables

5.3.1. Original variables

The Excel files created from the SAP system values provide a number of variables that all play a role in The host company's forecasting process. All variables useful for analyses, as well as variables that have been implicitly mentioned in the thesis, are corroborated on and can be found in Table 3. These variables are represented as rows in the SAP system, as opposed to the Excel files, in which they are turned into columns (see Appendix 2).

The variables' names originally are in Dutch, but the thesis only refers to them in their translated equal. Abbreviations remain unaltered, and thus reflect the original Dutch name. This approach has been chosen to avoid confusion between system and analysis data and reduce the chance of mix-ups. All variables are constructed on a weekly basis. The unit of sales referred to is boxes of products.

Name	Abbreviation	Description	Type
Historical regular sales	HRV	All orders that have been delivered on a regular sales basis in a certain week (realized demand)	Integer
Historical action sales	HAV	Orders linked to a promotional contract	Integer
Intercompany sales	ICV	Sales of storage kept elsewhere	Integer
Average regular sales	GRV	Historical average of HRV and ICV (MA)	Integer
Manually adjusted regular sales	MGRV-i	If the GRV is not representative, the forecaster can apply an adjustment here	Integer
Manually adjusted regular sales	MGRV	If a value for MGRV-i is entered, the SAP system uses it here, otherwise GRV is copied	Integer
Seasonal factor	SF	A seasonal factor for correcting the MGRV	Decimal
Regular prognosis	RP	MGRV multiplied with SF	Decimal
Confirmed pre-/after-dip	PD/AD	An expected dip in sales due to a future or past promotional event	Integer
Corrected regular prognosis	GRP	RP corrected for PD or AD	Decimal
Confirmed action prognosis	BAP	The confirmed amount of promotional sales for a period	Integer
Extra promotional demand	EV	BAP – GRP	Decimal
Action backlog	AA	Promotional units leftover from preceding week	Integer
Customer orders	KLO	Orders placed by customers at the end of the week	Integer
Deliveries	LEV	Orders delivered to customers at the end of the week	Integer
Salesplan	SLS	GRP + EV	Integer
Salesplan + 50% rule	SLS-i	Add 50% to SLS if many last-minute orders come in	Integer
Salesplan + action backlog	SLSAA	SLS + AA	Integer

Manual salesplan	DSLS	If the SLS (or SLSAA) is not representative, the forecaster can apply an adjustment here (handbrake variable)	Integer
Corrected salesplan	GSLS	If a value for DSLS is entered, the SAP system uses it here instead of the SLSAA	Integer

Table 3: Variables utilized by the SAP system for producing a demand forecast

From the original set of variables; HRV, GRV and MGRV(-i) will be most important. They respectively reflect the realizations of demand, statistical forecasts and adjustments of the regular sales series. Table 3 clarifies the route that is taken from the (adjusted) regular sales forecast to the final forecast GSLS.

GSLS is the variable that production planners actually consider, four weeks ahead of demand realization. The forecast for a certain week is consulted only once more by the planning department, on the last day before a new week begins. At that moment the handbrake variable DSLS is checked by the production manager for any major alterations entered during the forecasters' meeting. This means that any large adjustments after the four week ahead mark are better implemented in the DSLS variable. If at times last-minute information would force a handbrake adjustment, no distinction can be made between a systematic correction or a last-minute forecasting adjustment. Apart from the adjustment itself, no other information about the handbrake adjustment is stored.

5.3.2. Additional variables

In addition to the initial variables, a collection of variables has to be created that allow for analysis of the data. These variables will be constructed from the variables present in the SAP system and are applied to the master data set. They will reflect the adjustment decision and the adjustment size, the relevant forecasting errors, the corresponding mean absolute errors and the week in which the forecast was produced. This provides the opportunity to see the effect of adjustments made long before demand realization and of this adjustment being altered again closer to demand realization. Moreover, it is then possible to assess the adjustment characteristics that prove to be most beneficial and offer the highest contribution to accuracy.

Name	Notation	Description	Type
Month of the run	Maand_run	The month of the year in which the forecast was made, numbered from 1 to 12	Integer
Number of the run	Num_run	The consecutive, overall number of the run	Integer
Week of the run	Week_run	Specifies the week of the year in which the forecast was made	Integer
Year of the run	Year_run	The year in which the forecast was made	Integer
Year to be forecast	fcst_year	The year of the period that is to be predicted	Integer
Week to be forecast	fcst_Period	The week number in the year of the period that is to be predicted	Integer
Forecast lag	fcst_lag	The number of weeks between the moment that the forecast was produced	Integer

		and the period that is predicted by the forecast	
Row ID	row_ID	A forecast identifier, which is a combination of the article number, the year and week to be forecast and the client number	String
Error	error	The difference between the forecast produced in a run (GRV or MGRV) and the realized demand of the forecasted period (HRV)	Integer
Mean absolute forecasting error	MAE	A measure of forecast error for the row, calculated by taking the absolute value of 'error'	Integer
Mean absolute percentage forecasting error	MAPE	A measure of forecast error for the row, calculated by dividing the absolute value of the forecasting error by the realized demand.	Percentage
Statistical mean absolute percentage forecast error	stat_MAPE	The forecasting error for the initial statistical forecast GRV, calculated identical to MAPE	Percentage
Adjustment size	size	The size by which a statistical forecast is altered in a judgmental adjustment decision	Integer
Adjustment direction	direction	A dummy variable indicating whether the judgmental adjustment was made in the upwards or downwards direction	Binary
Adjustment direction correctness	direction_correct	A dummy variable which indicates whether the judgmental adjustment was made in the right direction, regardless of size	Binary
Best forecast	sys_plan	A dummy variable, that is given the string value "STAT" if the statistical forecast resulted in the lowest forecasting error and the string value "JA" if the judgmental adjustment was closer	

Table 4: Additional variables created for the dataset

6. Results

After completing the master dataset, the validity of the hypotheses formulated in chapter 4 is to be examined. The hypotheses are repeated for completeness in this chapter, after which the explanation regarding their validity is presented. Graphs are created first to obtain visual information. Then, the hypotheses are tested by means of statistical analysis (see Appendix 5 for more information). Finally, the report will reflect on the legitimacy of the hypotheses put forward in chapter 4.

6.1. The adjustment decision

By analysis of the historical sales, every product can be assigned to its ABC category. This division among categories is applied at multiple instances in the research. It is found that, based on the two-year total sales, there are 53, 50 and 54 products in category A, B and C respectively. The categories sum up to 79.7 % for A, 15.2 % for B and finally 5.04 % for C, which is close to the original values set at The host company. The observations in the set of adjustments are assigned to their respective categories, which produce the results given in Table 5.

Category	Products	Adjustments (%)
A	53	51,388 (55.1)
B	50	28,605 (30.7)
C	54	13,235 (14.2)
Total	157	93,228 (100)

Table 5: Size characteristics of ABC product categories for the full dataset of adjustments¹

Hypothesis 1: Forecasters are more likely to judgmentally adjust the statistical sales forecasts, as compared to accepting them.

X

The number of individual sales periods in the data set can be found by looking for the unique number of row IDs in the cleaned historical dataset, which turns out to be 43,603. Every period has multiple forecasting moments that correspond with it, made over the course of time. These forecasting moments are opportunities at which a forecast can be adjusted. The amount of weeks that spans between the period the forecast was made in, and the period the forecast was made for is defined by the forecast lag (Table 4). The cleaned master set is essentially the collection of all the forecast lags larger than zero. In total, this adds up to 1,185,847 different forecast data points. This means that in total 1,185,847 moments have passed to which an adjustment could have been made to the data. Of these 1.2 million forecast lags 93,288 periods have been adjusted, leaving the other 1.1 million observations to be the statistical forecasts.

The dataset cannot directly provide information regarding the amount of data points at which the demand planner actively accepted the statistical forecast. The only direct information available is of the data points that were adjusted, as the adjustment itself is stored in the dump files. This means that the only way to identify the ratio between the adjusted- and unadjusted forecasts is to be derived logically.

¹ For the entirety of chapter 6, if not indicated otherwise the tabulated results are based on data from the the Z-score corrected dataset with forecast lag equal to four.

As discussed earlier, a general structure is applied in the judgmental adjustment process. Products from category A and B are checked on a regular basis. This regular time interval is once every four weeks, at which the demand planner can decide to either accept or reject the statistical forecast. The cleaned master set contains roughly 1.2 million rows of code. Categories A and B together are responsible for 1,055,500 rows of data in the master set. Data points are generated on a weekly basis, but the demand planner only checks them once every four weeks. Every week new data points are created, irrespective of the actions of the demand planner. This results in the majority of data points never have being observed and not being evaluated by the demand planner.

The regularity of the schedule does allow for computations regarding the acceptance ratio of statistical adjustments. Since products in category A and B are checked every four weeks, only a quarter of the 1,055,500 data points is actually observed. This means that $\frac{1055500}{4} = 263,875$ data points have been observed. As the total amount of adjustments made to category A and B is 82,871 (Table 5), the amount of statistical forecasts that is rejected rather than is $\frac{79,993}{263,875} = 0.303$, which is equal to 30.3%. When category C is assumed to be unobserved and an adjustment to it is considered to be “extra”, the ratio of accepted adjustments is about 35.3%. Both of these results disprove H1, implying no bias towards actively adjusting forecasts exists.

Hypothesis 2a: Judgmentally adjusted forecasts will have a higher level of accuracy as compared to the initial, statistically produced forecasts.



For an initial assessment of the added value of judgmental adjustments, a simple graph is plotted in Figure 6. The green line indicates the Mean Absolute Error (MAE) of the adjusted forecasts, which is lower than the MAE for the initial values of these adjustments for almost all forecast lags. From this graph, three deductions can be made that relate to this section of the literature study under investigation.

Firstly, forecasters are well able to detect products that require adjusting. The line of the initial statistical forecasts is above the green line for the majority of the graph, implying that the forecaster is able to pick up on erroneous statistical forecasts. Figure 6 is therefore consistent with findings by Fildes et al. (2009), who found that forecasters have the ability to identify the statistical forecasts that are in need of an adjustment.

Secondly, H2a is seemingly supported by Figure 6 as the judgmentally adjusted forecasts possess a lower error value than the statistical forecasts they were based on. This means that the judgmental adjustments have reduced the forecasting error and have added value to the forecasting accuracy.

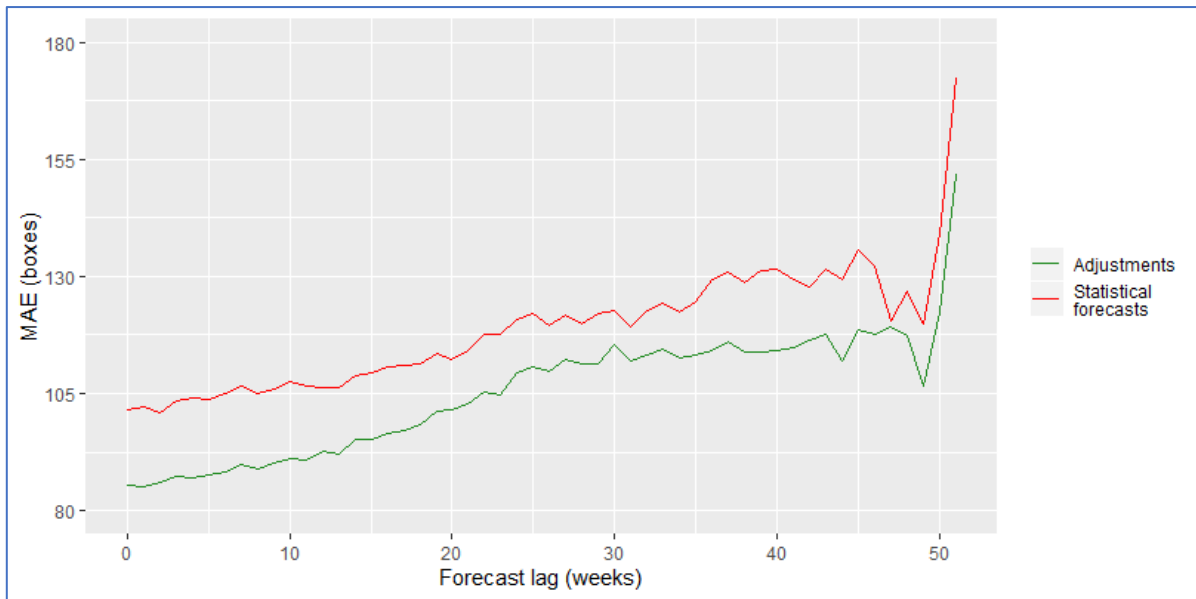


Figure 6: Mean Absolute Error per forecast lag for statistical and adjusted forecasts for the corrected full dataset

As a third and final observation, the distance between the red and green line increases as the forecast gets closer to demand realization. This would imply that adjustments provide an increasing amount of improvement to the statistical forecasts when the decision is made closer to the sales moment. This phenomenon would support H7, which assumes an increase in accuracy over time. H7 will be discussed later in this chapter.

In Table 6, the information from Figure 6 is given in tabular form. It shows that the improvement through judgmental adjustments is 21.2 boxes averaged over all forecast lags. The MAE for the adjusted set is lower than the MAE for the unadjusted set, supporting H2a.

	Observations	MAE (boxes)
Statistical forecasts	93,228	112
Adjusted forecasts	93,228	98.4

Table 6: Forecasting error measure for the statistical and adjusted forecasts for the full datasets

After more thorough analysis of the dataset, the frequency counts of which forecasting method is the most accurate can be created. For the cleaned, four week ahead forecasts it is found that the adjustments are more accurate in 59.1% of the cases, as is presented in Table 7. The scores are found to be independent of one another after applying a Chi-square test of independence, $\chi^2(1) = 106$, $p < 0.001$, and it is therefore safe to conclude that forecasts are closer to the eventual demand value after an adjustment has been made. The difference is noticeable, as the MAE is reduced from 97.6 to 82.2. This implies an average improvement of 15.4 units in terms of the MAE.

Both the statistical and adjusted forecasts errors are not normally distributed, as the Shapiro-Wilk test points out, $W = 0.791$, $p < 0.001$. This means a non-parametric test is required. After testing with a Wilcoxon Paired Signed Rank test, the difference in means turns out to be highly significant, $V = 2,964,018$, $p < 0.001$. By combining the results of both tests, H2a can be confirmed.

	Most accurate forecast (%)	Mean Absolute Error
Statistical forecast	1307 (40.9)	97.6
Adjusted forecast	1891 (59.1)	82.2
Statistical test	Chi-Square test	Wilcox paired rank sum test
Test statistic	$\chi^2 (1) = 106$	$V = 2,964,018$
P-value	$< .001$	$< .001$

Table 7: Hypothesis tests for the accuracy of adjusted versus unadjusted forecasts

Hypothesis 2b: Judgmentally adjusted forecasts will be more effective for products higher in the ABC hierarchy, compared to adjustments made to lower ranked products.

X

As the hypothesized characteristics of the adjustments of H2a have been confirmed, assessing the same hypotheses on a deeper layer could prove insightful. H2b assumes just this and expects to find a pattern to emerge on the previously defined category level. For the purpose of evaluation, the characteristics of the categories are displayed in Table 8 and Table 9.

		A (%)	B (%)	C (%)
Best forecast	Statistical	651 (39.5)	416 (42.0)	240 (42.8)
	Adjusted	996 (60.5)	574 (58.0)	321 (57.2)
Test statistic: Pearson Chi-square		$\chi^2 (1) = 71.8$	$\chi^2 (1) = 24.9$	$\chi^2 (1) = 11.4$
P-value		$< .001$	$< .001$	$< .001$

Table 8: Hypothesis tests for best forecast statistics per ABC category

	A	B	C
Statistical MAE (boxes)	140	54.0	50.8
Adjusted MAE (boxes)	120	45.9	36.8
Improvement in MAE (boxes)	20.3	8.10	14.0
Test statistic: Wilcoxon PSR (V)	777,229	288,653	91,315
P-value	$< .001$	$< .001$	$< .001$

Table 9: Hypothesis tests for MAE statistics per ABC category

As stated in H2b, the effectiveness of adjustments is expected to increase with the importance of the categories. In other words, category A adjustments are expected to outperform those of category C. When assessing Table 8, this effect seems to emerge as the percentage of adjustments that outperform the statistical forecast increases. As the differences per category are significant, the pattern does hold.

