Improving an operationally oriented forecasting process by analyzing the interaction between algorithm and analysts

Klein, E.P.A.

Award date:
2022

Department of Industrial Engineering & Innovation Sciences

Improving an operationally oriented forecasting process by analyzing the interaction between algorithm and analysts

Master Thesis

E.P.A. Klein
Student Number: 0960688

Supervisors:
dr. P.P.F.M. van de Calseyde
dr. ir. R.J.I. Basten
dr. S. Rispens

Eindhoven, November 2022

Series Master Theses Operations Management and Logistics

Keywords: Forecasting, Rolling forecast, Aggregated, Workload, Employee scheduling, Judgmental adjustments, Heuristics and biases, Forecast accuracy

Abstract

For efficient employee scheduling, accurate forecasts of the expected workload are crucial. In this study, the forecasts of the expected workload for the operational domains of Coolblue, a Dutch e-commerce company, are analyzed. The forecasts are made on an aggregated level based on the operational processes that are required, using a two-step forecasting method. First, based on historical data, a statistical forecast is generated by a machine learning algorithm. Second, analysts can make judgmental adjustments to this forecast to account for exceptional circumstances of which the algorithm is not aware. Although this can increase forecast accuracy, judgmental adjustments can also lead to bias in the forecast and reduce forecast accuracy. To investigate whether, which, and when judgmental adjustments contribute to forecast accuracy, 111,852 judgemental adjustments are analyzed. Overall, the quality of operationally oriented forecasts is improved by judgmental adjustments, and adjustments related to disruptions, in this case Covid-19, outperformed the adjustments that were not related to disruptions. Some types of adjustments appeared to be more effective than others. Adjustments in absolute numbers generally decreased forecast accuracy, in contrast to percentage adjustments and overrides. In addition, subsequent adjustments were effective in approximately half of the cases, and especially subsequent upward and subsequent large adjustments were detrimental. As of 6 weeks before the forecast date, analysts were able to significantly improve the forecast accuracy relative to the statistical forecast. Based on the findings, several new adjustment procedures are designed, taking into account the behavioral tendencies of the analysts. Combining the skills of the statistical model and the skills of analysts, and dampening generally detrimental adjustments appeared to be effective in reducing the forecast error. The forecast could be further improved by assigning different weights to the forecasts, and the complete elimination of certain types of adjustments also turned out to be beneficial. Based on the insights, time and resources should be allocated to the type of adjustments that have the biggest positive impact on forecast accuracy.

Management Summary

Introduction

Coolblue is a Dutch e-commerce company, and for its four operational domains: the warehouse, customer service, stores and delivery, it forecasts the expected workload. Based on the forecasts of the expected workload, expressed in the number of orders, customer contacts, visitors and delivery trips, employees are scheduled. The operational forecasts are made on an aggregated level based on the operational processes that are required. The forecasting system is based on a rolling forecast and consists of two steps. First, on a daily basis, a statistical forecast, referred to as *raw forecast*, is generated by a machine learning algorithm, utilizing historical data. Second, analysts can make adjustments to these raw forecasts when they possess contextual knowledge that the algorithm does not have, resulting in a *adjusted forecast.* This adjusted forecast is eventually communicated to the stakeholders and serves as input for their planning process. Multiple adjustments can be made for the same forecast over time, and they can be made in three ways: a percentage adjustment, an adjustment in absolute numbers or an override. The first adjustment for a particular forecast is made to the pure raw forecast and after that, subsequent adjustments are made to a forecast which consists of the raw forecast with all previous adjustments processed in it, referred to as the *adjusted forecast morning.* Although it can be beneficial to let analysts make adjustments to process contextual information in the forecast, it can also lead to bias and therefore a reduced forecast accuracy. It is therefore important to know which adjustments do and do not contribute to a higher forecast accuracy. Currently, Coolblue does not have insight into this. The first research question, therefore, is formulated as follows:

RQ 1*: What are the specific strengths and weaknesses of analysts when adjusting algorithm-generated forecasts, with regard to the characteristics of adjustments, and when do these adjustments increase and decrease the accuracy of operationally oriented forecasts?*

Based on the insights, it should be investigated how the algorithm forecast can be best combined with analysts' judgement, taking into account the behavioral tendencies of the analysts. The second research question, therefore, is formulated as follows:

RQ 2*: Given the strengths and weaknesses of analysts when adjusting algorithm-generated forecasts, what is the most effective procedure to make adjustments to algorithm-generated forecasts?*

The aim is to investigate whether a more effective adjustment procedure is possible to ultimately increase forecasting accuracy and reduce the time spent on making adjustments.

Literature Review

A literature review is conducted, and based on that, hypotheses were formulated regarding the behavioral tendencies of analysts when making judgmental adjustments. The first hypothesis focuses on the accuracy of judgmentally adjusted forecasts. Several studies have demonstrated that forecast accuracy is increased by judgemental adjustments (Fildes, Goodwin, Lawrence & Nikolopoulos, 2009; Syntetos, Nikolopoulos, Boylan, Fildes & Goodwin, 2009). Despite the biases that judgmental adjustments are subject to, the processing of contextual information, of which statistical models are not aware, is considered effective to improve the accuracy of forecasts (Fildes et al., 2009; Lawrence, Goodwin, O'Connor & Önkal, 2006). It is therefore hypothesized that judgmental adjustments increase the quality of operationally oriented forecasts.

The second set of hypotheses focused on the direction of judgmental adjustments, i.e. upward and downward. Adjustments should be made in the right direction to provide accurate forecasts, as wrong-sided adjustments in particular have a major detrimental impact on forecast accuracy (Fildes et al., 2009). Based on the existing literature it was expected that analysts are generally competent in making adjustments in the right direction. Furthermore, it is frequently observed that the frequency with which upward adjustments are made is higher than that of downward adjustments, and that forecasts that are adjusted upward are much less likely to increase accuracy than those that are adjusted downward (Fildes et al., 2009; Syntetos et al., 2009; Trapero, Pedregal, Fildes & Kourentzes, 2013). This is mostly explained by an optimism bias, causing upward adjustments to be excessively high or wrongly directed. Moreover, according to Fildes et al. (2009), the information on which adjustments are based is often more reliable in the case of downward adjustments. In addition, acting upon an asymmetric loss function, by confusing forecasts and decisions, is another possible contributor to the bias in upward adjustments. In the context of this study, it was expected, assuming an asymmetric loss function where under-forecasting is more costly than over-forecasting, in combination with an optimism bias, that analysts were more likely to adjust a forecast upward than to adjust a forecast downward, and that downward judgmental adjustments are in general more likely to decrease the forecasting error of operationally oriented forecasts than upward judgmental adjustments.