The effectivity of judgmental adjustments is not only based on the ratio of winning forecasts, but can be reflected in the reduction of the MAE as well. Per category, the unadjusted MAE is compared to the MAE after the adjustment has been made. For all categories, a Shapiro-Wilk test is applied to test for normality. The statistical MAE values are found not normally distributed for all three categories: category A ($W = 0.743, p < .001$), category B ($W = 0.579, p < .001$), category C ($W = 0.505, p < .001$). Additionally, the adjusted MAE values were also found not to be normally distributed: category A ($W = 0.750, p < .001$), category B ($W = 0.588, p < .001$), category C ($W = 0.559, p < .001$). This means a Wilcoxon PSR test will be applied. The decrease in MAE instantiated by the adjustment is found to be highly significant for all categories, as presented in Table 9.

The results from Table 9 clarify the decrease in MAE for all categories, as the Wilcoxon PSR test shows the results have different distributions. As the calculated improvement is indeed meaningful, there is an opportunity to check if the value of the improvement that results from the judgmental adjustments decreases when going from category A to category C. This is tested by comparing the mean improvement in MAE value between categories against the alternative that the mean for category that is named first is greater. Results for these tests are provided in Table 10, which can be found below.

	A versus B	B versus C	A versus C	A versus BC
Test statistic: Wilcoxon Rank Sum (W)	$8.42 \cdot 10^5$	$2.83 \cdot 10^5$	$4.82 \cdot 10^5$	$1.32 \cdot 10^6$
P-value	.0820	.248	.0606	.0372

Table 10: Hypothesis tests for comparison of MAE improvement between categories

With the results of the Wilcoxon tests from Table 10 and the mean improvement values from Table 9, it becomes clear that the category A has the highest improvement value. The test results for the Wilcoxon test are not significant for the 95% confidence level. As category A takes up most of the observations, the smaller sample size of both category B and C prevents the individual comparisons to reach a value that is significant. This changes however when category B and C are combined.

After creating a subset with all products that belong to category B and C, a set of 1,551 rows is what results. This set has an unadjusted MAE of 52.8 and an adjusted MAE of 42.6, implying an average improvement of 10.2 in terms of MAE. When Wilcoxon's rank sum test is performed for category A versus the combination of categories B and C (BC), the resulting p-value is indeed significant. This means the improvement in MAE arising from performing an adjustment for category A is on average higher than of category BC. It is thus safe to conclude that adjustments to category A are more effective than those to products in the other categories.

Hypothesis 2b is still not proven as the mean improvement for B is considerably lower than that of category C, which is supported by a Wilcoxon rank sum test ($V = 91,315$; $p = 8.83 \cdot 10^{-5}$). A mean improvement that is higher for category C than it is for B contradicts hypothesis 2b, as the test implies that an adjustment to category C will on average generate a bigger improvement than one done to a product in category A. Even though results presented in Table 8 and Table 9 look promising, hypothesis 2b cannot be deemed proven. Therefore, hypothesis 2b is to be rejected on the grounds of the improvement not being fully consistent with the product category importance.

6.2. Adjustment characteristics

Once more, the result section will start by examining a graph in order to visualize the data at hand. Figure 7 portrays information about the hypotheses that will be examined in this section and can give us an impression of the results that are to be expected from hypothesis 3 up to and including hypothesis 6. It contains the adjustment size on the x-axis and the forecast error on the y-axis. A negative adjustment size implies a downward adjustment, whereas a positive size implies an upward adjustment. The forecast error is depicted rather than the MAE, as to convey information about over forecasting in this image. MAE cannot do this, since it contains absolute values rather than positive or negative error values.

Finally, color is applied to express information about the data. In this graph, a green data point was an adjustment made in the correct direction, while a red data point was an adjustment made in the incorrect direction.



Figure 7: Forecast error for adjustments plotted against the size of the adjustment

Firstly, it is evident that there are more adjustments made in the correct direction rather than the incorrect direction. The graph is populated with far more green dots than red ones, indicating quite effective selection of direction. This provides evidence for sustaining H3.

Secondly, the amount of adjustments upwards seems to outnumber the amount of adjustments downward. Judging from the image, the right side of the y-axis is populated more than the left side is, implying a propensity to adjust upwards rather than downwards. This is exactly the phenomenon that H4a expects. Additionally, the forecast errors remain quite a bit more spread out for upward than downward adjustments, which could hint at H4b being supported as well.

Finally, the adjustment size could indeed be linked to the forecasting accuracy, especially for downward adjustments. The further the data points are removed from the y-axis, the closer the downward points seem to stick to the x-axis. This pattern does not transfer over to the upward adjustments, where the larger adjustments still carry hefty error. This would prove H6, but only when looking at the downward adjustments.

Now that a visual representation of the hypothesis testing has been established, the hypotheses will be examined in tabular form to numerically prove or refute them.

Hypothesis 3: Judgmental adjustments to statistical forecasts are more likely to be made in the correct direction.



The surface of this hypothesis has been scratched before, while investigating the preceding hypotheses. The variable that measures which forecast (statistical or judgmental) is closer to the eventual demand realization covers a considerable part of the implications conclusions on H3 can have. Naturally, the ability to outperform the statistical forecasting system relies for a great part on the ability to select the right direction for the intended adjustment. However, analyzing direction

additionally as a standalone factor in adjustment performance will provide more in-depth knowledge of forecasting performance.

Firstly, the figures reflecting the correct directions are summed in Table 11. It is quite clear that the ratio is in favor of the decision being correct. This observation is strengthened by Chi square test of independence, which shows the two variables are indeed not independent (H0 of the test is rejected) and the relationship is thus not based on sheer chance.

	Number of adjustments (%)
Correct direction	2172 (67.9)
Incorrect direction	1026 (32.1)
Statistical test	Chi-square test
Test statistic	$\chi^2 (1) = 411$
P-value	< .001

Table 11: Hypothesis test for correctness of adjustment decision direction

As the correct direction is selected in the vast majority of cases, namely for more than two thirds of the adjustments, this is evidence in favor of hypothesis 3. This evidence is useful, but more is required to irrefutably confirm H3. For this purpose, Table 12 has been created and was tested by means of a Chi-square test.

	Adjusted up (%)	Adjusted down (%)
Upwards required	1140 (70.0)	462 (30.9)
Downwards required	489 (30.0)	1,032 (69.1)
Test statistic: Pearson Chi-square	$\chi^2 (1) = 260$	$\chi^2 (1) = 217$
P-value	< .001	< .001
Test statistic: Pearson Chi-square	$\chi^2 (1) = 474$	
P-value	< .001	

Table 12: Hypothesis tests for required adjustment direction versus actual adjustment direction

The table above contains information about the frequencies with which the four combinations of outcomes appear within the dataset. Note that the entries in Table 12 do not add up to the totals in Table 7 and Table 8. This is due to the set-up of the underlying data from The host company and the calculations regarding the independent variables in Table 12. The dump files with which the dataset is constructed do not differentiate between a situation in which the statistical forecast has not been observed, and a situation in which the statistical forecast has been accepted implicitly (as discussed in 5.2.3). The dataset therefore does not contain any “acceptance” adjustments and thus no “acceptance required” category is included in the tests. This results in 75 observations being eliminated for this section of the thesis, which amounts to approximately 2% of the data points that are included in the preceding sections of hypotheses testing.

When evaluating the p-values for the upward and downward adjustment categories, it is evident that the probability that the correctness of the adjustment is not based on chance and that the results are highly significant. Additionally, it is interesting to look at the results for the contingency table as a whole. The p-value for Table 12 as a whole is highly significant as well, which means that null hypothesis of the Chi-square test of independence is rejected. This in turn implies that the selected direction and the required direction are related in one way or another. As the required direction of adjustment is an intrinsic value and the direction is selected by the forecaster, it is safe to conclude

that the forecaster is able to select the correct direction of adjustment and that this does not happen based on chance (randomly). With this, hypothesis 3 has been proven to be true.

Hypothesis 4a: Forecasters will be more inclined to adjust the initial statistical forecasts upwards than they will be to adjust them downwards.



For H4a, the number of upward and downward adjustments is required. Looking at Table 13, the results are quite comprehensive. Firstly, it is found that the required adjustments are very close to a 50/50 distribution. This is confirmed by the p-value of the Chi-square test, indicating the two directions are independent.

For the adjustments, a different conclusion is found. The amount of upward adjustments 9.7% higher than the amount of downward adjustments, which is in line with the expectations derived from literature. Additionally, a Chi-square test is performed on the frequency of the adjusted directions. The resulting p-value indicates that this difference is indeed significant, which implies an inclination to adjust upwards. This means that the probability of an upward adjustment is not equal to the probability of a downward adjustment.

	Required adjustments	Actual adjustments
Upward	1,602 (51.3)	1,673 (52.3)
Downward	1,521 (48.7)	1,525 (47.7)
Test statistic:	$\chi^2 (1) = 2.05$	$\chi^2 (1) = 6.78$
Pearson Chi-square		
P-value	0.152	$9.34 \cdot 10^{-3}$

Table 13: Hypothesis tests for adjustment direction propensity

H3 has shown that forecasters are able to identify the required adjustment direction. However, the ratio between the upward and downward adjustments does differ for the required and actual adjustments. Even though the difference is small, clearly the actual adjustments are more often made upwards than the required adjustments are. Consequently, H4a is accepted.

Hypothesis 4b: Downward adjustments are more beneficial for the forecasting accuracy than upward adjustments.



So far, it has become clear that adjustments on average will decrease the error associated with the forecast and that adjusted forecasts are more capable of approaching the value of demand actualization than the statistical forecast (H2a). Apart from producing information to validate claims from the literature review, H4b will provide a deeper understanding of a possible discrepancy between the two adjustment directions. Based on the optimism bias that has been found in previous research, and was discovered again in the analysis of H4a, it is expected that upward adjustments might not be as beneficial as downward ones and could even prove to be detrimental.

Firstly, the ability of the two directions to forecast the actualized demand is compared. It is found that an upward judgmental adjustment has forecast the exact demand realization 51 times, which 273 times for the downward adjustments. Keeping into mind that there are 9.7% more upward than downward adjustments, this result makes it very clear that downward adjustments are better able to

identify the correct future demand. It is important to note that the downward adjustments have a slight advantage here, as just over 200 products have a forecast value of one. This happens when a certain product for a certain client is on a regular sales basis of one box per week, but the statistical forecast has not reached this point yet (see 3.2.2). As this is a vital benefit of human adjustment to the statistical forecast, implementing information that the system does not have, these adjustment values will not be removed in the analytical part of the thesis. In the simulation section, they will be removed in order to increase the usefulness of the results proposed measures to simulate.

Next, the same approach as the analysis for H2a is applied. However, instead of expressing the best performing categories, the best performing direction is examined in Table 14. The values indicate the amount of adjustments that made the forecast better, so decreased the statistical MAE, or worsened the forecast and thus increased the initial statistical MAE.

	Adjusted up (%)	Adjusted down (%)
Adjustment improved	967 (57.8)	924 (60.6)
Adjustment worsened	706 (42.2)	601 (39.4)
Test statistic:	$\chi^2 (1) = 40.4$	$\chi^2 (1) = 68.0$
Pearson Chi-square		
P-value	< .001	< .001

Table 14: Hypothesis tests for the best forecast per adjustment direction

As Table 14 shows, the differences for both the upward and downward adjustments. As one can see quickly, the downward adjustments have a slightly better chance of being correct than the upward adjustments are. This would point toward the hypothesis holding up. However, the results are still pretty identical so further examining is required. Consequently, the MAE values for both directions are calculated and presented in Table 15.

Required adjustment direction	Adjustment direction	N (%)	Δ MAE	Δ MAE total
None	Upward	44 (1.38)	11.0	
Upward	Upward	1,140 (35.7)	-48.5	- 16.3
Downward	Upward	489 (15.3)	56.3	
None	Downward	31 (0.969)	5.52	
Upward	Downward	462 (14.5)	29.3	- 14.4
Downward	Downward	1,032 (32.3)	-34.5	

Table 15: Effect of correctness adjustment direction on MAE (boxes)

By studying the table above, it becomes evident that the values for the upward adjustments are more extreme than those for the downward adjustments. The correct adjustment has a higher average benefit associated with it, while selecting the wrong direction results in bigger loss of accuracy. The result of both adjustments is very similar, as the upward adjustments lead to an average improvement of 16.3 boxes and the downward adjustments to an average improvement of 14.4 in terms of MAE. These values are very close to each other and when tested with a Wilcoxon rank sum test, they are found to be equal ($V = 1.26 \cdot 10^6$, $p = .493$). This means that from a statistical point of view, there is no large enough distinguishable difference between them.

Looking at the variance of both directions, the variance for improvement values of upward adjustments is higher than that of the downward adjustments. This is an expected effect, as the values

that characterize the upward adjustments are more extreme. As a result of these extreme values, the adjustments that are mistakenly made in an upward direction will “pollute” the forecasting information more than the downward ones. This means that both adjustment directions are potentially beneficial, but do so in a less risky manner.

The hypothesis is therefore not completely acceptable in terms of the error measure reduction, but the ability to exactly forecast demand does make downward adjustments very useful. All things considered, the hypothesis cannot be proven to the largest extent and is rejected.

Hypothesis 5: Adjusted forecasts are expected to be consistently higher than the realized demand value.



For H5, an inclination to over forecast (to forecast higher than) the demand level at actualization is investigated. The research differentiates between correct and incorrect adjustments and how these relate to the phenomenon of over forecasting. This way, a more complete picture of this phenomenon will arise and moreover, a better understanding of the different variance levels found in research H4b can possibly be created.

First, initial expectations for over- and under forecasting conclusions can be drawn from a simple table. Below, Table 16 expresses the median and mean for the error values for both adjustment directions.

Direction	Mean error		Median error	
	Up	Down	Up	Down
Unadjusted	-58.3	26.6	-12	6
Adjusted	20.0	-13.9	2	0

Table 16: Bias characteristics for both adjustment directions

A few observations can be made from this table. Most notably, the ability to correct statistical forecasts in the right direction is quite visible here. Statistical forecasts that are adjusted upwards have a negative mean error associated with them, while the opposite holds for downward adjustments. Also, the forecast error is lower after adjustment than before adjustment.

The values of the median tell something about the bias that a certain forecast has. For the pre-adjustment forecasts, the same pattern is observed. The statistical forecasts that are later adjusted upwards turn out to be centered on values that are too low, as the negative median shows. Once again, the opposite holds for the downward adjustments.

When looking at the median values after adjustment, a surprising phenomenon is observable. The downward adjustments seem to be unbiased, as the distribution of the forecast errors is centered around zero. This means that, from a probabilistic perspective, there is no propensity to either over or underforecast. It is important to note that this does not mean downward adjustments are not unbiased, as the adjustment volumes might tell a different story. The probabilities however that a downward adjustment over- or undershoot demand are equal.

The median for upward adjustments is not centered around zero, implying a propensity to over forecast. Following the same logic as before, the magnitude of over- or under forecasting remains unknown. For upward forecasts however, it is clear that the probability of overshooting the demand is higher than undershooting it.

The absolute numbers in Table 17 tell the same story as the median and mean values in Table 16. The upward adjustments tend to be higher than the actualized demand value by a few percent. This had been derived from the median value as well, as it is not as large as the mean but still larger than zero. For the downward adjustments, the median error was zero and this can be explained by the percentages given here. The middle of the number line for forecast errors produced by downward adjustments is populated by the correct forecasts. Many zeroes therefore exist in the middle of the dataset and thus the median takes this value. However, the conclusion that an equal propensity to over- and underforecast is invalid, since the percentages do differ quite drastically.