Next to the direction of adjustments, there may be a correlation between the size of an adjustment and the accuracy of forecasts. Large adjustments are typically more successful than small adjustments in improving forecast accuracy (Baecke, De Baets & Vanderheyden, 2017; Van den Broeke, De Baets, Vereecke, Baecke & Vanderheyden, 2019). A potential reason for this is that large adjustments are most typically linked to reliable information, and if forecasters consider information to be unreliable, they probably reduce the size of the adjustment. Additionally, there is a propensity for small adjustments to be made for no apparent reason. For example, to have a sense of control (Goodwin, 2002), or because of forecasters 'tinkering with the data' just to show that they are doing their job (Fildes et al., 2009). It was therefore hypothesized that relatively large judgmental adjustments are more likely to decrease the forecasting error than relatively small judgmental adjustments.

Finally, concerning the timing of adjustments, it was hypothesized that the forecast error of adjusted forecasts increases with a decreasing time horizon. The possible reason for this is that forecasters may overreact to new information that becomes available closer to the forecast date (Lawrence, O'Connor & Edmundson, 2000). In addition, it was expected that the forecast accuracy improvement, i.e. the amount of error that analysts can reduce relative to the statistical forecast, decreases with a decreasing time horizon. Due to a rolling forecast, where the statistical forecast is updated every day, it was expected that the statistical forecast already reflects the most up-to-date information, and that therefore the added value that analysts can still contribute decreases over time (Van den Broeke et al., 2019).

Methodology

The main methods used in this study are hypothesis testing and simulation. To retrieve the data for the analysis from the database, queries are written using Structured Query Language (SQL). The data used in the study runs from 20-05-2020 to 20-05-2022, in terms of days for which a forecast has been made. For evaluating the accuracy of judgmental adjustments, the Median Absolute Percentage Error (MdAPE) is used. For data cleaning and preparation, observations with negative values for a forecast or the target have been removed. The target is the workload that was actually there on a given day. Furthermore, extra (dummy) variables have been created for, for example, the timing, direction, and size of an adjustment. In addition, an outlier analysis is performed by applying the Interquartile Range (IQR) method. The final data set used in this study consists of 111,852 judgmental adjustments. The data set is divided into a train and test data set by a time-based split of 80%/20% based on the forecast dates. Hypotheses have been tested on the train data set using the program RStudio version 4.2.1. Non-parametric tests are applied since the data was not normally distributed. Exploratory research is done on Covid and Non-Covid adjustments, the type of adjustments (absolute, percentage, override), and the characteristics and effects of subsequent adjustments. Based on the insights of the data analysis, new forecasting procedures are designed and trained on the training data set using simulation. Subsequently, the performance is evaluated on the test data set.

Results and Conclusion

The results of the analysis show that, in general, the quality of operationally oriented forecasts is improved by judgmental adjustments. This applies for percentage and override adjustments, but not for absolute adjustments. Furthermore, subsequent adjustments improve the quality of forecasts in approximately half of the cases, and judgmental adjustments related to Covid had a greater positive impact on forecast accuracy than adjustments not related to Covid. Moreover, analysts proved capable of making adjustments in the right direction, and the ratio of upward and downward adjustments was equal. In general, downward adjustments were more likely than upward adjustments to reduce the forecast error. Subsequent upward adjustments were found to be generally detrimental. For first adjustments, to the pure raw forecast, relatively large adjustments generally had a better performance than relatively small adjustments. For subsequent adjustments, the relatively large adjustments, in general, increased the forecast error. Finally, with a decreasing time horizon, i.e. closer to the forecast date, the forecast error of adjusted forecasts decreased, and the amount of error that analysts reduced relative to the raw forecast increased. As of 6 weeks before the forecast date, analysts were generally able to significantly improve the forecast accuracy relative to the raw forecast.

The results of the simulations show that new forecasting procedures that take into account the behavioral tendencies of analysts can lead to an improvement in forecast accuracy. The simulation models were targeted at subsequent upward adjustments, subsequent large adjustments, and absolute adjustments, since these were generally detrimental to forecast accuracy. By damping or even completely eliminating generally detrimental adjustments, and utilizing the generally effective adjustments, the forecast accuracy was improved. Both a strategy in which the skills of the statistical model are combined with the skills of analysts, and a strategy fully focused on dampening generally detrimental adjustments were the basis for this. Assigning different weights to the forecasts showed that this can further improve the forecast accuracy. The most effective procedure for making adjustments differs per operational domain.

Implications, Recommendations and Future Research

Largely the same results are seen in terms of adjustments characteristics and forecast accuracy in the context of this study, where forecasts are made for the expected workload at an aggregated level, compared to the commonly seen context in the literature of sales forecasting. In addition, exploratory research showed that the performance of adjustments that were directly linked to disruptions, in this case the COVID-19 pandemic, was better than the performance of adjustments that were not directly linked to disruptions. This indicates that human judgment can be of extra added value when forecasting in periods with disruptions. Furthermore, this research shows that subsequent adjustments increase the forecast accuracy in approximately half of the cases, and that this is highly dependent on the characteristics of such subsequent adjustments. Especially subsequent upward and subsequent large adjustments were generally detrimental. Moreover, a difference was found in the effectiveness of different types of adjustments. Absolute adjustments, in contrast to percentage and override adjustments, were generally detrimental. Also, adjustments proved to be more effective closer to the forecast date. It is recommended that time and resources should be allocated to the type of adjustments that have the biggest positive impact on forecast accuracy. A topic for future research is to investigate whether, rather than the way of adjusting, the reasons for particular types of adjustments could underlie the forecasting performance.

Finally, dampening or completely eliminating adjustments that are generally biased, and using the complementary strengths and weaknesses of analysts and algorithm, led to an improved forecast accuracy and can decrease the time spent on making adjustments. Improved forecast accuracy can lead to a more efficient employee scheduling. This, in turn, can ensure that not too few employees are scheduled, which can lead to higher customer satisfaction through, for example, on-time delivery of products or answering customer questions without too long waiting times. In addition, it can prevent too many employees to be scheduled, which can lead to FTE savings. A subject for future research is to translate the error reduction caused by judgmental adjustments into cost savings of a decreased number of FTE required, due to efficient employee scheduling, powered by accurate forecasts of the workload.

Table of Contents

Abstract ii

Management Summary iii

List of Abbreviations ix

List of Figures x

List of Tables xiii

1 Introduction 1

1.1 Company Description 1

1.2 Problem Context 1

1.3 Problem Statement 3

1.4 Research Questions 4

1.5 Research Scope 4

1.6 Objectives 4

1.6.1 Practical Objective 4

1.6.2 Academic Objective 5

2 Literature Review 6

2.1 Statistical Forecasting 6

2.1.1 Quantitative Forecasting Methods 6

2.1.2 Limitations of Quantitative Forecasting Methods in Isolation 8

2.2 Integrating Forecasting Methods 8

2.2.1 Integration of Human Judgment and Quantitative Methods 9

2.2.2 Integration Methods 9

2.3 Judgmental Adjustments 10

2.3.1 Motivations for Judgmental Adjustments 11

2.3.2 Accuracy of Judgmentally Adjusted Forecasts 12

2.3.3 Direction of Judgmental Adjustments 14

2.3.4 Size of Judgmental Adjustments 17

2.3.5 Timing of Judgmental Adjustments 18

3 Forecasting Process 20

3.1 Adjustment Procedure 20

3.2 Example Forecast Calculation 21

3.2.1 Updated Raw Forecast 21

4 Methodology 22

4.1 Data Extraction 22

4.1.1 Data Transformations 22

4.1.2 Data Choices 23

4.2 Data Sets 23

4.2.1 Additional Information Data Sets 24

4.2.2 Initial Variables 24

4.3 Error Measure 24

4.4 Data Cleaning and Preparation 25

4.4.1 Added Variables 26

4.4.2 Outlier Analysis 26

4.4.3 Train set and Test set 27

4.5 Data Analysis 27

4.5.1 Hypothesis Testing 27

4.5.2 Exploratory Research 28

4.5.3 Improve Adjustment Procedure 28

5 Results 29

5.1 Descriptive Statistics 29

5.2 Accuracy of adjusted forecasts 29

5.3 Direction of judgmental adjustments 35

5.3.1 Right vs. wrong direction of adjustment 35

5.3.2 Frequency of upward and downward adjustments 37

5.3.3 Performance of upward and downward adjustments 40

5.4 The size of adjustments 45

5.5 The influence of timing 53

5.6 Main Findings 57

6 Improve adjustment procedure 58

6.1 Improvement Strategies 58

6.1.1 Improvement Strategy A 58

6.1.2 Improvement Strategy B 58

6.2 Improvement Models 59
6.2.1 Model 1 - Subsequent Upward Adjustments 59
6.2.2 Model 2 - Subsequent Large Adjustments 59
6.2.3 Model 3 - Absolute Adjustments 60
6.2.4 Performance Models 1, 2, 3 60
6.2.5 Model 4 - Combination Model 60
6.2.6 Performance Combination Model 61
6.3 Domain Specific Models 62
6.4 Main Findings 66

7 Conclusion 67
7.1 Behavioral Tendencies of Analysts 67
7.2 Improvements Adjustment Procedure 68

8 Discussion 69
8.1 Implications 69
8.1.1 Academic Implications 69
8.1.2 Practical Implications 71
8.2 Limitations 72
8.2.1 Academic Limitations 72
8.2.2 Practical Limitations 73
8.3 Recommendations 74
8.4 Future Research 74

References 76

Appendices 80

A Statistical Tests 80
A.1 Wilcoxon signed-rank test 80
A.2 Kruskal-Wallis Test 80
A.2.1 Pairwise Wilcox Test 81
A.3 Chi-square goodness-of-fit test 81
A.4 Effect Sizes 82

List of Abbreviations

SKU Stock-Keeping Unit

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

APE Absolute Percentage Error

MdAPE Median Absolute Percentage Error

IQR Interquartile Range

AF Adjusted Forecast

RF Raw Forecast

AFM Adjusted Forecast Morning

PO Products Ordered

OD Orders Delivered

VR Visitors Registered

CO Contacts Offered

List of Figures

1.1 An example of an operationally oriented forecast with adjustments 2

3.1 Forecasting Process 20

4.1 Isolated Effect of Adjustments 28

5.1 MdAPE RF and AF per data set 30

5.2 MdAPE AFM and AF per data set 31

5.3 MdAPE RF, AF and AFM per No. of adjustment 32

5.4 MdAPE RF, AF per adjustment type for all adjustments 32

5.5 MdAPE RF, AF per adjustment type 33

5.6 MdAPE AFM, AF per adjustment type 33

5.7 MdAPE RF and AF per domain 34

5.8 MdAPE AFM and AF per domain 34

5.9 Percentage of right and wrong direction per data set 35

5.10 % Adjustments right direction per No. 36

5.11 % Right direction per month ahead 36

5.12 Percentage of right and wrong direction per adjustment type 36

5.13 Percentage of right and wrong direction per domain 37

5.14 Percentage of upward and downward adjustments per data set 38

5.15 % Upward adjustments per No. 39

5.16 % Upward adjustments per month ahead 39

5.17 Percentage upward and downward adjustments per adjustment type 39

5.18 Percentage upward and downward adjustments per domain 40

5.19 MdAPE RF and AF upward and downward for all adjustments 41

5.20 MdAPE RF, AF for up and down 41

5.21 MdAPE AFM, AF for up and down 41

5.22 MdAPE RF, AF upward and downward per adjustment type for all adjustments 42

5.23 MdAPE RF, AF per adjustment type 43

5.24 MdAPE AFM, AF per adjustment type 43

5.25 MdAPE RF, AF per domain for all adjustments 44

5.26 MdAPE RF, AF per domain 44

5.27 MdAPE AFM, AF per domain 44
5.28 MdAPE RF, AF for Large and Small 46
5.29 MdAPE RF, AF for Large and Small 46
5.30 MdAPE AFM, AF for Large and Small 46
5.31 Median size of adjustments per No. adjustment 47
5.32 MdAPE RF, AF large and small per adjustment type for all adjustments 48
5.33 MdAPE RF, AF per adjustment type 48
5.34 MdAPE AFM, AF per adjustment type 48
5.35 MdAPE RF and AF large and small for upward and downward for all adjustments 50
5.36 MdAPE RF, AF large, small, up, down 50
5.37 MdAPE AFM, AF large, small, up, down 50
5.38 MdAPE RF and AF per domain for large and small for all adjustments 51
5.39 MdAPE RF, AF per domain, up/down 52
5.40 MdAPE AFM, AF per domain, up/down 52
5.41 MdAPE RF, AF per month ahead 53
5.42 MdAPE AFM, AF per month ahead 53
5.43 MdAPE RF, AF 13 weeks ahead 54
5.44 MdAPE AF, AFM 13 weeks ahead 54
5.45 MdAPE RF and AF one month ahead 55
5.46 MdAPE AF and AFM one month ahead 55
5.47 RF AF 13 weeks ahead CO 55
5.48 AF AFM 13 weeks ahead CO 55
5.49 RF AF 13 weeks ahead OD 56
5.50 AF AFM 13 weeks ahead OD 56
5.51 RF AF 13 weeks ahead PO 56
5.52 AF AFM 13 weeks ahead PO 56
5.53 RF AF 13 weeks ahead VR 57
5.54 AF AFM 13 weeks ahead VR 57

6.1 CombiModel A (0.4/0.6) 1 year - Train 61
6.2 CombiModel A (0.4/0.6) 13 weeks - Train 61
6.3 CombiModel A (0.4/0.6) 1 year - Test 62
6.4 CombiModel A (0.4/0.6) 13 weeks - Test 62
6.5 Model 1B, 1 year ahead OD - Train 63
6.6 Model 1B, 13 weeks ahead OD - Train 63
6.7 Model 1B, 1 year ahead OD - Test 63