	Total number of adjustments (%)	Adjustments upward (%)	Adjustments downward (%)
Overforecast	1,383 (43.2)	878 (52.5)	505 (33.1)
Correct	324 (10.1)	51 (3.0)	273 (17.9)
Underforecast	1,491 (46.6)	744 (44.5)	747 (49.0)

Table 17: *Over- and under forecasting characteristics for all adjustments in the dataset*

Even when the “easy wins” for the forecaster (where demand is lowered to one) are removed from the dataset, this would not impact the probabilities to over- or undershoot demand as these adjustments are correct. It appears as though it is safe to conclude that H5 does not hold for both directions. For the upward directions, it is evident that there is a tendency to overshoot the demand value, but it is more likely to undershoot demand when an adjustment is made downward. However, the data over under- and overshooting is polluted by the adjustments that are made in the incorrect direction. A subset of data that would include only the directions made in the correct direction could shed a better light on the phenomenon under investigation.

Table 18 is constructed from the subset of data from Table 17, for which the adjustments were actually made in the correct direction. The figures for the adjustments that have been made in the correct direction, show quite different results compared to the preceding table. For both the upward and downward adjustments, the propensities have switched. An upward adjustment in the correct direction is now expected to undershoot demand, as opposed to downward adjustments that are now more likely to overshoot. If the easy-wins for downward adjustments are removed, both directions have similar results for over and undershooting the demand.

	Total number of adjustments (%)	Adjustments upward (%)	Adjustments downward (%)
Overforecast	850 (39.1)	345 (30.3)	505 (48.9)
Correct	324 (14.9)	51 (4.47)	273 (26.5)
Underforecast	998 (45.9)	744 (65.3)	254 (24.6)

Table 18: *Over- and under forecasting characteristics for correct adjustments*

The pattern emerging here is therefore one of adjustments that are not too extreme. The forecaster is “hedging” the adjustment against the possibility of the adjustment being in the wrong direction. As a result, the forecasts that are in the appropriate direction are not able to reach the correct target and generally fall short of the realized demand. This means that there indeed is a propensity to over forecast when adjustment is made in the downward direction, instead of for the upward direction. Overall however, the over forecasting pattern does not hold and H5 is consequently rejected.

Hypothesis 6: Large adjustments are more effective for decreasing forecasting error than small adjustments.



In the literature study, a relationship between the size of an adjustment and its resulting accuracy is hypothesized. Upon inspection of Figure 7, this relationship becomes quite debatable for the upward adjustments. For the downward adjustments on the other hand, the pattern seems to hold as the largest adjustments are less far apart from each other vertically than adjustments closer to the y-axis. The adjustment errors for upward adjustments are spread out more over the entirety of the corresponding part of the graph, but these forecasts may have possessed high statistical error values before. Therefore, additional graphing is required to investigate the expected pattern of H6.

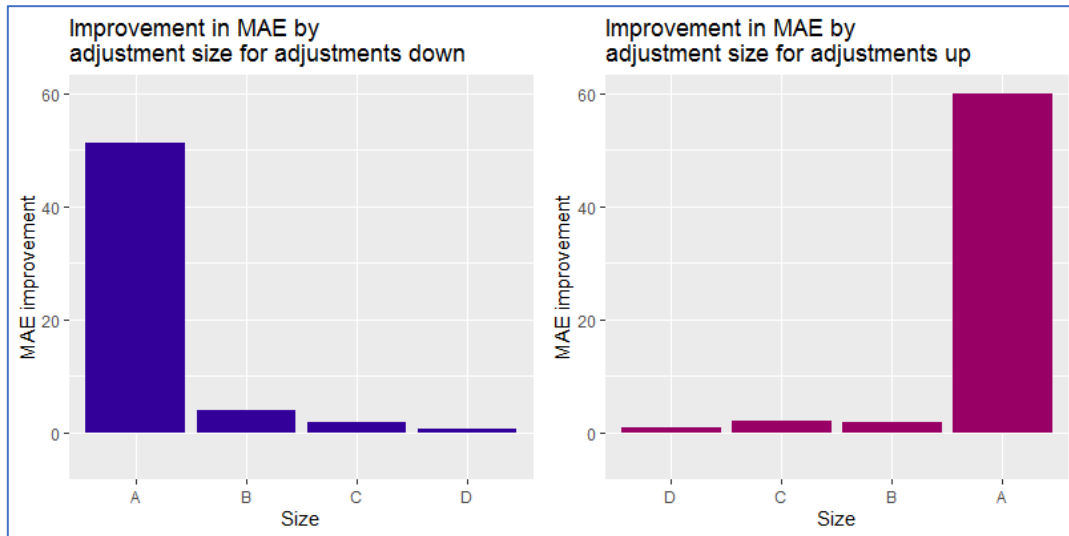


Figure 8: Effect of adjustment size on the improvement in MAE

To provide evidence for H6, Figure 8 was plotted. The graphs in this figure express the improvement that the respective size quartiles result in. In both graphs, the largest adjustments are located in quartile A and the smallest in quartile D. The graphs show a clear trend that for the largest 25% percent of adjustments in a certain direction, the largest improvements are found. Not only was this result expected from the literature study that suggests larger adjustments are accompanied by higher certainty from the forecaster, but also by simply mathematical deduction. The maximal improvement that an adjustment can deliver is equal to the adjustment's size. An adjustment of size ten cannot increase the accuracy by more than ten units. This means that all improvement values are equal to or lower than the adjustment size that they are accompanied by.

With this last piece of information in mind, it becomes obvious that large adjustments will be more effective in reducing forecasting error. They simply are able to reduce more error, as their magnitude is larger. Additionally when looking back at Figure 7 once again, it is distinguishable that large adjustment values are generally more correct than smaller ones. The further away from the y-axis, the fewer red dots are spotted. This means large adjustments do not only possess larger potential improvement, they also seem less likely to be incorrect.

	Downward adjustments				Upward adjustments			
Quartile	QA	QB	QC	QD	QD	QC	QB	QA
Direction correct	69.8%	67.4%	68.9%	64.6%	67.3 %	70.0%	64.8%	70.3%

Table 19: Percentage of correctly selected adjustment direction per size quartile

This visual observation is supported by the numerical evidence in Table 19 that shows that larger adjustments quartiles are more likely to contain adjustments in the correct direction than the smaller quartiles do. Both the A quartiles have the largest proportions of adjustments in the right direction, with 70.3% for upward adjustments and 69.8% for downward adjustments. This is larger than the other quartiles, which contributes to the higher improvement values for these quartiles. Combined with the larger adjustment size for these quartiles, it is easy to see that adjustments made in the largest 25% of the both directions are more efficient than those to the smaller quartiles. Considering all the figures presented, H6 is confirmed.

Hypothesis 7: Judgmental adjustments will be increasingly beneficial for the forecasting accuracy, the closer they are made to the moment of demand realization



The first step in investigating the final hypothesis has already been already taken in the beginning of the previous section. Figure 6 clearly shows a decline in the value for MAE, which looks promising. On the other hand, the evolution of the line of the statistical forecast shows a similar downward trend the closer it gets to the demand actualization. After all, an adjustment is based on the statistical forecast and the decline in MAE for adjustments closer to the demand actualization could thus be due to consistent adjusting over time. In order to discover the hypothesis' truthfulness, the improvement value should once again be evaluated alongside the regular error measures.

For the creation of the graphs below, the full set of adjustments is used. This set has been corrected in a similar fashion to the four-week lag set, by removing all error values that exceeded a Z-score of three. When assessing the graphs below (Figure 9), the means of the judgmentally adjusted MAE per forecast lag follows a linear trend. This is confirmed through a linear regression over all the data points, which finds that the forecast lag provides a good fit for the means of the adjusted MAE ($R^2 = 0.84$, $p < .001$). One should keep in mind that these values explain the relationship between the means of the forecast lags, not between the individual adjustments and the forecast lags. Still, a noticeable decrease in adjustment error can be detected.

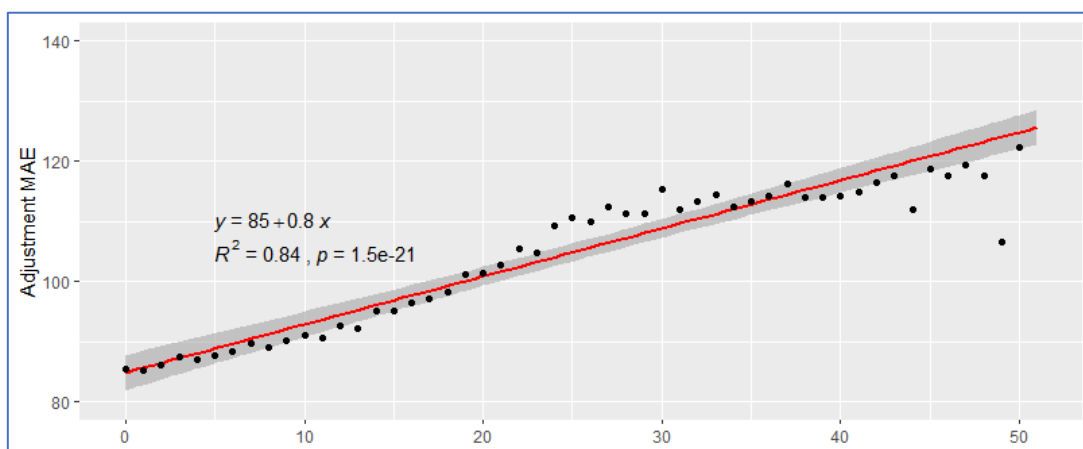


Figure 9: Mean MAE values for all forecast lags with linear regression

Next, the values for the forecast improvement values are evaluated by means of regression. For the improvement versus the forecast lag, a loess regression was applied to see if a relationship between the variables might exist. A type of third order polynomial was expected to exist for the improvement data as the loess regression has a shape similar to this kind of expression.

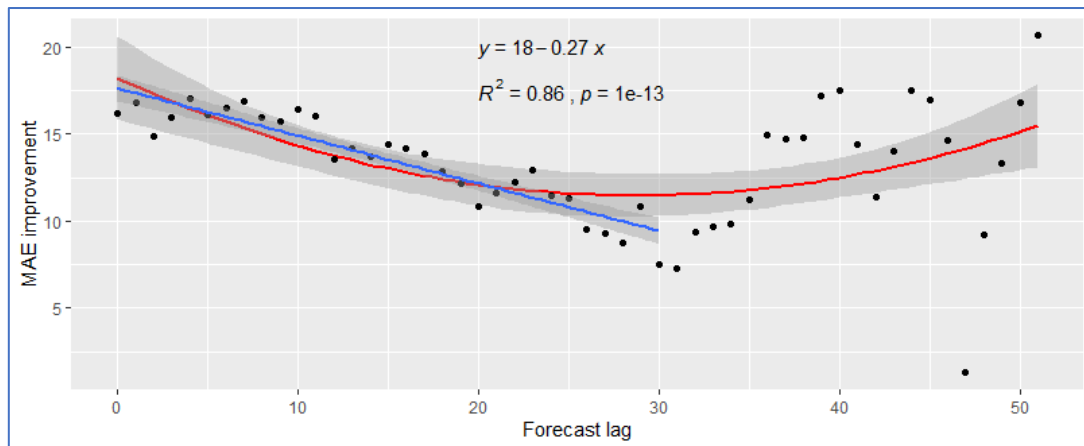


Figure 10: Mean MAE improvement value for all forecast lags with a loess regression

After checking the regression fit for the forecast lag versus the improvement in MAE for several expressions, the third order polynomial (of kind $y = x + x^2 + x^3$) did not provide the best fit. The final model to fit the means of the forecast improvement is presented in Figure 10. The R^2 value for the polynomial is low at 0.288, which is most likely due to the noise from the middle to the right of the observations. If the noise is disregarded and only the weeks with a lag of thirty weeks or lower are included, the resulting regression line provides an excellent fit. As the dump files span either the first or second half of a year, the forecasts are more meaningful. The R^2 value is high and the variable is highly significant. These results imply an increasing value for the improvement an adjustment offers the closer it is made to demand realization.

Unfortunately, the dataset does not allow for analysis of sequential adjustments to a single demand instance. This means that no statements can be made about the effects of changing a forecast after it has been made earlier. Judging from the image above, one would expect that these adjustments to earlier forecasts are more accurate as they are made closer to demand realization. This can however not be confirmed.

The evidence presented in this section leads to confirming the seventh hypothesis. Not only does the MAE decrease over time, the improvement an adjustment offers increases over time as well. This is the result of more accurate forecasting decisions and thus H7 is confirmed.

6.3. Redesigning the forecasting procedure

With the results uncovered in the two previous sections, implications can be derived and a step towards a more accurate and efficient forecasting procedure can be made. Looking at the hypotheses, it is possible to establish a pattern of where adjustments are most beneficial and where adjustments provide less added value. This will result in different forecasting settings that will be simulated in R, in order to identify a forecasting strategy that provides the most benefits in terms of forecasting

efficiency (less time required to forecast, or more effective use of time applied to forecasting) and forecasting error. The different situations are explained below and summarized in Table 20.

6.3.1. Avoiding small adjustments

In this simulation, the result of hypothesis 6 will be translated into a set of adjustment rules that prescribe no small adjustments are allowed to be made anymore. As Figure 8 shows, the largest 25% of adjustments in both upward and downward direction provide the most benefits when an adjustment is made. The smaller adjustments carry less improvement potential, resulting in a less efficient decision as they deliver less reduction in error per adjustment. In order to take full advantage of the benefits large adjustments have to offer, a situation in which small adjustments are discontinued will be simulated. A cut-off value will be based on the percentile range of the adjustment size histogram, to exclude all unnecessary adjustment size values.

The downside to this cut-off value could be a form of gaming when the cut-off value demand is not fulfilled, in which the forecaster will weigh the potential error of not adjusting and the error if the adjustment is too large. The result could be artificially inflated adjustments to satisfy the cut-value and make the adjustment possible. This might become problematic if this model is implemented in a company setting.

The result of the simulation and its analysis will be a comparison of the total error before and after excluding the smallest adjustment sizes and how the efficiency of the system has changed, as well as a comparison with the other simulated situations.

6.3.2. Reducing the erroneous direction adjustments

In Fildes et al. (2009), a simple but seemingly effective approach is applied in an attempt to increase the usefulness of judgmental adjustments. It is based on the theory that it is in fact possible to eradicate 50% of the wrong-sided adjustments by means of more thorough market research. In the research, the adjustments made in the wrong direction were allocated to either a 'change' and a 'no change' group on a random basis. As a result, 50% of the incorrectly sided adjustments kept their adjustment (change group) and the other 50% had its judgmental adjustment value removed (no change group). This implies that the statistical forecast was applied for these adjustments.

The same approach will be adapted in this research, to see if the potential benefits can also be detected here. As the observations that will be deleted are erroneous, the expected average improvement value will increase. The results of this simulation will be compared to the current situation, as well as to the other simulations in order to review the new forecast error and how well this fares to the alternative options.

6.3.3. Blattberg-Hoch model

A third option to decrease the error associated with judgmental adjustments is the Blattberg-Hoch model (50/50 heuristic, see paragraph 4.2). This model will be assigned to all the adjustments made, in order to test if it possesses any benefits for the set under review. The algorithm will construct a mean of the statistical forecast and the judgmental adjustment and check if this results in a lower forecast error. Theoretically, this approach is helpful in that it removes any extremely large adjustment values by averaging it with the statistical forecast. In a situation where over forecasting due to human intervention is an issue, this approach would be beneficial by reducing the impact of the adjustment. But as the research has already shown, there is no structural problem of over forecasting present. Judgmental adjustments are being hedged (as shown in the analysis of H5), so the model is not used to its full capabilities in this case.

However, by averaging the adjustment with the statistical forecast, adjustments in the wrong direction are also hedged by this algorithm. Needless to say, using 50% of a mistake is reducing the damage it causes substantially. Therefore, the Blattberg-Hoch model situation will still be in this research as it can be applied as an easier to apply solution than the reduction of erroneous direction adjustments. Moreover, this model can be applied no matter what and does not need marketing innovation investments that the previously presented method would require.