6.8 Model 1B, 13 weeks ahead OD - Test 63
6.9 CombiModel B, 1 year PO - Train 64
6.10 CombiModel B, 13 weeks PO - Train 64
6.11 CombiModel B, 1 year PO - Test 64
6.12 CombiModel B, 13 weeks PO - Test 64
6.13 CombiModel A, 1 year VR - Train 65
6.14 CombiModel A, 13 weeks VR - Train 65
6.15 CombiModel A, 1 year VR - Test 65
6.16 CombiModel A, 13 weeks VR - Test 65

List of Tables

3.1 Fictional Example of the Adjustment Sheet 20
3.2 Numeric Example Forecasting Process 21

4.1 Available data per domain 23
4.2 Variables initial data set 24
4.3 Median Size of Adjustments per Data Set 26
4.4 Amount of Train and Test Data per Domain 27
4.5 Statistical Tests 27

5.1 Average Number of Adjustments per Key and Date 29
5.2 Number of Adjustments per Adjustment Type 29
5.3 MdAPE RF vs. AF per data set 30
5.4 MdAPE RF and AF Covid vs. Non-Covid data 30
5.5 MdAPE AFM vs. AF per data set where No. adjustment >1 31
5.6 Performance of the AF in percentages compared to the RF and the AFM 31
5.7 MdAPE RF vs. AF per adjustment type 32
5.8 MdAPE RF vs. AF per adjustment type for first adjustment 33
5.9 MdAPE AFM vs. AF per adjustment type 33
5.10 MdAPE RF vs. AF per domain 34
5.11 MdAPE AFM vs. AF per domain 34
5.12 Performance of adjustments relative to the RF and AFM per domain 34
5.13 Percentage of right and wrong direction per data set 35
5.14 Percentage of right and wrong direction per adjustment type 37
5.15 Right and wrong direction upwards and downward adjustments 37
5.16 Right and wrong direction per domain 37
5.17 Percentage of upward and downward adjustments per data set 38
5.18 % upward and downward adjustments for first and subsequent adjustments . . . 38
5.19 Percentage upward and downward adjustments per adjustment type 40
5.20 Percentage upward and downward adjustments per domain 40
5.21 MdAPE of RF and AF per data set for upwards and downwards adjustments . . 41
5.22 MdAPE for RF and AF per data set for first upwards and downwards adjustments 41
5.23 MdAPE of AFM, AF per data set for adjustments up and down, No. >1 42

5.24 Performance of the AF in percentages for upward and downward adjustments . . 42
5.25 MdAPE of RF and AF per adjustment type for upwards and downwards 43
5.26 MdAPE for RF, AF per adjustment type for first up- and downwards adjustments 43
5.27 MdAPE of AFM and AF per adjustment type for upward and downward, No. >1 43
5.28 Performance AF relative to RF, AFM in % for up and down per adjustment type 44
5.29 MdAPE of RF and AF per domain for upwards and downwards adjustments . . 44
5.30 MdAPE for RF and AF per domain for first upwards and downwards adjustments 44
5.31 MdAPE of AFM and AF per domain for upward and downward, No. >1 45
5.32 Performance AF relative to RF, AFM in % for up and down per domain 45
5.33 MdAPE of RF and AF per data set for large and small adjustments 46
5.34 MdAPE for RF and AF per data set for first large and small adjustments 46
5.35 MdAPE of AFM and AF per data set for large and small adjustments, No. >1 . 47
5.36 Performance of the AF in percentages for large and small adjustments 47
5.37 MdAPE of RF and AF per adjustment type for large and small adjustments . . . 48
5.38 MdAPE for RF and AF per adjustment type for first large and small adjustments 49
5.39 MdAPE of AFM and AF per adjustment type for large and small, No. >1 49
5.40 Performance AF relative to RF, AFM in % for large and small per adjustment type 49
5.41 MdAPE of RF and AF for large and small, upwards and downwards adjustments 50
5.42 MdAPE for RF, AF for first large and small, up- and downwards adjustments . . 50
5.43 MdAPE of AFM and AF for large and small, upward and downward, No. >1 . . 50
5.44 Performance AF relative to RF, AFM in % for large, small, upward, downward . 51
5.45 MdAPE of RF and AF per domain for large and small adjustments 51
5.46 MdAPE for RF and AF per domain for first large and small adjustments 52
5.47 MdAPE of AFM and AF per domain for large and small adjustments, No. >1 . 52
5.48 Performance AF relative to RF, AFM in % for large and small per domain . . . 52
5.49 Timing of adjustments 53

6.1 MdAPE and median %p difference relative to the Adjusted Forecast (AF) for adjustment types addressed in the models 60
6.2 Best performing model on train data per domain 62

A.1 Wilcoxon signed-rank effect size 82
A.2 Cramer's *v* effect size 82

1. Introduction

In this chapter, a description of the company is given in Section 1.1. In Section 1.2, the problem context is described, and problem statements are formulated in Section 1.3. In Section 1.4, the research questions are given, and in Section 1.5 the scope of the research is discussed. Finally, in Section 1.6, the objectives of the research are stated.

1.1 Company Description

Coolblue is a Dutch e-commerce company headquartered in Rotterdam, The Netherlands. The company was founded in 1999 by Pieter Zwart (CEO), Paul de Jong and Bart Kuijpers. Coolblue provides end-to-end solutions for its customers and has an installation and delivery service called CoolblueDelivery, including a bicycle delivery network, CoolblueBikes. The company currently has 20 physical stores, a warehouse in Tilburg, The Netherlands, and is operating in three countries: The Netherlands, Belgium, and Germany. Its product range is very diverse, from computers and tablets, household appliances, kitchen equipment, and sports items to telephony products and garden tools. More than 7,000 employees are working towards the goal of Coolblue: making customers happy. Hence its motto: "Anything for a smile". Coolblue's brand values are: unconventional, friends, go for it, flexible, and simply amaze. These brand values define the culture at Coolblue and are guiding for decision making.

1.2 Problem Context

To ensure that the work within Coolblue's operational departments can be completed, it forecasts the expected workload for the operational departments such that enough but not too many employees can be scheduled. There are four operational departments, referred to as *domains*: the warehouse, customer service, stores, and delivery. For the warehouse, it is forecasted how many orders have to be processed and for the customer service domain, forecasts are made for how many customer contacts they can expect. For the physical stores, a forecast is made of how many visitors they expect to receive, and for the delivery domain, it is predicted how many trips they and their delivery partners expect to have to make.

These forecasts are made over different dimensions. For instance, a distinction is made for the warehouse between parcel, XL packages and white goods, and for the customer service, the customer contacts are divided into voice and non-voice customer contacts. This is done because the operational processes are different for the different dimensions. All operational forecasts are therefore aggregated to the operational dimension for which it is relevant. As a result, percentage-wise, the error of the forecasts is generally lower because the variance is lower for the average of a group of independent, identically distributed random variables than each random variable's variance (Nahmias & Olsen, 2015).

The forecasts for all operational domains are made for one single day. A statistical forecast is generated by a machine learning algorithm from the data science team, which is utilizing historical data from the past two years. The forecasting system of Coolblue is based on a rolling forecast, which means that forecasts are made repeatedly. As the forecasting period approaches, the statistical forecast is updated (Fildes, Goodwin & Lawrence, 2006; Huang, Hsieh & Farn, 2011). Its algorithm is run every night, which means that an update of the statistical forecast is available every day. Generating a statistical forecast starts one year in advance. That means, for a particular key, for a particular date for which a forecast is made, 365 statistical forecasts are generated, one on every single day in that year. Here, the *key* indicates what the forecast is for, e.g. 1_159 is the key for The Netherlands (1) White Goods (159).

There is a team of five business analysts who can make adjustments to the statistical forecasts. The purpose of these judgmental adjustments is to improve the accuracy of the forecasts based on the knowledge that the analysts have and the algorithm does not. This can be an advantageous collaboration since in this way the analysts can account for the effects of special events such as future product promotions, while a statistical algorithm can excel at detecting patterns in huge volumes of data (Fildes et al., 2006).

Adjusting the statistical forecast can and may be done several times and at any time if the analysts have reason to do so. By continuously reviewing forecast data based on the most recent information, the degree of forecast accuracy can be improved and uncertainty in demand forecasting can be reduced (Huang et al., 2011). The analysts can make adjustments to a statistical forecast in three ways: a percentage adjustment, an adjustment in absolute numbers, or override a forecast by hard coding. Override adjustments are mainly applied when stability in and control over the forecast is needed, such as during important campaign periods. Absolute adjustments are, for example, used for a new product release, based on how many of that product will be in stock. In all other cases, percentage adjustments are usually used. An adjustment can apply for a single day, but it can also be implemented for a longer period, for example for two weeks or a month.

In practice, in about 80% of the cases, several adjustments are made for a given key and date. To clarify the adjustment process, Figure 1.1 is shown. This figure shows, for a specific key, the forecast for every day of a month. Two lines can be seen: the green line shows the statistical forecast, and the black line shows the forecast after adjustment(s). If no adjustments (yet) have been made, the statistical forecast and the adjusted forecast are at exactly the same level. However, in this example, it can be seen that for every day of this month one or more adjustments have been made since the black line deviates from the green line for every day.

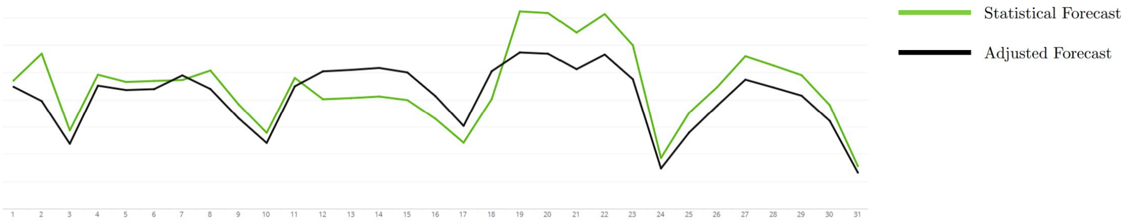

Figure 1.1: An example of an operationally oriented forecast with adjustments

The first adjustment for a specific key and date is made to the pure algorithm-generated statistical forecast. From then on, the statistical forecast and the adjusted forecast are no longer at the same level. For all subsequent adjustments for the same key and date, the adjustments are made to the adjusted forecast, using the statistical forecast as a reference. For the analysts, the most recent statistical forecast (green line) and the most recent adjusted forecast (black line) are displayed at all times.

Usually, on Mondays and Thursdays, the team of business analysts looks quite actively at the forecast for a somewhat shorter-term, up to about one month ahead. These forecasts focus on short-term employee scheduling for the operational domains. Once a month they look more closely at the long-term forecasts from 6 months to a year ahead and make adjustments where necessary. These forecasts focus on long-term capacity planning and are used, for example, as input for hiring/firing decisions, to ensure that capacity is at the right level for the future workload in the operational domains. The team of business analysts is based in The Netherlands but also does the forecasting for Belgium and Germany. Within this team, everyone has a specialism and is responsible for (a part of) the forecasts of an operational domain.

The stakeholders who receive the forecasts do not make any adjustments to them. The forecasts of the expected workload, expressed in the number of orders, customer contacts, visitors and delivery trips, serve as input for their planning process, and based on this they determine how many employees they schedule. Stakeholders usually start with employee scheduling about two to three weeks in advance. This means that the forecast, adjusted or not, that is in the system two to three weeks in advance is crucial. After this, it is still possible to adjust the forecast, but it is less effective because from that moment on scaling up and down employees is significantly more complex. So, in theory, an adjustment can be made up to a day in advance, but the effect of this is minimal. However, some operational parts of the process have more flexibility than others, so the impact of such a late adjustment can differ per operational domain.

1.3 Problem Statement

As mentioned, Coolblue uses a two-step forecasting method. First, a statistical forecast is generated by its algorithm, based on historical data, and then analysts can adjust these statistical forecasts. Because this allows the analysts to take into account exceptional circumstances of which the algorithm is not aware, it can improve the accuracy of the forecasts (Fildes et al., 2009). However, judgmental adjustments, on the other hand, can bias forecasts and reduce forecast accuracy (Fildes & Goodwin, 2007). Optimism, wishful thinking, and lack of consistency are examples of these possible biases (Sanders & Ritzman, 2001).

Accurate forecasting can contribute to increased competitiveness, significant cost savings, and higher customer satisfaction (Moon, Mentzer & Smith, 2003). In addition, for decisions such as human-resource planning, guidance can be provided by accurate short-term forecasts (Fildes & Goodwin, 2007). It is therefore important that there is insight into which judgmental adjustments to algorithm-generated forecasts increase or decrease the forecasting accuracy. However, currently, Coolblue does not keep track of the effects of its adjustments. Although there is a strong expectation within the forecasting team that these adjustments increase the accuracy of its forecasts, there is no clear insight into whether and which adjustments do contribute to the forecast accuracy. Therefore, the following problem statement can be formulated:

> *Problem Statement 1:* Currently, Coolblue has no insight into which judgmental adjustments to algorithm-generated forecasts increase or decrease the accuracy of its operationally oriented forecasts.

Coolblue's adjustment method may be enhanced to raise the accuracy of its operationally oriented forecasts, based on the results of which judgmental adjustments increase or decrease the accuracy. To this end, it must be investigated how analysts' judgment can be combined as effectively as possible with the algorithm's forecasts. There are several possible methods to improve the accuracy of adjustments. Which one is most suitable, depends on the inefficiencies and biases that may be identified. For example, statistical procedures can be applied to correct adjustments or restrictions can be built into the software with which adjustments are made (Fildes et al., 2009). It is not yet known what the best procedure for combining analysts' judgment with the algorithm-generated forecasts is for Coolblue. Therefore, the following problem statement can be formulated:

> *Problem Statement 2:* Since different judgmental adjustments may have a different effect on the forecast accuracy, Coolblue wants to know the most effective procedure for combining the algorithm-generated forecast and analyst judgment.