One should note that in this forecasting environment, the 50/50 name is not completely valid. The adjustment and the statistical forecast are not independent of one another, as the statistical forecast acts as an anchor for the adjustment. Fildes et al. (2009) represent this by the following equation:

$$0.5 \cdot (\textit{System forecast}) + 0.5 \cdot (\textit{System forecast} + \textit{Adjustment}) = \textit{System forecast} + 0.5 \cdot \textit{Adjustment}$$

Equation 1: Blattberg-Hoch formula as described by Fildes et al. (2009)

This clearly demonstrates the dampening of the adjustments that occurs as a result of the Blattberg-Hoch model, which is most beneficial in cases of adjusting with too large volumes and wrong-sided adjustments. The results of this simulation will be compared to the current situation and, more interestingly, to the performance of the wrong-sided direction reduction model.

6.3.4. Upward indication model

The final forecasting situation that will be examined concerns the upward indication model. It is based on the principle of replacing a part of the forecaster's responsibility with an algorithm. The algorithm will ask the forecaster to indicate in which direction the adjustment is going to have to be made. If the forecaster indicates the adjustment is to be made upwards, the algorithm will generate an expected demand value based on known characteristics of the product and the statistical forecast as calculated before. If the adjustment has to be made downwards, the algorithm will not intervene and leave the adjustment task to the forecaster.

This model is based on the results of H3, which suggest that forecasters are well able to identify the required direction for adjustments, and the results of H5 that suggest forecasters have trouble forecasting exactly the actualized demand value. The set-up of this model has some useful benefits for implementing it in a business environment.

Firstly, the model only has to calculate expected demand values in case of an upward adjustment, which reduces the complexity of the model and consequently increases its ease of use and comprehension for employees.

Secondly, the loss of control experienced by the human forecaster is limited as fewer tasks are transferred compared to the previous model (as discussed in 3.3). The best properties of human judgmental adjustments are still applied in this setting, as H3 has shown that downward adjustments are a safer means of forecasting than upward adjustments. Algorithm aversion is potentially small in this variant, as the forecaster still performs a large part of the adjustment task (Arvan, Fahimnia, Reisi, & Siemsen, 2018). The human forecaster can thus work at full potential in this model and is relieved from less successful efforts by the algorithm. The model equation for adjustments is as follows:

$$\begin{aligned}
 \text{Final forecast} &= [(1 + x) \cdot \text{Statistical forecast}] \cdot y_u + \text{Judgmental adjustment} \cdot y_d \\
 y_u + y_d &= 1 \\
 y_u, y_d &\in \{0,1\}
 \end{aligned}$$

Equation 2: Formula applied for calculating the forecast value in the direction indication model

Three variables are present in Equation 2: x , y_u and y_d . The first variable is x , which expresses the percentage of increase applied to the statistical forecast as a decimal. The other two variables y_u and y_d are binary variables that determine which part of the equation is used as the final forecast depending on the direction of the adjustment. If an adjustment is made upward, y_u becomes 1 and the judgmental adjustment is ignored as y_d becomes zero. The opposite occurs when a downward adjustment is made. The statistical forecast multiplied by the increment value is rounded upwards, to prevent decimal amounts of boxes.

All models put forward in this section will be compared to the other models presented on the grounds of accuracy of results and applicability in a company setting, to come to a final conclusion on the best intervention method for increasing the forecasting performance.



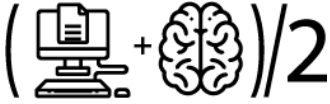
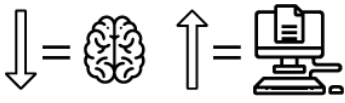
	 No small adjustments	 Reduce wrong-sided adjustments by 50%	 Blattberg-Hoch model	 Human forecaster for direction and downward adjustments
Focus	Make adjustments only in large quantities and thus avoid the smallest of adjustments	Extensive market research efforts to reduce wrong-sided adjustments by 50% by increasing knowledge	Reduce adjustment impact by averaging it with the forecasting value that it was based on	Let the human forecaster adjust downwards and indicate upward direction when it is required
Advantages	Capitalize on the higher payoffs for large adjustments and abandon less beneficial, less efficient and more risky, small adjustments	The market research will increase forecasting skill by providing detailed and profound insight into customer needs and market development	Forecasts adjusted with a volume too large will be corrected by the statistical forecast by assigning equal weight in determining the expected demand	Downward adjustments are helpful, so leaving this for the human forecaster increases the efficiency of work and asks for little transfer of control
Disadvantages	This approach could encourage “volume gaming” to minimize possible demand loss if an adjustment is smaller than the cut-off value	The targeted reduction might not be reached if the market research does not obtain the desired effects, meaning wasted effort, time and money	In case of hedged or safe adjustments, the model could dampen the positive effects of the adjustments as much as the negative ones	Like the previous model, though to a lower degree, this approach requires complex calculations that can hurt acceptance and understanding by staff

Table 20: Summary by focus, advantages and disadvantages of the proposed adjustment simulations for the research

7. Execution and evaluation of simulations

Chapter 7 describes the method of execution of simulations performed for the situation sketched in section 6.3, Table 20. After explaining the construction of the simulation, the section devoted to the particular model will then show the results and indicate how meaningful these results are.

7.1. Adjustment size cut-off value

The first model that will be considered revolves around avoiding the smallest adjustments present in the current situation. Fildes et al. (2009) allow only adjustments larger than a fifth of the statistical forecast in their model, in order to assess the potential increase in forecasting accuracy and the decrease in time spent on forecasting. Similar to their research, a set of small adjustments will be removed as to weigh the gains against the losses.

Figure 8 has shown that the adjustments that correspond to the largest 25% of the adjustment size for a direction on average bring about the largest accuracy improvement in terms of MAE. The smallest 25% of adjustments contribute very little to this goal, which raises the question if they are possibly redundant. Therefore, the smallest 25% of adjustment sizes are removed from the set. Removing adjustments is done for the entire set of adjustments, and for the two directions separately.

	Complete set	Remove 25% of total	Remove 25% up	Remove 25% down
N	3198	2396	1225	1141
Cut-off value		5	8	5
Absolute adjustment size	60.2	79.4	106	53.2
Statistical MAE	97.6	125	147	103
Adjusted MAE	82.2	105	125	84.5
Improvement in MAE	15.4	20.3	21.9	19.1

Table 21: Error measurement values for disallowing small adjustments to statistical forecasts

The results of removing the smallest 25% percent of adjustments are presented in Table 21. Due to the removal algorithm's design, both sets have removed slightly more than 25% of observations. This is 25.1% for the test set that removes from the total set of observations, while the second test set has removed 26.8% of upward and 25.2% of downward adjustments and thus totals 26.0%.

Table 21 shows that the average improvement for the reduced sets is larger than for the complete set. Additionally, the table shows that larger adjustments are performed to forecasts that possess a higher intrinsic error value. In other words, there statistical forecast is less accurate for these observations. Possibly this phenomenon could be explained by more volatility in the demand values for the subset of products (Sanders & Ritzman, 1992), but cannot be proven due to the structure of the dataset.

Removing 25% of the total set results in an average improvement of 20.3 in MAE and a corresponding total of 48,622. Separating the two directions and removing the smallest 25% results in an average improvement of 20.5 and a total improvement of 48,590. The difference between removing the smallest 25% of adjustments sizes from the total set, or from the two adjustments directions separately thus influences the average and total MAE improvement little.

The total improvement value resulting from the judgmental adjustments is 49,285 in terms of MAE for the original set, which is close to the improvement totals of the reduced sets. Removing from the entire set has thus resulted in a decrease of 663 units, or 1.35%, in total improvement value. For the separated removal approach, the decrease is 695 or 1.41%. This implies that by removing 25% of adjustments, the accuracy has only decreased marginally.

The effect of removing the smallest adjustments is slightly more beneficial for downward as it is for upward adjustments. The full set of upward adjustments provided an improvement of 16.3 in MAE, while this improvement is 14.4 for the downward adjustments Table 15. This means that by the removal algorithm described above, upward adjustments have become more accurate by a value of 5.6 in MAE against 6.2 for the downward ones.

7.2. Avoiding wrong-sided adjustments

Figure 7 already depicted impact of wrong-sided adjustments have on the forecasting error. Every unit adjusted in the wrong direction immediately adds to the forecasting error, wasting the time and effort that was invested into the adjustment. If these wrong-sided adjustments can be avoided due to some kind of investment, this would improve the forecasting performance drastically. As proposed by Fildes et al. (2009), half of these adjustments will be removed and the gains will be assessed. The analytical results, presented in Table 22, confirm the anticipated effect.

	Complete set	Avoid 50% of wrong-sided adjustments
N	3198	2685
Absolute adjustment size	60.2	64.4
MAE statistical forecast	97.6	100
MAE adjusted forecast	82.2	74.6
Improvement in MAE	15.4	25.7

Table 22: Error measurement values when 50% of wrong-sided adjustments is removed

The deletion algorithm picks a random set of observations from the subset of wrong-sided adjustments. Then, the observations that are not removed are added back to the set of adjustments that are right-sided.

It becomes quite evident that when half of the wrong-sided adjustments is removed, the improvement value of an adjustment is increased. The removal of these wrong-sided adjustments was performed on a random basis. On average, the improvement per adjustment is increased by 10.3 units. Thus, by deleting only 16% of the observations, the average improvement per adjustment has risen noticeably.

The relatively higher payoff for the model presented in this section compared to the previous was expected. Naturally the impact of a wrong-sided adjustment is always negative, while small adjustments in the correct direction can offer value to the system. Therefore, the increased improvement for the model in Table 22 is explicable. It should be considered that one cannot know beforehand whether an adjustment will be in the wrong direction or not. However, the results do show the magnitude of reduction in forecasting error if erroneous forecasts can be avoided. By avoiding adjustments to products that have an above average propensity to be wrong-sided, the average effectivity of adjustments could be improved.

7.3. Combining statistics and judgment directly

7.3.1. Blattberg-Hoch model

The third proposed model is one by Blattberg and Hoch and is based on assigning equal weight to the judgmental forecast and the statistical forecast (Blattberg & Hoch, 1990). This implies the average of the two is applied as the final adjustment. This model will reduce the impact of the judgmental adjustments and thus decrease the negative effect wrong-sided adjustments. The results for this Blattberg-Hoch (BBH) model are given in the table below.

	Complete set	Blattberg-Hoch
N	3198	3198
Absolute adjustment size	60.2	30.1
MAE statistical forecast	97.6	97.6
MAE adjusted forecast	82.2	83.8
Improvement in MAE	15.4	13.8

Table 23: Error measurement values for BBH model

After assessing the outcomes for the BBH model, it is apparent that it does not raise the improvement value of the adjustments. On the contrary, the improvement value drops as a result of the implemented model. To better grasp these results, the formula from 6.3.3 is reexamined in a more generalized form, which is as follows:

$$\text{Final forecast} = [x \cdot \text{Judgmental adjustment} + (1 - x) \cdot \text{Statistical forecast}]$$

Equation 3: General form of Blattberg-Hoch model formula for introducing different weights

The value of x in the formula above is the weight assigned to the judgmental adjustments (the expected demand value as indicated by the planner) in determining the final forecast. The rest of the weight, which can be between 0 and 1, is assigned to the statistical forecast. This model is a type of smoothing technique. It reduces the magnitude of adjustments by smoothing them with the statistical forecast, which is the arithmetic mean of the previous 20 periods of demand.

The original setup for the BBH model, provides no accuracy improvement as compared to the current situation. However, the average adjustment size has decreased which does provide the nice bonus of less deviation of the statistical forecast line. This means there is less variation in the final forecast and thus less variation for the production department to incorporate in their production planning.

By decreasing the weight put on the judgmental adjustment, the improvement value of the adjustment has subsequently decreased. A weight of 0% assigned to the adjustment, will result in an improvement of zero, as the final forecast is simply the statistical forecast. No change is made, so no improvement can occur.

Assigning a weight of 100% to the judgmental adjustment, which is identical to the current situation, results in the improvement of 15.4 boxes as indicated in Table 23 most recently. The 50% weight assignment to the judgmental adjustment returns an improvement of 13.8. Therefore, the improvement value does not have a linear relationship with the weight allocated to the two components of the final forecast. Plotting this formula does indeed return a curved shape, which is depicted in Figure 11.

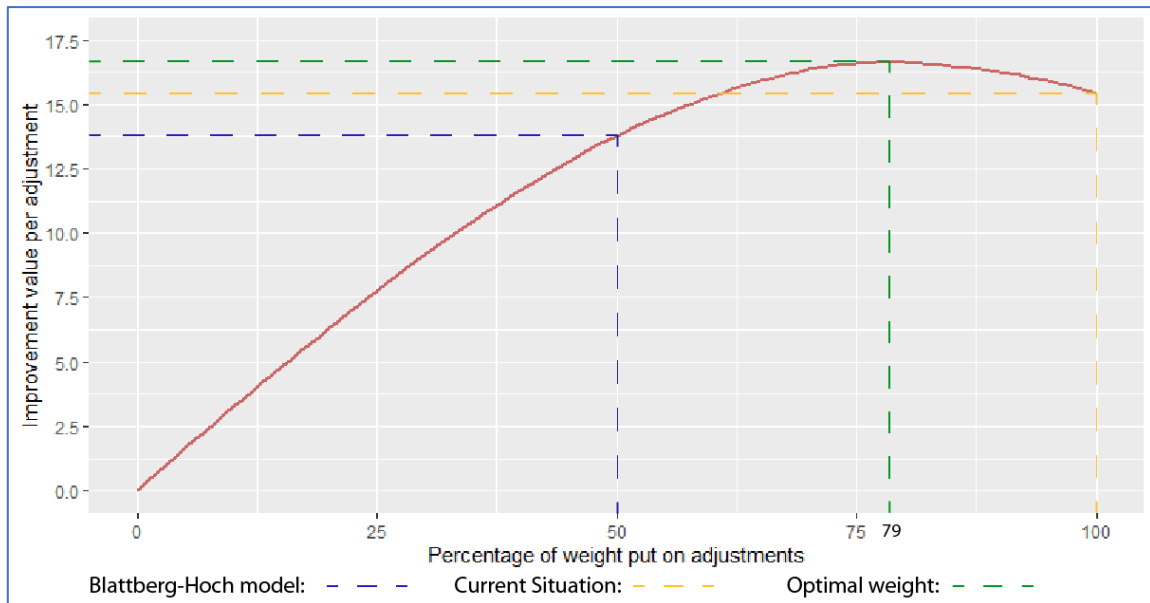


Figure 11: Improvement values plotted by the corresponding weight assigned to the judgmental adjustment

Looking at the plotted figure in Figure 11, there is a clear paraboloid shape detectable. This implies that there is an optimal value for the weight that is assigned to the adjustments. This optimal weight is calculated with R and turns out to be an x-value of 79. This means that rather than the 50% proposed by the BBH model, 79% would be the optimal weight to assign to the judgmental adjustment and thus to the adjustment size. The statistical forecasts contributes 20% to the final forecast, by which an improvement value of 16.6 in terms of MAE is realized. This optimal value balances on the edge of reducing the effect of over forecasting, and the problem of shrinking the size of adjustments that are too small.

7.3.2. Different weights for different categories

In the preceding section, the possibility of assigning a weight to the judgmental adjustment has been identified. In order to maximize the effect of this approach, the weights could be specified for subgroups within the dataset. This in turn should lead to higher judgmental adjustment improvement values. The weights will be based on a learning set and then applied to a simulation set to see if the results hold for new data.

Firstly, the learning set is created from the set that has been applied previously. This set will start with the 3,198 observations used previously, and subtract the 220 perfect downward adjustments discussed in section 6.2. These observations are removed, as they are generated by a single customer and are all adjusted to an expected demand value of 1 that occurs every time. This means they have no predictive value, as their behavior is different from regular demand forecasting. Then, 20% is removed from the set to be used later as the testing data. As a result, 80% of the subset is applied for learning purposes.

An optimal weight will be based on the total set of observations in the learning set. Additionally, separate weights will be determined for different grouping categories, such as the article number and adjustment direction. The grouping of observations will allow for more effective results, as different categories have different properties. If a grouping category contains less than 5 observations, it will be removed from the learning set. A weight would in that case be based on very little information,

which also has little predictive value and is therefore removed. If a group in the simulation set has no optimal value, i.e. it was not in the learning set, a weight of 100% will be assigned. This is equivalent to the current situation.

Finally, the simulation adjustments are corrected with the optimal weights for the judgmental adjustment. These observations are the 20% of data that was removed from the cleaned set earlier. After correcting the adjustments, the additionally gained forecast accuracy is calculated and reported.

7.3.3. Simulating the forecasts

For the learning set, the optimal weight for the judgmental adjustments is found to be 76%. This means that weight given to the statistical forecast is 24%. The corresponding average improvement in accuracy per adjustment is 1.60 in terms of MAE. This value is the average additional forecast improvement that is obtained through implementation of the optimal weight, compared to the current procedure of making judgmental adjustments.

Splitting the learning- and simulation set over different products, does not improve all adjustments. The simulation set contains 94 products for which adjustments have been made, 51 of which have 5 or less observations in it. The average added value of the model is -.683 per adjustment in terms of MAE. The mean increase per group however is .418, with 69 groups showing a nonnegative improvement as a result of applying the weighted judgmental adjustments. This means that for the majority of simulated groups the improvement is either zero or larger, but these groups contain a small set of observations as the change per adjustment is negative.

Removing the simulation groups that have less than 5 observations, results in a simulation of $94 - 51 = 43$ groups of which 30 have a nonnegative change. The average change now is -1.06 per product group and -0.535 per individual adjustment in units of MAE. This implies that some groups with a large model improvement value have been removed, and thus the average drops below zero.

The average change per adjustment has increased, which is the result of product groups with a large, negative model outcome also being removed.

The simulation- and learning- set can also be divided into groups of clients. There are 13 client groups in the simulation set, of which one has fewer than 5 simulation values. When the results of the 12 simulation groups with weights assigned to specific clients are assessed, the change in accuracy is positive. For 9 groups, the change is larger than zero with an average of 1.71 in terms of MAE per group. The weighted mean of all adjustments is even 2.65, implying that the client groups with positive model change values are larger than the smaller ones.

When the set is separated into upward and downward adjustments, the model improvement is found to be an average of 2.05 per adjustment. This is a result of an improvement of 3.35 in terms of MAE for upward adjustments and .0697 for downward adjustments. The model improvement for the downward adjustments almost equals zero, as the weight assigned to downward adjustments is 99%. This means that the algorithm leaves downward adjustments almost completely to the forecaster, but corrects upward adjustment by a weight of 72% to arrive at a more accurate forecast. A model applying a similar approach will be examined in the next section.

7.4. Indicating direction and downward adjustment size

The previous model showed that for upward adjustments, not allowing the full size of the upward adjustment increases the accuracy. Simultaneously, the model in 7.3.3 allows the forecaster to almost fully control the size of the downward adjustment. In this section, a similar solution will be explored. The forecaster is still allowed to determine the size of the downward adjustments, but will only be able to indicate if an adjustment is to be made upward or not. The size of the adjustment is to be left to an algorithm, that balances the effects of over- and underforecasting. This approach will be applied to the whole set of adjustments as well as to sections of the set, which is similar to the previous section.

Firstly, finding an average percentage for all adjustments upward does not produce an improved result. An optimal forecast improvement is found for an upward adjustment value of 6%. At this value, the model reduces the forecasting accuracy by 3.08 in terms of MAE. This means that forecasts actually are less accurate if this model is applied.

The same pattern is found when product subsets are created for the simulation set. For every product, an optimal upward adjustment percentage for adjusting upwards is calculated. If the simulation contains products that were not in the learning set, the original judgmental adjustment will be kept. This is what would be most realistic, as the system has no information on the required average upward adjustment. In this case, either the average optimal percentage for the whole set could be applied. However, in this case the upward adjustment is accepted completely as it is closest to the current forecasting procedure.

If all simulation groups are allowed in the simulation, the mean increase per group is .0934 in units of MAE, but -3.57 for the arithmetic mean of all adjustments. In line with the approach in 7.3.3, the simulation observations smaller than 5 are removed. The mean increase in forecast improvement per group now is a decrease of 4.44 in terms of MAE, while the average accuracy of all adjustment decreases by 4.94.

Finally, grouping is done on a client basis. The model output once again is negative compared to the current forecasting procedure, with a decrease of 9.90 for the average per group and a decrease of 9.65 for the arithmetic mean in terms of MAE. Disallowing the small simulation group does not help this grouping procedure, as the average per group drops to a decrease in forecast improvement of 12.0 in units of MAE. The arithmetic mean of all adjustments remains comparable, with a decrease of 9.65 per adjustment.

7.5. Comparison of simulation output

In sections 7.3.3 and 7.4, simulations have been performed for different change propositions to increase the forecasting accuracy of the host company's forecasting procedure. In this section, the results will be tabulated for a comprehensive overview.

Grouping	Variable weight model		Upward indication model	
	Group average	Arithmetic mean	Group average	Arithmetic mean
None	x	1.60	x	-10.2
Direction	1.74	2.06	x	x
ABC	.921	1.48	-7.58	-9.66
Client (all)	1.58	2.65	-9.90	-9.65
Client (all group sizes > 5)	1.71	2.65	-12.0	-9.70
Product (all)	.417	-.683	.0934	-3.57
Product (all group sizes > 5)	-1.06	-.535	-4.44	-4.94

Table 24: Forecast improvement in units of MAE as compared to the current forecasting procedure for variable weight model and upward indication model

As shown in Table 24, the model effectiveness of the variable weight model presented in 7.3.3 is higher for all categories compared to the upward indication model. With the exception of 1 category, all simulated results have a negative model outcome for the upward indication. This is in stark contrast to the variable weight model, that has generally has positive model outcomes.

When assessing the results for the variable weight model, the highest improvement values are found for setting the weights by direction or by client. Grouping by the ABC categories also provides additional forecasting accuracy, but is outperformed on both metrics by the previously mentioned groupings and by the non-grouped alternative. The model benefits for grouping by product are lowest, as the results are generally negative.

The upward indication model has the highest results for the product based adjustment percentage, even though these results are mostly negative. The other categories generate results very similar to each other, all reducing the forecast improvement that the judgmental adjustments provide. It is therefore safe to conclude that, based on the results in Table 24, the variable weight model provides a more beneficial addition to the judgmental forecasting process than the upward indication model does.

8. Discussion

This chapter of the thesis is devoted to the discussion of the results produced by the research presented in this thesis. It will consider the implications of the research and introduce recommendations for the host company. The limitations of the research are discussed and possibilities for future research are derived from the implication and limitations.

8.1. Implications

The thesis has presented the results of its research in chapters 6 through 8, by investigating the truthfulness of the various hypotheses put forward in chapter 4. By doing so, some claims based on the literature study have been proven, while others have been rejected on the basis of too little evidence or disproven. Both by proving and disproving claims, the thesis brings about some new knowledge, both in the realm of scientific literature and in that of business. The implications for both types of knowledge will be discussed here, accompanied by comments about the rigidity of the results.

8.1.1. Pragmatic implications

As intended in the pragmatic objective, the research has uncovered new information about the forecasting process within The host company. This information is mostly linked to the reduction of forecast error that is achieved by adjusting system-generated forecasts and by which characteristics of the adjustments the reduction of error is increased or decreased. The most important implication for the logistics management staff, is that the thesis shows that the judgmental adjustments do in fact increase the accuracy of the forecasting process. By improving the forecast quality, the discrepancy between expected demand and demand actualization is reduced. This in turn saves error related cost like extra holding costs and lost sales.

Moreover, it was discovered that the forecasts on average are in the correct direction and that large adjustment sizes bring about larger forecast improvements. Simulations were set out to capitalize on these phenomena. Firstly, the amount of small adjustments present in the set of adjusted forecasts was reduced by removing the bottom 25% of adjustment sizes. This resulted in a set that was 25% smaller in the number of adjustments, the total forecast improvement in MAE over all adjustments decreased by 1.41%. This implies a low improvement in forecast accuracy is associated with the smallest adjustments. Unfortunately, adjustment size is an unknown factor before the adjustment is made. Hence, it is recommended to find an approach to judgmental adjusting that is able to identify products for which the adjustment's size is likely to be small. By disallowing these adjustments to be made, time is freed for the forecaster to invest in other, larger adjustments.

A downside could be that in order to accomplish the forecasters' envisioned result, the adjustment is artificially inflated. This would involve raising the adjustment size to a point that it would reach the minimal required adjustment size. Proper education on the purpose and outcomes of the intervention could help towards attaining more understanding for the rule and insight into how it increases an adjustment's effectiveness.

Secondly, a simulation was performed to see if the indication of adjustment direction and the statistical forecast together would allow an optimization algorithm to find a suitable adjustment value. This adjustment value would be a certain percentage of the statistical forecast being added to the statistical forecast in case of an upward adjustment. For downward adjustments, the forecaster would retain the full adjustment responsibility. This approach captures the ability of forecasters to identify

the required direction of adjusting and decreases the time a forecaster has to invest in the forecasting task. Unfortunately, the direction indication model did not perform well. All model improvement values except one were negative. This means that the model did not increase the added value of judgmental adjustments. This is most likely due to the variability of the forecast sizes and the existence of wrong-sided adjustments.

The wrong-sided adjustments have been proven to be less common than correctly-sided adjustments, but still are a large source of forecast error. Decreasing the presence of wrong-sided adjustments would directly increase the average forecasting improvement, as shown in a simulation, and would increase the potential of the direction indication model. Therefore, decreasing the amount of wrong-sided adjustments is recommended if this is both attainable and affordable.

Another simulation that shows potential is the extension on the Blattberg-Hoch model. A percentual weight is assigned to the judgmental adjustment and the remainder to the statistical forecast. An optimization algorithm is able to identify the optimal weight per category, which provides positive model improvements for most categories. The method shows it has the potential to increase the effectivity of judgmental adjustments and thus optimizes the collaboration between the forecaster and the forecasting system. A downside for this model is that the amount of time a forecaster has to spend on adjusting is not reduced. Even though the effectivity of the adjustment is optimized, the efficiency of adjustments could possibly be improved still. The direction indication model aimed to reduce this, but failed in its envision task.

All in all, it is recommended to pursue the weighted adjustment model as a means of increasing overall forecasting accuracy. The model requires little research for implementation, requires little transfer of responsibility from the forecaster to the system and allows for collaboration instead of working parallel to each other. Additionally, it is advisable to investigate how the size of an adjustment can be forecast, since disallowing small adjustments increases the efficiency of the forecasting process.

8.1.2. Academic implications

The second set of objectives introduced aimed to contribute to the existing literature on judgmental adjustments. More specifically, it set out to generate knowledge regarding the collaboration between human forecasters and forecasting systems. The academic implications of the thesis are related to the degree to which the research supports or disproves established knowledge, and are related to the introduction of new information and ideas to the set of scholarly knowledge.

The most valuable finding is that the adjusted forecasts are, on average, more accurate than the system-generated forecasts. In line with the results of research by Willemain (1991), judgmental adjustments to a simple statistical forecasting procedure resulted in increased accuracy of the final forecast. Fildes et al. (2009) demonstrated forecasters' abilities to identify system-generated forecasts that require adjusting. Both of these conclusions are supported by the results of this thesis. Additionally, the ability to identify the correct direction was proven after testing, which supports claims made by Mathews and Diamantopoulos (1990).

Literature on the effect of adjustment size on accuracy of the adjustments suggested that a correlation exists between them (Diamantopoulos & Mathews, 1989). This finding was supported in the thesis, by means of analysis and simulation. Fildes et al. (2009) concluded that the smallest adjustments would hurt the forecasting accuracy, while research by Baecke et al. (2017) found that the correlation

between size and accuracy would eventually decrease for the largest adjustments sizes. Both of these findings were not supported by the thesis, though the error metric applied does differ which could possibly lead to different results.

Findings by Fildes et al. (2009) and Baecke et al. (2017) regarding the superior accuracy of downward adjustments over upward adjustments could not be replicated here. This implies that the optimism bias is in this study only related to the propensity to adjust upwards, which was a pattern found by Fildes et al. (2009) and Franses & Legerstee (2009). Moreover, the anticipated increase of forecasting accuracy with product importance as suggested by Fildes et al. (2009) for was not proven. Even though some evidence was found to support the claim, the hypothesis is rejected by the significance tests.

The phenomenon of consistent over forecasting, or overpredicting (Lawrence, Edmundson, & O'Connor, 1985) could also not be repeated in this research. Overpredicting was present, but not as drastically as previous research would suggest (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). This could be explained by the Anchoring and Insufficient Adjustment (AIA) bias, initially presented by Schweitzer and Cachon (2000). They found that human decision makers order too many of a low-profit product, while simultaneously ordering too little of a high-profit product. Gavirneni and Robinson (2017) show that the AIA bias can best be characterized by a combination of risk-aversion and shortage cost. The risk-aversion affects the small orders for high-profit products, while the low-profit products are overproduced due to the reduction of shortage costs. The absence of over forecasting in the reviewed setting could thus be more susceptible to the AIA bias, rather than the optimism bias.

The expected result of increased adjustment accuracy over time was confirmed. Not only does the accuracy of the adjustments increase over time, the improvement they offer increases over time as well. This implies that the adjustments improve the accuracy not with a constant, but with an increasing trend. This confirms recommendations by Ghiani, Laporte & Musmanno (2013), Nahmias (2013) and Syntetos et al. (2009), who all conclude a combination of objective and subjective forecasting is the most effective at a medium to long term forecasting horizon.

Finally, The host company applies a seasonal correction to their sales which results in the regular sales being seasonless. Corrections regarding promotional sales are applied afterwards, leaving the regular sales to be a set of deseasonalized values that contains all other properties of the signal (Nahmias, 2013). This has been an addition to the scope of academic knowledge, as previous literature has examined the effects of judgmental adjustments on final sales values.

8.2. Shortcomings

8.2.1. Limitations of the study

Although the research has included a substantial number of observations, there are some limitations that do show up. First and foremost, the research has been performed in a case study fashion. This implies that the results found here are indeed useful for the host company, but are not necessarily generalizable to a much broader setting.

Additionally, the forecasting process examined was based on a single forecaster that was supervised by a logistic manager. Adjustments made are most likely to be made by only one person, which in itself is a limitation as a single person is not representative of the entire population of human forecasters. At control moments that the logistic manager scheduled to do, adjustments could be rejected still or changed. This means that the adjustments could also be the result of two people

adjusting separately. Once again, the data storage problem makes it impossible to trace who did what in the system and the information is lost.

Moreover, a large portion of information that was lost due to the method of data storage. An analysis regarding the acceptance decision could not be performed, as this information is not stored in the dump files. As these files are simply a reflection of the SAP system at that point, there are no direct indicators of human interference. The analysis performed in the thesis was simply a difference analysis, while more conclusions could have been drawn if the required information had been present.

An analytical issue was using the ABC categories for indicating product importance, a division that is based on total sales. It is possible that the ABC categories are not indicative enough of exposure to a product to fully grasp the relationship that was hypothesized in part by Fildes et al. (2009).

Finally, the direction indication model could have been more consistent with the rest of the research. Instead of applying the absolute values for changing the system-generated forecast after an indication of an upward adjustment, a percentage is applied. This is not in line with the current forecasting process, but was not as burdensome on calculation models as upward adjustments are potentially unbounded. A percentage can be applied to all statistical forecasts evenly, whereas absolute values would have to be selected from a certain range based on pre-adjustment characteristics. Even though this might increase the model's usefulness, it was not applied as time and processing power were of the essence.

8.3. Advice for future research

The previous sections have identified the recommendations for the host company, as well as the limitations of the research presented in the thesis. By assessing them, interesting potential for future research can be identified.