1.4 Research Questions

Two research questions have been formulated. These research questions indicate the design of the research and aim to ultimately achieve the research objective. This research objective is to propose solutions for the problem statements described in Section 1.3 and will be further elaborated on in Section 1.6. The behavioral operations research goals as defined by Donohue, Özer and Zheng (2020) underlie the research questions. The first research question aims to understand the role of human behaviour, in this case behaviour of the analysts, in operations. This question is formulated as follows:

> *Research Question 1:* What are the specific strengths and weaknesses of analysts when adjusting algorithm-generated forecasts, with regard to the characteristics of adjustments, and when do these adjustments increase and decrease the accuracy of operationally oriented forecasts?

The second question is focused on designing an adjustment procedure that takes into account the behavioral tendencies of the analysts to improve the accuracy of the operationally oriented forecasts at Coolblue. This question is formulated as follows:

> *Research Question 2:* Given the strengths and weaknesses of analysts when adjusting algorithm-generated forecasts, what is the most effective procedure to make adjustments to algorithm-generated forecasts?

1.5 Research Scope

The scope of the research is determined in order to make the research feasible within the limited time available for a master's thesis. Within the scope of this research, the focus will be on Coolblue's operational forecasting process. In particular, the research will focus on adjustments that are made to forecasts for which an algorithm-generated forecast is available. Because the forecasting process is the same for all four operational domains of Coolblue, they will all be part of the scope of this research.

1.6 Objectives

The objective of this research is twofold. Firstly, there is a practical objective that focuses on generating concrete insights for Coolblue about its operationally oriented forecasting process. The second objective is academic and is aimed at contributing to the judgmental forecasting literature. These objectives are further elaborated in Section 1.6.1 and Section 1.6.2.

1.6.1 Practical Objective

The first part of the practical objective of this research is to gain insight into the effects of the adjustments that analysts make to the algorithm-generated forecasts. More specifically, which characteristics of these adjustments play a role and affect the accuracy of the forecasts. This provides a clearer understanding of the current interaction between the analysts and the statistical forecasts created by the algorithm of the data science team, and is in line with Research Question 1. The second part of this practical objective is to propose an improved procedure for adjusting algorithm-generated forecasts, in line with Research Question 2. For this purpose, the insights obtained concerning the characteristics of the adjustments will be used. The ultimate goal is to increase the accuracy of Coolblue's operationally oriented forecasts and to reduce the time spent on making adjustments.

1.6.2 Academic Objective

The academic objective is focused on the contribution of this research to the existing literature. Much research has already been done into judgmental adjustments to statistical forecasts, almost always focusing on sales forecasting at a highly disaggregated, Stock-Keeping Unit (SKU) level (Davydenko & Fildes, 2013; Fildes & Goodwin, 2007; Khosrowabadi, Hoberg & Imdahl, 2022; Syntetos et al., 2009). In the context of this research, forecasts are made on the expected workload for the operational domains on an aggregated level. A contribution to the literature will be whether the same characteristics of adjustments will play a role when forecasting in this context, and what the effects of these adjustments are on the accuracy of the forecasts.

In addition, in most studies, adjustments are made to the statistical forecasts in only one way. This research looks at different types of adjustments: percentage, absolute, and override, and the effects of them on forecast accuracy. Moreover, there is not yet much research into rolling forecasts within the judgmental forecasting literature. This study examines the timing of adjustments and its influence on accuracy. It also focuses on the effects of subsequent adjustments to forecasts that already have been adjusted, a topic on which only minimal research has been done (Van den Broeke et al., 2019; Perera, Hurley, Fahimnia & Reisi, 2019). Furthermore, one of the objectives of this study is to contribute to the existing literature on judgmental adjustments by investigating the difference in the characteristics and performance of adjustments that are directly linked to a period with disruptions, in this case the COVID-19 pandemic, and the adjustments that are not directly linked to disruptions.

2. Literature Review

In this literature review, first statistical forecasting and integrating forecasting methods are discussed, in Section 2.1 and Section 2.2, respectively. After that, in Section 2.3, judgmental adjustments to statistical forecasts will be discussed in detail.

2.1 Statistical Forecasting

For many companies, forecasting plays a key role in their operational function. Based on forecasts, all business planning is determined (Nahmias & Olsen, 2015; Petropoulos, Kourentzes, Nikolopoulos & Siemsen, 2018). For example, to guide workforce planning decisions, it is necessary to make accurate forecasts (Fildes & Goodwin, 2007). These forecasts can be made in several ways, one of which is statistical forecasting also called quantitative forecasting. With this type of forecasting, data analysis is used to create forecasts (Nahmias & Olsen, 2015). In Section 2.1.1, several quantitative forecasting methods will be discussed.

2.1.1 Quantitative Forecasting Methods

There are a variety of quantitative forecasting methods, examples are exponential smoothing and its variants, autoregressive integrated moving average (ARIMA) models, neural networks, and a lot more (Petropoulos et al., 2018). Forecasting accuracy can be significantly enhanced if the most suitable model is selected. According to Fildes and Petropoulos (2015b), even improvement margins of 25 to 30 percent are possible. Four main forecasting approaches are suggested by Fildes, Nikolopoulos, Crone and Syntetos (2008) of which three are quantitative forecasting methods: extrapolation methods, causal and multivariate methods, and computer-intensive methods. Each of these methods will be discussed below.

Extrapolation Methods

Extrapolation is a method of forecasting exclusively based on historical data of the time series (Fildes et al., 2008). This method is also called 'time series method' or 'naive method'. A time series is a set of observations gathered at discrete, generally equally spaced moments in time. By inferring information from the pattern of previous observations, future values of the series can be forecasted (Nahmias & Olsen, 2015). When dealing with enormous amounts of data, extrapolation models are frequently utilized (Petropoulos, Makridakis, Assimakopoulos & Nikolopoulos, 2014). Exponential smoothing is a classic extrapolative method that is commonly used in industry (Fildes, 1992). There are many variants of exponential smoothing. According to Hyndman, Koehler, Ord and Snyder (2008), the exponential smoothing 'family' includes thirty distinct methods. Contrasting a simple moving average, where simply the unweighted average of the N most recent observations is used (Nahmias & Olsen, 2015; Petropoulos et al., 2014), all variants of exponential smoothing have the property of giving relatively more weight to recent observations compared to older observations. The forecast is a weighted combination of these historical observations. As observations get older the weight decreases exponentially as reflected in the name 'exponential smoothing' (Hyndman et al., 2008). The main advantages of exponential smoothing methods are that they are easy to implement, the computational power required is relatively low, and there is no requirement for long series (Petropoulos et al., 2014).