For The host company, research should be set out to what type of collaboration between forecaster and system would be most optimal. This research has provided them with the tools to perform such an investigation, which is indispensable as a new ERP system will be installed. The investigation should seek out what type of adjustments are most beneficial and what type of integration, voluntary or mechanical, would present the most benefits.

In the domain of scholarly research, future investigations could focus on applying a different proxy for product importance in investigating its relationship with forecasting accuracy. The relationship found here was based on the ABC categories and total sales, but it might be true that larger sales volumes come with larger forecast adjustments and thus increase forecasting accuracy.

The simulations that were performed in the thesis only had the host company data as their learning and simulation values. It would be interesting to see if the variable weight model can increase the accuracy of adjustments in other settings as well. The same holds for the direction indication model, that might prove to be very beneficial in a different forecasting setting.

A very simple model was able to increase the forecasting accuracy by a fair amount, implying that more complex and optimized models could perhaps increase the forecasting accuracy even more. This is not necessarily the case, but the integrative forecasting design clearly outperformed its restrictive base. A more in-depth look into the possibilities that more sophisticated models, built around the same concept as discovered in this thesis, could be able to provide additional forecasting accuracy.

Based on factors such as recent sales volume, recent forecast error and recent judgmental forecast error, a more optimal model could be obtainable. Therefore, efforts into investigating these possibilities would be advisable.

Finally, forecast improvement in this study has been assessed by evaluating the outcomes of judgmentally adjusting forecasts and what the characteristics of the observation are before the adjustment is made. Human bias in adjusting has only been considered as a possible explanation for hypothesized behavior, but has not been measured directly. Therefore, the research cannot differentiate between a downwardly adjusted that is too high as a result of over-optimism or as a result of conservative adjusting. Future research could look into these phenomena more directly and could make an attempt at preventing or steering these effects to increase forecasting accuracy.

9. Conclusions

This thesis has presented proof that the collaboration of a human forecaster and a statistical forecasting algorithm can be highly beneficial. Even though ample literature is available on the topic, this setting in the confectionary sector with a rolling forecast has been quite unique.

Evidence is provided that the adjustments are not only usually in the right direction, but the adjustments directly decrease the error associated with the forecast. Additionally, the adjustments generally are better at approaching the actual demand value than the statistical forecast.

Even though there is much room for future research, the evidence that speaks in favor of judgmental adjustment is paramount in this research. It also showed that there is still room for improvement and that the quality of the adjustments can still be optimized even further. The variable weight model would be a good starting point for future research, as it has shown to be well able to increase the value of a judgmental adjustment.

Bibliography

- Andreassen, P., & Kraus, S. (1990). Judgmental Extrapolation and the Saliency of Change. *Journal of Forecasting*, 9(4), 347-372.
- Arvan, M., Fahimnia, B., Reisi, M., & Siemsen, E. (2018). Integrating human judgement into quantitative forecasting methods: A review. *Omega*, 86, 237-252.
- Baecke, P., De Baets, S., & Vanderheyden, K. (2017). Investigating the added value of integrating human judgment into statistical demand forecasting systems. *International Journal of Production Economics*, 191, 85-96.
- Bell, W. (1984). An introduction to forecasting with time series models. *Insurance: Mathematics and Economics*, 3(4), 241-255.
- Blattberg, R., & Hoch, S. (1990). Database Models and Managerial Intuition: 50% Model + 50% Manager. *Management Science*, 36(8).
- Burdette, W., & Gehan, E. (1970). *Planning and Analysis of Clinical Studies*.
- Cachon, G., & Terwiesch, C. (2012). *Matching Supply with demand: an introduction to Operations Management* (3rd ed.). McGraw-Hill Education Europe.
- Chatfield, C. (2005). Time-series forecasting. *Significance*, 2(3), 131-133.
- Cooper, H. (1984). The integrative research review: A systematic approach. *Applied social science research method series*, 2.
- Cooper, H. (1988). Organizing Knowledge Synthesis: A Taxonomy of Literature Reviews. *Knowledge in Society*, 1.
- Cox, D., & Hinkley, D. (1974). *Theoretical Statistics*.
- Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7, 1-30.
- Diamantopoulos, A., & Mathews, B. (1989). Factors affecting the nature and effectiveness of subjective revision in sales forecasting: An empirical study. *Managerial and Decision Economics*, 10(1), 51-59.
- Dietvorst, B., Simmons, J., & Massey, C. (2016). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, 64(3), 1155-1170.
- Edmundson, B., Lawrence, M., & O'Connor, M. (1988). The use of non-time series information in sales forecasting: A case study. *Journal of Forecasting*, 7(3), 201-211.
- Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 3(23).
- Franses, P., & Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting*, 25, 35-47.
- Gavirneni, S., & Robinson, L. (2017). Risk aversion and implicit storage cost explain the Anchoring and insufficient Adjustment bias in human newsvendors. *45*, 191-198.
- Ghiani, G., Laporte, G., & Musmanno, R. (2013). *Introduction to Logistic Systems Management* (2 ed.).
- Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, 16(1), 85-99.
- Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega*, 30(2), 127-135.
- Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, 12, 37-53.

- Kemphorne, O., & Folks, L. (1971). *Probability, Statistics and Data Analysis*.
- Kotteman, J., Davis, F., & Remus, W. (1994). Computer-Assisted Decision Making: Performance, Beliefs, and the Illusion of Control. *Organizational Behavior and Human Decision Making Processes*, 57, 26-37.
- Langer, E. (1975). The Illusion of Control. *Journal of Personality and Social Psychology*, 32(2), 311-328.
- Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, 43(2), 172-187.
- Lawrence, M., Edmundson, R., & O'Connor, M. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting*, 1, 25-35.
- Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: a review of progress over the last 25 years. *International Journal of Forecasting*, 22(3), 493-518.
- LeCompte, L., Klinger, J., Campbell, S., & Menke, D. (2003). Editor's Introduction. *Review of Educational Research*, 73(2), 12.
- Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., & Lewandowski, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2), 111-153.
- Mancini, S. (2019, 1 18). *Dit zijn de Beste Werkgevers per branche van 2018-2019*. Retrieved 6 7, 2019, from Effactory: <https://www.effactory.nl/kennis/blog/dit-zijn-de-beste-werkgevers-per-branche-van-2018-2019/>
- Mathews, B., & Diamantopoulos, A. (1990). Judgmental revision of sales forecasts: Effectiveness of forecast selection. *Journal of Forecasting*, 9(4), 407-415.
- Mathews, B., & Diamatopoulos, A. (1987). Alternative Indicators of Forecast Revision and Improvement. *Marketing Intelligence and Planning*, 5(2), 20-23.
- Moon, M., Mentzer, J., & Smith, C. (2003). Conducting a sales forecasting audit. *International Journal of Forecasting*, 9, 5-25.
- Nahmias, S. (2013). *Production and Operations Analysis*. Mcgraw-Hill Education - Europe.
- O'Connor, M., Remus, W., & Griggs, K. (1993). Judgmental forecasting in times of change. *International Journal of Forecasting*, 9(2), 163-172.
- Patt, A., & Zeckhauser, R. (2000). Actions Bias and Environmental Decisions. *Journal of Risk and Uncertainty*, 21(1), 45-72.
- Randolph, J. (2009). A Guide to Writing the Dissertation Literature Review. *Practical Assessment, Research and Evaluation*, 14(13).
- Rowley, J., & Slack, F. (2004). Conducting a Literature Review. *Management Research News*, 27(6), 31-39.
- Royall, R. (1986). The Effect of Sample Size on the Meaning of Significance Tests. *The American Statistician*, 40(4), 313-315.
- Sanders, N., & Ritzman, L. (1992). The Need for Contextual and Technical Knowledge in Judgmental Forecasting. *Journal of Behavioral Decision Making*, 5, 39-52.
- Sanders, N., & Ritzman, L. (1995). Bringing judgment into combination forecasts. *Journal of Behavioral Decision Making*, 13, 311-321.
- Schweitzer, M., & Cachon, G. (2000). Decision Bias in the Newsvendor Problem with a Known Demand Distribution. *Management Science*, 46(3), 404-420.
- Syntetos, A., Babai, Z., Boylan, J., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, 252(1), 1-26.

- Syntetos, A., Nikolopoulos, K., Boylan, J., Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgment into intermittent demand forecasts. *International Journal of Production Economics*, 118(1), 72-81.
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124-1131.
- Tyebjee, T. (1987). Behavioral biases in new product forecasting. *International Journal of Forecasting*, 3(3-4), 393-404.
- van Berkum, E., & Di Bucchianico, A. (2007). *Statistical Compendium*. Eindhoven.
- Webby, R., O'Connor, M., & Edmundson, B. (2005). Forecasting support systems for the incorporation of event information: An empirical investigation. *International Journal of Forecasting*, 21(3), 411-423.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.
- Willemain, T. (1991). The effect of graphical adjustment on forecast accuracy. *International Journal of Forecasting*, 7(2), 151-154.
- Yap, B., & Sim, C. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12), 2141-2155.

Appendices

Appendix 1 - Literature Review Setup

Methodology

In the preceding chapters, an introduction into the nature and purpose of the thesis has been given. Now, the literature review is to be executed. A literature review can be regarded as a summary of a subject field, that underpins the necessity of the research question (Rowley & Slack, 2004). By performing the literature review, one can discover leading researchers in the field, get acquainted with jargon, identify gaps in research and find possible trends that might exist (LeCompte, Klinger, Campbell, & Menke, 2003).

The first step is to define a methodology by which the literature review is to be conducted. Applying an existing methodology saves time, while still allowing a structured research and review process. In this review, the stages of a literature as identified by Cooper are applied (Cooper, 1984). These stages and the actions that correspond to them for proper conduction of the review, are explained in this chapter.

Stage 1 – Problem formulation: In the problem formulation stage, the research questions that are to be answered with secondary literature are created. Then the inclusions and exclusion criteria by which articles are selected are created. Finally, the research questions and the inclusion and exclusion criteria are examined by the thesis mentor for approval.

Stage 2 – Data Collection: The second stage encompasses the documented search for relevant articles and data from selected sources and those that conform to the selection criteria are to be used for the remainder of the literature review process.

Stage 3 – Data evaluation: In the third stage, information is evaluated and extracted from the article if the data are beneficial for the literature review.

Stage 4 – Data analysis and interpretation: Stage four requires the extracted data to be categorized and an analysis of the extracted data is synthesized. This will allow the data to be made sense of and to interpret possible trends or other patterns.

Stage 5 – Public presentation: After the research itself is completed, the review is put together in its final form in the fifth and final stage. Information that is deemed more interesting can be prioritized and less interesting information can be moved to the background or completely left out. The order of presentation is also determined at this time.

The first stage can be regarded as a preparatory stage, which will determine the success and scientific rigor of the literature review. Stages 2 through 4 make up the analysis part of the literature review process, which all culminates to the final stage which is the structuring of the review. This is illustrated schematically in Figure 12.

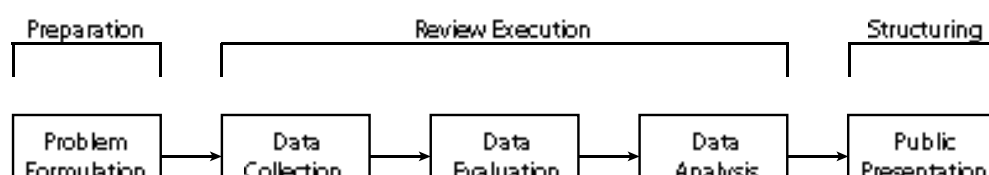


Figure 12: Schematic representation of literature review process

Problem Formulation

Now that the methodology of the literature review has been selected and the set-up is known, the literature review can be initiated. The problem formulation stage is divided into two phases. In the first phase, the research questions are generated, while in the second phase the selection criteria for articles are instated.

Research questions

To start the literature review, the research questions will first be determined. They serve as guidelines throughout the review, as they point towards the knowledge that is to be gathered in the literature review process and serve as the purpose of the literature review (Randolph, 2009). The research questions will be based on the goal and focus of the review, as they are described in Cooper's Taxonomy of Literature Reviews (Cooper, 1988), and on the problem statement and research questions generated in chapter 2.

Research goal: Identify central issues in the field of statistical demand forecasting models, especially with respect to human adaptations to the outputs of these models and integrate previous research findings.

Research focus: The literature review should focus on the outcomes of previous research on human adaptations to forecast model outputs and on the methods they applied to arrive at these outcomes.

With the goal and focus identified, the main research questions and the sub research questions can be formulated. These questions are as follows:

Main research question: What is, according to literature, the expected influence of human adjustments to forecasts on the accuracy of forecasting results in the statistical demand forecasting process?

Number	Research Question Literature Review
1	What are the most common forms of statistical demand forecasting models and what are their properties?
2	How accurate are these models generally in predicting the dependent variable they aim to forecast?
3	What are the most commonly applied methods of adjusting the results produced by forecasting models?
4	How has research in literature examined the effect of human intervention with respect to the outcomes on the accuracy of statistical demand forecasting models?
5	What is the (expected) influence a planner would want to exercise on the forecasts generated by the model?
6	When would planners be inclined to adjust forecasts generated by a forecasting model?
7	What does previous research say about how planners would change the forecast produced by a model?

Table 25: Research questions for the literature review

These research questions will serve as a guide throughout the entirety of the literature review, as their answers converge to answering the main research question.

Inclusion and exclusion criteria

With the research questions now defined, the selection criteria are to be determined. These criteria will determine which articles are suitable for use in the literature review and which will be discarded. The criteria will decrease the amount of articles to consider by immediately excluding obsolete articles and information and simultaneously reduce the amount of time required for the literature review.

The selection criteria for this literature review are as follows:

Number	Selection Criterion
1	The article has been written either in Dutch or English, to ensure no time is lost by, or mistakes are made in, translating articles from other languages.
2	Only databases acknowledged by the TU/e library will be used
2a	Article search will only occur through the databases Web of Science, Scopus and ScienceDirect, as these include results from the fields of engineering and natural sciences as well as social sciences.
3	Only articles published in scholarly and research journals will be used, as to guarantee (peer) reviewed content (Rowley & Slack, 2004). Moreover, books used by the TU/e for courses on relevant topics can be used for more knowledge of principal theory.
4	The first 30 articles in every search query after sorting for relevance are eligible for review, as well their references.
5a	The title is read for the first 30 articles and their references so to determine their relevance.
5b	If these titles seem relevant for further consideration in the literature review, their abstract is read for a definitive verdict.

Table 26: Criteria applied for data collection

Note that in Table 26, the criteria 1 – 4 can be reviewed epistemically objectively, as opposed to criteria 5a and 5b, which are subject to the reviewer’s interpretation and evaluation of the title and abstract. This means that if the literature review was to be completed by a different reviewer, the final selection of relevant articles could differ from the selection in this literature review. Therefore, documentation of article discovery will be extensive, so the thoroughness of the literature can be reviewed by the thesis mentor.

Data Collection

In chapter 0, the preparation of the literature review was initiated and finished. Now comes the search for useful literature and evaluation of relevance of the search results.

University Course Literature

Firstly, a search was conducted among the academic books that are mandatory for students in the Operations Management and Logistics program, as well as books from the bachelor Industrial Engineering and Innovation Sciences. The books that have been selected and checked for relevant chapters are from courses that handle the topics of human decision making processes, statistical forecasting models and mathematical theories on probability and operations management. For conciseness sake, book chapters included will also be referred to as articles.

Scholarly Articles and Literature

To find appropriate literature, clear search queries are to be utilized. These queries will be documented in this chapter, along with the amount of results each query returns in the database they were used in. Moreover, they will be linked to the respective research question they aim to find relevant literature for. The search was executed on the March 12, 2019. All numbers found in Table 27 were recorded on this date and presented for future reference.