One of the exponential smoothing methods is Single Exponential Smoothing (SES). Here, no seasonal patterns or trends are assumed and only one smoothing parameter is employed. Stationary data can be forecasted quite accurately by simple exponential smoothing (Petropoulos et al., 2014). To expand this method and make it more suitable for data with a trend, Holt (2004) added a smoothing parameter for the trend. Here, a trend refers to a time series' propensity to display a steady pattern of growth or decline (Nahmias & Olsen, 2015). After that, by adding a seasonal smoothing factor, the Holt trended model was further expanded, called the Holt-Winters approach, where multiplicative or additive seasonality in the data is assumed (Winters, 1960). Here, seasonality means that there is a pattern in the data that at fixed intervals repeats itself (Nahmias & Olsen, 2015). And, to have more control over the trend's long-term extrapolation, a dampening factor was introduced by Gardner and McKenzie (1985). This method is known as damped exponential smoothing.

Time series extrapolative methods, such as exponential smoothing, have a successful track record in practice and are therefore still very attractive (Gardner, 2006). In addition, it is not necessarily the case that statistically more sophisticated methods outperform these more simple methods (Makridakis & Hibon, 2000). Furthermore, the fact that these methods are simple and transparent might influence the likelihood that people will use them because, for example, it could be that they feel more comfortable with them than with more complex and less transparent methods (Dietvorst, Simmons & Massey, 2015).

Other more complex extrapolation methods that are widely utilized are the Auto-Regressive Integrated Moving Average (ARIMA) method and its variations, also known as Box-Jenkins models. Trends, seasonality, errors, and non-stationary characteristics of a time series, can all be accounted for using these methods (Box, Jenkins, Reinsel & Ljung, 2015). Autoregression considers the correlation of series observations with other observations, separated by a fixed number of periods, to find repeating patterns in data. These ARIMA models need a huge amount of data to be meaningful, but for forecasting these models have shown to be very powerful (Nahmias & Olsen, 2015).

Causal and Multivariate methods

For time series forecasting, causal models have been commonly utilized (Armstrong, 2006). Causal and multivariate methods, in contrast to extrapolation methods, incorporate data of other variables than the time series itself that is in some way linked to the forecasted value (Nahmias & Olsen, 2015). For these methods, which are forms of regression analysis, it is assumed that the dependent variable, the one that is being predicted, and one or more independent, explanatory variables have a cause-and-effect relationship (Armstrong, Green & Graefe, 2015; Fildes et al., 2008). For causal models to be applied, first the dependent and causal variables must be identified. Subsequently, the size and the direction of the relationships must be estimated. Causal models are effective for forecasting the impact of government and business decisions since they may include policy factors such as a product's price. For instance, if there is a major change in prices, causal models can be useful in such a situation to estimate the consequences of such a change (Armstrong, 2006). Furthermore, the issue of non-stationary time series can be dealt with by these methods, for example by accounting for a trend in the data (Nahmias & Olsen, 2015; Trapero et al., 2013).

Computer-intensive methods

Advances in computational power and intelligent software, such as machine learning, have led to new developments in computer-intensive forecasting methods (Fildes et al., 2008). Machine learning has its roots in artificial intelligence. It is described as computer algorithms that are able to learn to solve a problem, and improve problem-solving performance, through model building by utilizing training data and experience from previous executions (Grossmann & Rinderle-Ma, 2015; Mitchell, 1997). Although currently these computer-intensive methods are not extensively utilized in demand forecasting in supply chains, three major data mining techniques have shown to be useful in operations research. These are: Artificial Neural Networks (ANN), Support Vector

Machines (SVM), and decision tree classification with regression algorithms (Perera et al., 2019). In contrast to causal models, machine learning methods, including ANN, SVM and trees, do not imply a specific relationship between variables. These methods entail searching the functional form space and estimating parameters (Ali, Sayın, Van Woensel & Fransoo, 2009). Recently, the popularity of computer-intensive methods is growing in operations research (Perera et al., 2019). However, there are also potential drawbacks of using this computer-intensive methods for forecasting in terms of user acceptance. In contrast to time series methods, which are relatively easy to understand, machine learning methods are complex. Machine learning methods are frequently experienced as black boxes, with little or no information on how forecasts are made or information on which data elements were important in the generation of the forecasts. Nevertheless, for users, this information about forecasts is often critical (Sagaert, Aghezzaf, Kourentzes & Desmet, 2018).

2.1.2 Limitations of Quantitative Forecasting Methods in Isolation

In general, it is highly difficult to rely exclusively on statistical forecasting approaches due to volatile company dynamics and challenges with access to credible domain information. This is especially the case in situations where uncertainty is high and decisions can have a significant impact, such as with demand forecasting (Sanders & Manrodt, 2003). Even though a specific statistical forecasting approach can produce fairly accurate forecasts in most cases, there may be information regarding future demand that is not included in the series' historical data (Nahmias & Olsen, 2015). Therefore, integrating computer-based methods for forecasting with the broader organizational context is often advantageous for the company's operations. Judgment plays a considerable role in this integration process (Fildes et al., 2008). In practice, the forecasts generated by quantitative methods are frequently utilized as starting points which are subsequently adjusted, averaged, or otherwise combined (Perera et al., 2019). In the next section, it will be discussed how human judgment can be integrated in the forecasting process of companies.

2.2 Integrating Forecasting Methods

Statistical forecasts should not be utilized without including known information. Incorporating such known information into the forecast has to be done manually, and can, for example, consist of information about a future promotional campaign (Nahmias & Olsen, 2015). This processing of information in the forecast, which is not reflected in the forecasting model, is the primary manner that value can be added to the forecast by human judgment (Lawrence et al., 2006). This is why, despite the availability of a wide range of software and significant technological advancements, most organizations still forecast using a combination of statistical approaches and judgment. A survey by Fildes and Petropoulos (2015a) showed that only 28.7% of organizations solely used statistical methods for forecasting. And, even if companies make exclusively use of statistical forecasting methods, human judgment would still have a role to play. Judgment and forecasting are arguably inextricably linked. Human judgment plays a role in, for example, choosing the statistical forecasting method to employ as well as which data sets to use (Bunn & Wright, 1991; Goodwin, 2002; Petropoulos et al., 2018). But human judgment can play an even bigger and more direct role in the forecasting process of companies. There is a lot of research confirming that the integration of human judgment and statistical methods in the right manner may lead to more accurate forecasts. According to Goodwin (2002), this is not surprising given that the strengths and weaknesses of these two approaches to forecasting are complementary. Statistical methods can make the best use of vast amounts of data and are rigid yet consistent. Human judges, on the other hand, are adaptive and can consider one-time occurrences, but they are subject to cognitive biases, can only consider small amounts of data, and are inconsistent (Goodwin, 2002). Given the potential advantage of integration, the use of human judgment in the forecasting process should not be excluded. In Section 2.2.1 two different methods for integrating human judgment and quantitative methods will be discussed.