Query	Research Question	Scopus (no. of articles)	Web of Science (no. of articles)	Science Direct (no. of articles)
("forecast* model" AND introduction) OR ("forecast* model" AND comparison))	1	1.682	944	17.593
("history of forecast*" OR "history of time series models")	1	19	12	174
("demand forecast* models" OR "demand predict* models") AND (common OR general OR popular)	1	23	10	563
"forecast* model" AND (accuracy OR precision OR error)	2	4.830	4.434	15.390
("demand forecast* model") AND (accuracy OR precision OR error)	2	64	54	495
("human interv*" OR "human adapt*") AND (forecast* OR "demand forecast*")	3	146	63	66
("judgmental forecast*" OR "forecasting support system*") AND (introduction OR comparison OR application OR use)	3	46	118	484
"judgmental forecast*" AND (experiment* OR examin* OR analysis OR research)	4	152	89	436
"judgmental forecast*" AND (planner OR "operations planner")	5, 6, 7	3	1	69
("operations planner" OR planner) AND (bias OR inclination) AND (forecast* OR "judgmental forecast*")	5, 6, 7	20	35	3.576

Table 27: Search queries used for data collection with number of hits per database indicated

Scopus searches were limited to "journal article" searches, Scencedirect searches to contain only Review Articles and Research Articles. Web of Science searches was limited to return only articles and reviews. This in order to comply with criteria 4 from Table 26.

Data evaluation

For stage 3 of the literature review, the articles that appear useful for inclusion in the thesis and adhere to the inclusion criteria, are read thoroughly. A summary database is created for the articles read, which is a file containing all important information from the selected articles. Moreover, remarks are made about the quality of the results and the type of study that had been conducted (field study versus laboratory setting). This to minimize the amount of time the articles themselves are to be consulted after they have been read. The summary database was created in alphabetical order of the first author of every article.

Data analysis and interpretation

After reading and summarizing the selected articles, the information that results is overwhelming. To make sense of this information, it will be coded and sorted accordingly. This means that the summary database is analysed, and similar information is color-coded and stored in a separate file that is sorted according to the information it contains. This file is used as the main source of information for constructing the public presentation in the thesis, and consequently, for answering the thesis' research questions. This presentation will be given in the succeeding chapters.

Appendix 2 – Screen capture of the SAP forecasting module interface

Niveauplanning wijzigen

Kenmerk 0- kolom

Vestiging 0313 Material 16155 GESNEDEN ONTBUTKOEK 475G (X10)

Hoofdkantoor (HK) Subhoofdkantoor(SHK) Distributiectrm (DC)

Versie A00 Active version Actief

Peijnenburg forecast gebruikers key-figures

Geaggregeerde informatie	Eh	P 37.2018	P 38.2018	P 39.2018	P 40.2018	P 41.2018	P 42.2018	P 43.2018	P 44.2018	P 45.2018	P 46.2018
Historische reguliere verkopen (HRV)	DU	4464	2692	2049	3565	4587					
Historische actie-verkopen (HAV)	DU	918	4068	5436	2502	1566					
Intercompany verkopen (ICV)	DU										
Gemiddelde reguliere verkopen (GRV)	DU	5004	5008	4990	5012	4991	4982	4982	4982	4982	4982
Manuele reguliere verkopen (MGRV-i)	DU										
Gecorrigeerde reguliere verkopen (MGRV)	DU	5016	5009	4988	5013	5003	5006	5006	5006	5006	5006
Seizoensfactor (SF)	DU	1	1	1	1	1	1	1	1	1	1
Reguliere prognose (RP)	DU	5016	5009	4988	5013	5003	5006	5006	5256	5256	5256
Bevestigde pre- en afterdip (PD/AD BA)	DU	401	321	43	1580	724	204	405	326	244	177
Vrije pre- en afterdip (PD/AD NBA)	DU										
Gecorrigeerde reguliere prognose (GRP)	DU	4615	4688	4945	3433	4279	4802	4601	4930	5012	5079
Bevestigde actie-prognose (BAP)	DU	2475	5418	5578	495	1440	720		915	7387	8091
Vrije actie-prognose (ENBAP)	DU										
Extra activeraag (EV)	DU	1021	2235	2562	25	1130	410		665	3952	4426
Actie-achterstand (AA)	DU		1021	2421	959		276				
Klantorders en IC (KLO)	DU						2121	55	1	1	1
Leveringen (LEV)	DU	5382	6760	7485	6067	6153	1042				
Salesplan (SLS)	DU	5636	6923	7507	3949	5409	5212	4601	5595	8964	9505
Salesplan incl. 50% regel (SLS-i)	DU		7750		5924	5783					
Salesplan met actieachterstand (SLSAA)	DU	5636	7944	9928	4908	5409	5488	4601	5595	8964	9505
Manueel salesplan (DSLS)	DU			8228	5300	5700					
Gecorrigeerd salesplan (GSLs)	DU	5636	7944	8228	5300	5700	5488	4601	5595	8964	9505
DC Stock (DCSTCK)	DU	9052	8482	8918	5550	5342	3710				
Plant stock (PLSTCK)	DU	108	4447		1620	1080					
Geproduceerd (PROD)	DU	8827	3523	4895	5474	3840	3240				
Trendfactor(TF)	DU										

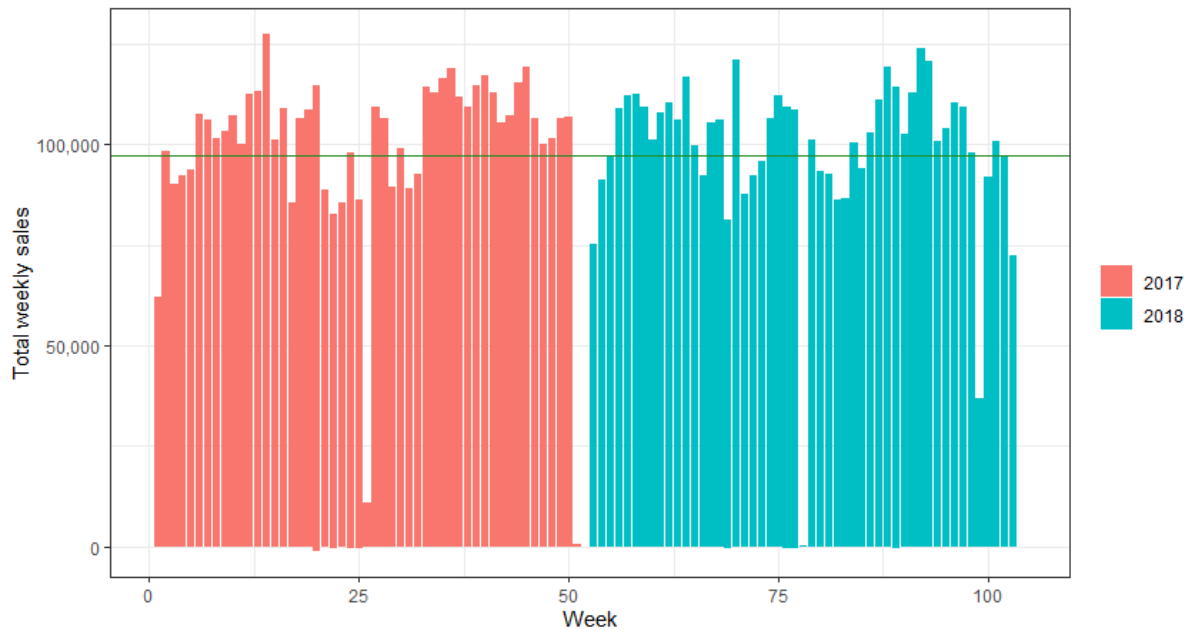
The image above shows the user interface for entering adjustment into the SAP system. The first row shows the weeks that currently on screen, for which the information is stored in column form. The names of the variables are shown in the first column, with the corresponding abbreviations in brackets.

As indicated in 5.3.1, there are two possibilities for adjusting regular demand, namely the MGRV-i and the DSLS. These are located in the fifth row from the top (MGRV-i) and fifth row from the bottom (DSLS), or more easily localized by the white cells instead of the usual grey ones.

As mentioned before in the main text, there is no indication of the impact of the adjustment. Information such as absolute adjustment size or the adjustment size relative to the statistical forecast are not visible and have to be consulted manually. Moreover, there is no indication of past forecasting performance. If an adjustment has been made, no value of accuracy, increase of accuracy or decrease of accuracy can be retrieved from the SAP system directly. This means that a lot of information is not conveyed immediately and if a forecaster wishes to be informed more about their performance or

about an adjustment they intend to make, they have to create this information themselves. This implies an additional responsibility for the forecasters and, from an efficiency point of view, requires additional time of their day that could have been spent differently.

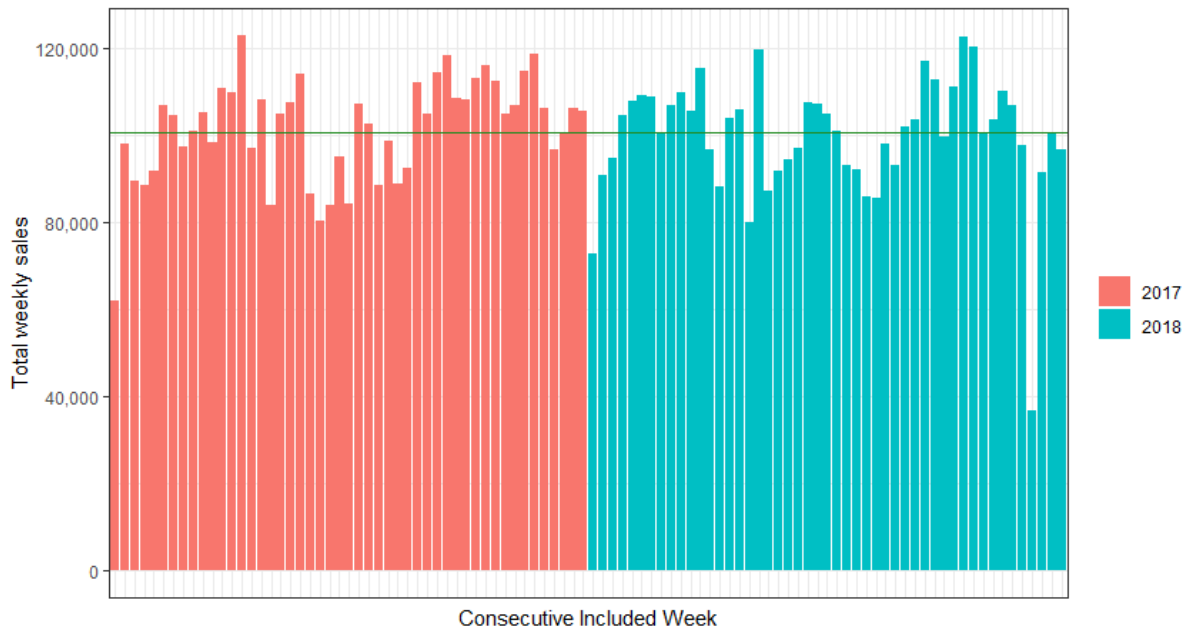
Appendix 3 - Cleaning of the Historical Regular Sales data



The image above shows the initial historical dataset, created by combining four dump files (see 5.2.1). This results in a dataset that spans 181,622 lines of code. Summing the total sales per week results in the histogram above. The weeks that contain incomplete information can be spotted easily when the graph is examined, as they are far lower than the rest of the bars. Something seems off about week 99 as well, but it will be left in the set for the time being. Moreover, due to the stacking of the demand value, some negative demand can quickly be spotted as well. The mean for this set is indicated by the horizontal green line, which has a value of 97,324.

6 weeks are removed from the dataset, namely week 26, 51 and 52 of both years. The rows that do not have both a positive sales value and a value of -1 for the judgmental adjustment are deleted. Finally, all periods of which the final forecast was 0 or lower are removed. After cleaning, the size of the file was greatly reduced to 43,777 rows of data, which is about 25% of the initial set. When plotted on a week basis, the graph below results.

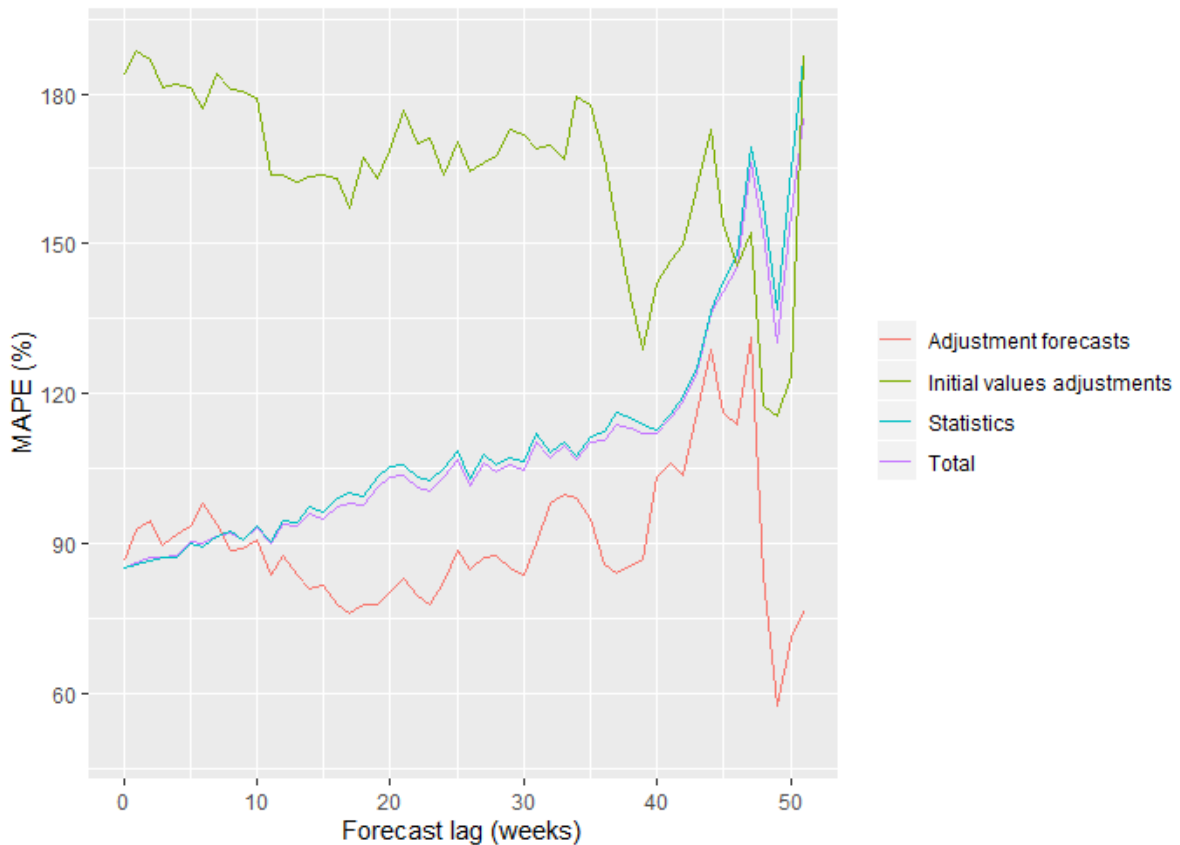
The mean now has shifted to 100,875 items on a weekly basis. Note that the lumping of the bars is a purely visual occurrence and is not related to the information that the graph is created from. This graph represents all the sales values known that occurred in 2017 and 2018. These values will be linked to the total set, in the process of creating the eventual fully cleaned set.



Appendix 4 – Additional graphs

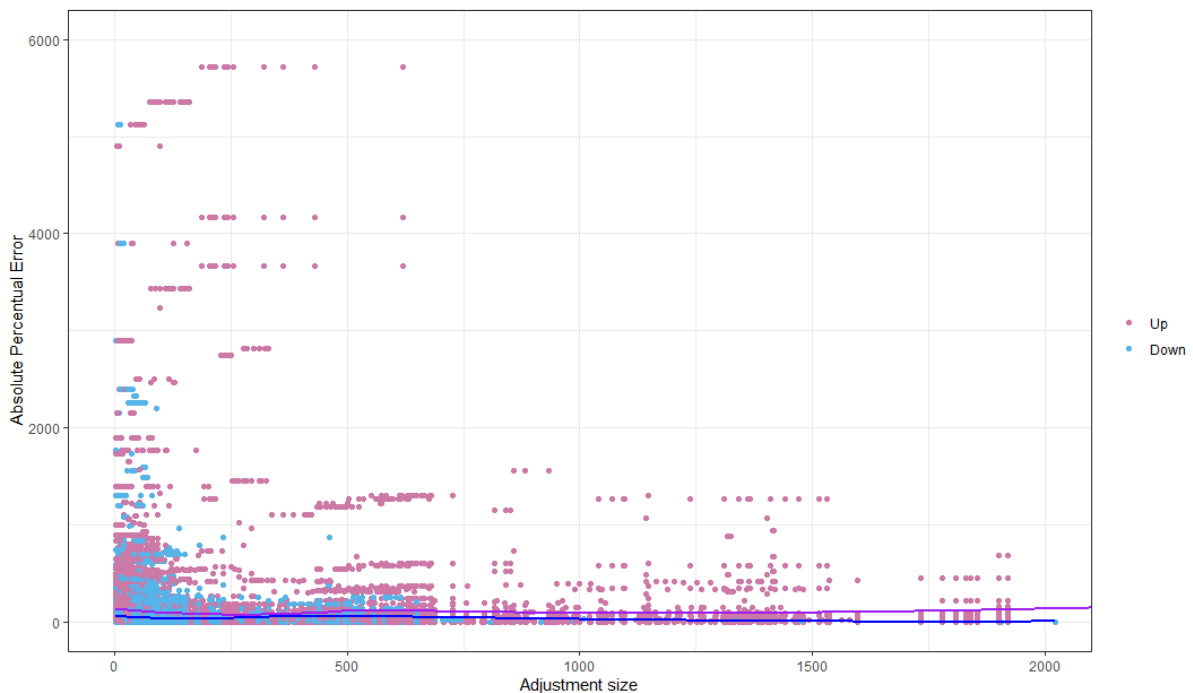
In chapter 6, the hypotheses analysis are performed. Not all graphs that have been constructed are included in this section, as their analytical value does not surpass that of the graphs that have been included or simply did not provide useful insights. In this appendix, the other graphs will be presented along with comments about their information.

Mean absolute percentage error versus forecast lag



The first plot included in the thesis is Figure 6, which shows the apparent relationship between the MAE and the forecast lag. Over time, the accuracy seems to increase, as the error measure decreases. The initial values seem to trend upwards even, thus increasing the importance of the adjustments. The image above is similar to Figure 6, but applies the MAPE. They are thus transformed into a percentage. As a result, the error measure becomes highly reliant on the volume of the eventual realized sales. MAPE seems to be increasing in popularity in companies, but requires additional cleaning based on the volume of the product sold and thus slims the available set of data. Examples of researches in which MAPE is applied are Goodwin (2000), Fildes et al. (2009) and Baecke et al. (2017). However, for this research the MAE has been selected as the leading measure.

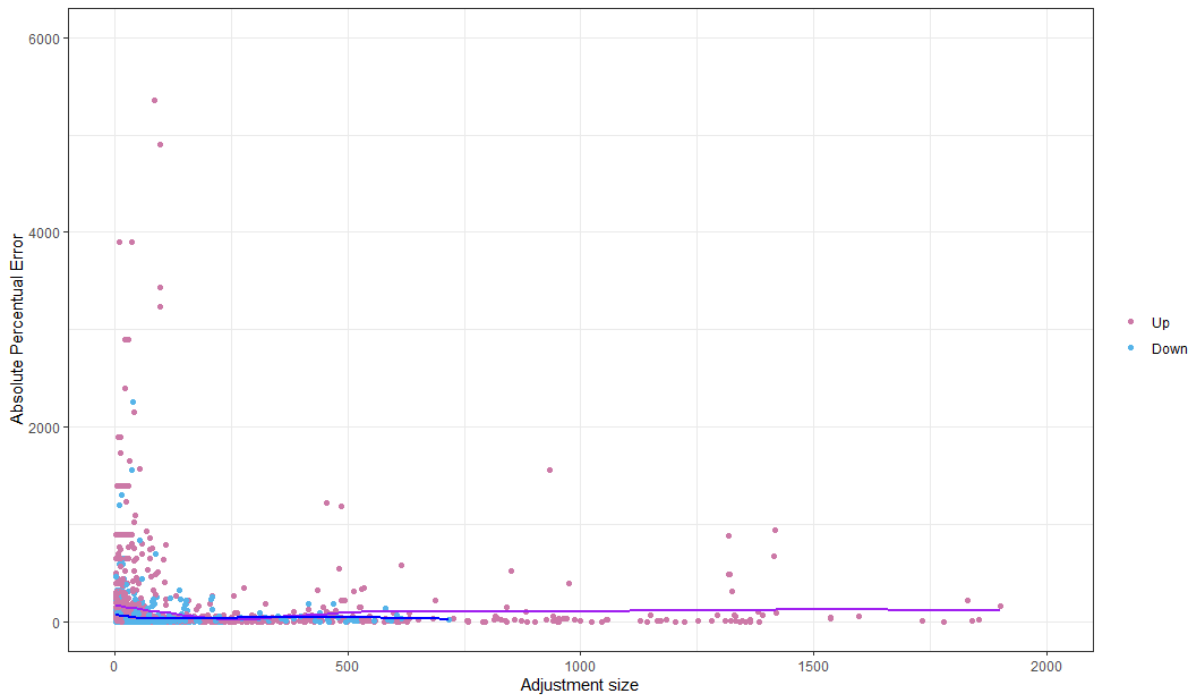
Absolute percentual error versus absolute adjustment size



The image above portrays the initial graph style that was selected for analysis of the effect of adjustment size on the forecasting error. In the end, the axis style layout of Figure 7 **Error! Reference source not found.** was selected since it allows for additional information to be conveyed. In this graph, the direction of the judgmental adjustment is indicated by means of color. This is not necessary in Figure 7, as the adjustment direction is indicated geographically by the position relative to the y-axis. Color can consequently indicate other kinds of information, for which the correctness of the decision was selected. Finally, the absolute values of the error were dropped, to obtain a less dense picture.

The image above uses all data points, rather than the four-week ahead forecast exclusively. As a consequence, the graph is very unclear and appear to make a smeared out impression as many consecutive forecast runs differ very little from each other. This problem disappears when a single run is selected as the source data.

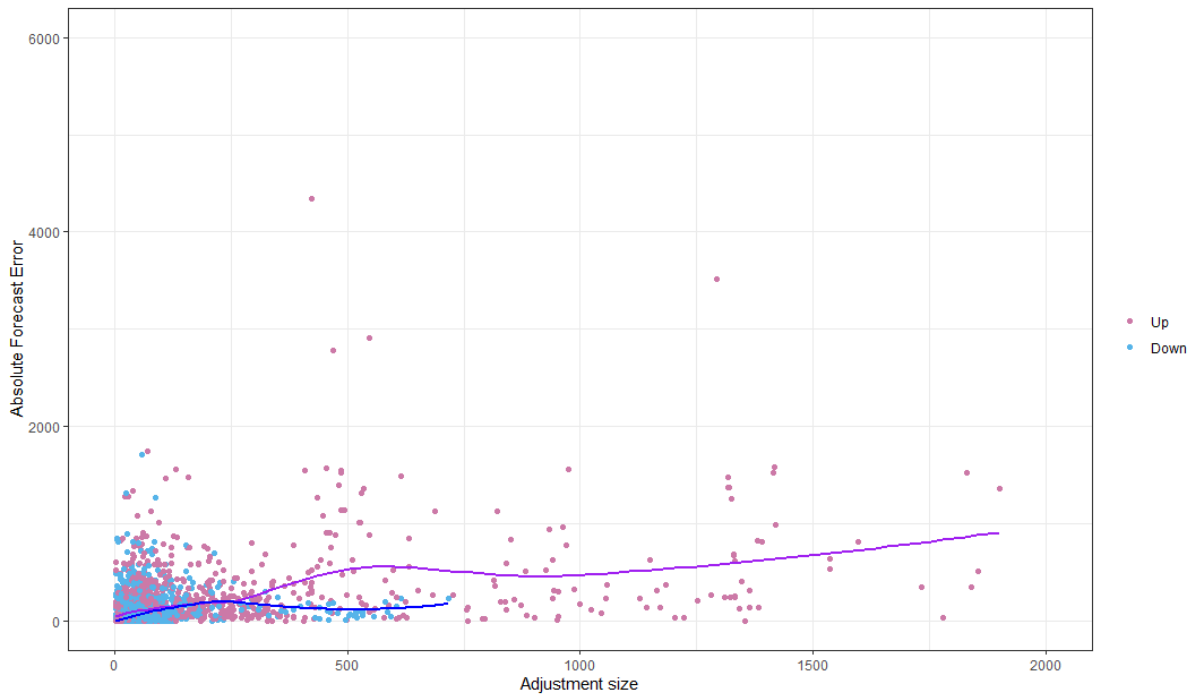
Two regression lines were included as well, but due to the use of absolute measures, the very much flat and provide no analytical value. The same holds for the image below, which applies the same style as just discussed, but only includes a single forecast run.



Absolute forecast error versus absolute adjustment size



In the third graph type (image above), the percentual error on the y-axis is replaced with the absolute value of the error. The dots behave fairly similar but are scattered more over the entire plot. Moreover, a relationship does seem to exist between size and error measure. The error seems to increase with adjustment size for the upward adjustments, while the downward adjustments show a more inverted U-shape with regards to the error. However, this graphing style combines too much information into a single quartile. As a result, information is lost and the regression lines provide little value as upward and downward adjustments have dissimilar properties.



In the sixth image, the values for the absolute error are displayed graphically for a forecast lag of four weeks. The regression line for the upward adjustments behaves similarly to the regression line for the full dataset, but for the downward adjustments, a slight positive trend is detectable at the end of the regression line. When the amount of observations near the end is considered, the shape holds little explanatory value. Similar to the three preceding graphs in this appendix, it was not included due to its incapability of conveying a clear message.

MAE improvement per ABC category



The final image presented, shows the distribution of the improvement in MAE per category of product importance. The position on the x-axis within the bar is random, but the MAE improvement value is

as presented. This picture clearly shows the higher improvement values for category A as compared to B and C. Additionally, the smaller bandwidth of downward adjustments also can be seen here. However, the section regarding the ABC category was already quite long and the image and was preceded and succeeded by sections with figures. Compared to those images, it carried less weight and the main findings from this figure are discussed in a later hypothesis. All these things considered, the figure was deemed to be of too little analytical importance and for the sake of readability not included in the main text.

Appendix 5 - Statistical concepts and methods for hypothesis testing

In this section of the thesis, the different statistical methods applied throughout the research will be explained. The reporting of actual test results can be found in the next section of the appendix, where the tests performed for hypothesis testing in chapter 6 are reported.

Degrees of freedom

In order to grip the hypotheses tests in this section, it is vital that the notion of degrees of freedom is explained. The degrees of freedom is the amount of values within a statistical calculation that can vary freely. Statistical systems are bound by certain mathematical laws that they adhere to. As a result, all but the last entries in that system are free to vary. If all but the last entries in the set have taken a value, the last entry is not free to vary and has to take a certain value in order to adhere to mathematical laws. In other words, the final entry is not free to take any value and is bound by the values of the others. The degrees of freedom is thus equal to the total set of entries in the system, minus 1.

P-value

Additionally, the p-value is explained in this section. The p-value is an important concept in the world of statistics, as it refers to the statistical validity of outcomes and quantifies the strength of the presented evidence (Kempthorne & Folks, 1971). The p-value expresses the probability that the observed data is a result of the distribution proposed under the null hypothesis (Cox & Hinkley, 1974). The significance test, and thus the p-value, can be interpreted as a quantification of the evidence the data poses against the null hypothesis (Royall, 1986). In this light, it is evident that a smaller p-value implies stronger evidence against H_0 than a larger p-value.

The following meanings are assessed to the p-value levels in testing the hypotheses in this study (adapted from (Burdette & Gehan, 1970)).

P-value	Meaning
$p\text{-value} < 0.01$	Very robust evidence against implication of the null hypothesis
$0.01 \leq p\text{-value} < 0.05$	Moderately robust evidence against implication of the null hypothesis
$p\text{-value} \geq 0.05$	No evidence to disprove the null-hypothesis

In the following sections of the Appendices, as well as in the main text, an asterisk or double asterisk can be found as superscript to a result. This asterisk indicates the significance level of the test that has been performed and thus of the test result. If no significance is found, this is reported explicitly.

Student t-test

The Student t-test was developed by William Gosset under his pseudonym Student, as his employer Guinness prohibited publishing scientific articles (van Berkum & Di Bucchianico, 2007). A t-test can be applied to discover if the observed series differs significantly from the series that was expected.

The Student t-test is based on a Student t-distribution, whose shape relies on the degrees of freedom. It is a symmetrically normalized distribution, implying it has a mean equal to zero and has the same positive and negative x-value for every point on the distribution's y-value. The distribution is very

similar to a normal distribution and follow a similar distribution shape. It technically varies only in that it applies not a set standard deviation, but applies a standard deviation it has deduced from the sample of data that is being applied on. The sample standard deviation is then applied to check for significant differences between different datasets.

Shapiro-Wilk test

The Shapiro-Wilk test is used to check if a sample is normally distributed. The null hypothesis of the test is that the observations in the sample do follow a normal distribution. It applies an analysis of variance for the observations in the sample. It calculates a value for the test statistic W based on the weight of the variance of the sample versus the variance of a standard normal distribution (Yap & Sim, 2011). The null hypothesis of the test is that the observations in the sample do follow a normal distribution. If the test statistic is unlikely to arise from a normally distributed variable, the p-value will be low and the null hypothesis can be rejected. The downside of the test is its unreliable results for large samples, which therefore should contain no more than five thousand observations. The Shapiro-Wilk test is generally considered to be the strongest test for normality (Yap & Sim, 2011).

Wilcoxon

Frank Wilcoxon in 1945 put forward two methods for comparing the test results of two treatments with one another (Wilcoxon, 1945). These tests are called the signed-rank test and the rank-sum test. Both tests can be applied as a replacement for their corresponding t-test, as a normal distribution of data is not required for the Wilcoxon test (Demsar, 2006).

Wilcoxon rank-sum test

The Wilcoxon rank-sum test (also known as the Mann-Whitney test) test the null hypothesis that the two samples are derived from the same distribution, and thus have an equal mean. The samples used for the rank-sum test are independent and do not require the sample sizes to be of equal size. It ranks all the observations that have been made from smallest to largest and checks how many times an observation from one sample is ranked lower than an observation from the other sample. Scores from both samples are then put together and ranked, with the lowest score receiving rank one, the second lowest score rank two et cetera. Equal scores get the same rank assigned. The test statistic is the smallest of the sum of ranks for the two sets, called W . The p-value is derived from the probability that the larger rank is from the same distribution given the alternative of being unequal, higher or lower.

Wilcoxon paired signed-rank test

The Wilcoxon signed-rank test calculates the differences between two paired (dependent) samples and ranks the differences whilst ignoring the signs, meaning it uses the absolute differences (Demsar, 2006). Once again, the sum of rank numbers is calculated and the p-value is based on the likeliness that the larger rank sum is unequal, larger or smaller than the small one, based on the alternative hypothesis.

Wilcoxon tests are robust to outliers, as there are no actual values used but only the ranks of the values. The magnitude of an outlier is subsequently lost.

Chi-square test of independence

The Chi-square test is a test for categorical variables presented in a contingency table (frequency table). The categorical variables are set out against the groups the independence test is made for. For every row and column, the totals are created. Based on these totals, the expected counts per cell are calculated. The test then is performed to check how likely the observed variables are given the expected values calculated before, which is expressed by the p-value. The null hypothesis is that the two variables are independent, meaning that they do not influence each other. In other words, the expected outcomes are close to the observed outcomes as the independence implies that observations could be made purely on chance.

A significant p-value means that the two variables are dependent, implying that the observed and expected observations are too far apart to be a coincidence.