2.2.1 Integration of Human Judgment and Quantitative Methods

The methods for the integration of human judgment and quantitative models can be classified into two main categories (Arvan, Fahimnia, Reisi & Siemsen, 2019). The first one is combining forecasts, this involves creating separate statistical and/or judgmental forecasts and then merging or averaging them, by either further judgment or a formal averaging mechanism, into a final forecast (Franses & Legerstee, 2011a). When this integration takes place by applying a statistical method, this is called 'mechanical integration' (Goodwin, 2002). Judgmentally adjusting statistically produced forecasts is the second and perhaps most extensively used strategy in modern business practice. This method consists of two steps. First, an automated baseline is set, which is generated by a system that uses statistical forecasting processes based on historical data. Subsequently, these initial forecasts can be adjusted judgmentally to manually integrate contextual knowledge and unmodeled factors (Alvarado-Valencia, Barrero, Önkal & Dennerlein, 2017). The term for this procedure is 'judgmental adjustment' and can be classified as 'voluntary integration' (Fildes et al., 2009; Goodwin, 2002). Some integration/combination methods will be discussed in Section 2.2.2, and judgmental adjustments to statistical forecasts will be discussed in detail in Section 2.3.

2.2.2 Integration Methods

The term 'integrated forecasting methods' refers to procedures where an expert's judgment is incorporated into quantitative forecasting models (Arvan et al., 2019). A variety of methods with differing levels of complexity is available, and they all have the same goal: improving the efficiency of judgmental forecasting and reducing the bias (Trapero et al., 2013). Some (mechanical) integration methods will be discussed below.

Blattberg-Hoch method
Blattberg and Hoch (1990) were the first to develop a method to integrate human judgment and statistical models. In this method, just as much weight is given to the statistical forecast as to the human judgment (50% model + 50% manager). This combination was found to be consistently outperforming each input separately in several contexts. For example, in a study of Franses (2011) conducted with data of a pharmaceutical company. In this study, a large database was analyzed, with model forecasts and expert forecasts, and the most accurate results were obtained when applying the 50%−50% rule. The fact that statistical models and human judges have different, but complementary strengths and weaknesses was found to be the reason for the fact that a strategy as simple as averaging is so successful (Blattberg & Hoch, 1990). For instance, managers can identify and evaluate exceptional circumstances, whereas models can consistently assign optimal weights to vast amounts of data. However, according to Fildes et al. (2009), the success of the Blattberg-Hoch method is highly dependent on the circumstances. In their study, they applied a variant of the method where the two forecasts were not independent and it appeared that only for positive adjustments this method improved forecast accuracy.

Divide-and-conquer method
Another integration method is the divide-and-conquer method. In this method, the forecaster is told how the system forecast was generated but does not gain insight into both the time series and the system forecast, i.e. the computer-modeled information. The forecaster can then indicate whether s/he would like to adjust the system forecast and by how much, based on possible additional information available to the forecaster. The idea behind this method is that information about historical data is already included in the system forecast and that it must therefore be prevented that the forecaster includes this again in the forecast, otherwise this historical data will be inefficiently overweighted (Alvarado-Valencia et al., 2017). In this method, the forecaster can focus all effort on important, unmodeled information to decide whether changes are required to the system forecast, by transferring the task of modeling the available structured data to the system (Jones & Brown, 2002). Biases like anchoring and adjustment, where people have a propensity to anchor too strongly on an initial value when making an adjustment (Tversky & Kahneman, 1974), may be reduced by the divide-and-conquer method. In an experiment conducted in a study of Jones, Wheeler, Appan and Saleem (2006), it was found that judgment

performance improved as the divide-and-conquer strategy was approached. Although this study was not performed in a demand forecasting context, they claim that judgment could be improved by applying these findings in numerous contexts. On the other hand, the advantages of this method might be offset by a lack of information availability. In the study of Alvarado-Valencia et al. (2017), they found that the use of information constraints to decrease forecasters' bias did not succeed. It turned out that the forecasters were unable to judge the quality of the system forecasts or the amount of adjustment required due to the lack of access to the system forecasts.

Bootstrapping method
Bootstrapping is another mechanical integration method, that is used to smooth out forecasters' inconsistencies when applying their knowledge. The goal of bootstrapping is to figure out how forecasters come up with their forecasts or adjustments (Arvan et al., 2019). By regressing a forecaster's forecasts against the information he or she has used, a model of a forecaster is developed. The idea is that '*the model of the man will be more accurate than the man*' since the model is more consistent in applying the man's rules (Armstrong, 2006). For cross-sectional forecasting, it is a well-established method. When there is a lack of or low-quality historical data on the variable to be forecasted, bootstrapping might be effective. And it is best used in complicated settings with a high degree of unreliability in judgments and some validity in expert judgments. Several studies have found that forecasts made using bootstrapping had a higher accuracy than forecasts made using unassisted judgment. Overall, there was a considerable improvement in accuracy, with the error rate dropping by roughly 6% (Armstrong, 2001). However, there are several drawbacks to bootstrapping, and some researchers have questioned its effectiveness. Lawrence and O'Connor (1996) conducted a study in a time series forecasting setting and when it came to forecasting accuracy, this study discovered that the model-of-man was not superior to man. One of the explanations they provide is that for each new time series man may utilize a different forecasting model, based on configuration and contextual cues available. As a result, using bootstrapping to analyze time series is less useful. The success of this method is therefore highly dependent on the forecasting task and its context.

The methods described in this section are three examples of integration methods but there are a lot more. It can be concluded that the accuracy of forecasts can benefit from the valuable and complementary contributions of statistical methods and human judges. However, the most suitable integration method will be determined by the particular circumstances of each context (Goodwin, 2002). In the next section, the currently most widely used integration method will be discussed: judgmental adjustments to statistical forecasts.

2.3 Judgmental Adjustments

It is common practice in industry to make judgmental adjustments to statistical forecasts, and for the effective planning of succeeding supply chain operations, the accuracy of these adjustments is critical (Fildes & Goodwin, 2007; Perera et al., 2019). However, depending on the person who is judgmentally adjusting the forecast, the available information, and a variety of other factors, the accuracy can be either improved or deteriorated by human involvement (Davydenko & Fildes, 2013; Fildes, Goodwin & Önkal, 2019). Whether judgmental adjustments to statistically produced forecasts are advantageous for accuracy and operational performance has therefore been the subject of debate in numerous studies in different contexts (Mathews & Diamantopoulos, 1986; Sanders, 1992; Fildes et al., 2009).

In a study by Fildes et al. (2009), they analyzed forecast data from four supply chain companies. Although the overall accuracy of forecasts improved through adjustments for three of the four companies, these accuracy improvements were highly dependent on the type of adjustment that was made. They discovered, for example, that relatively smaller adjustments generally reduce forecast accuracy, while larger adjustments tended to result in higher average increases in accuracy. Furthermore, it was significantly more likely for downward adjustments to improve accuracy than it was for upward adjustments. Upward adjustments were also more frequent in the wrong direction than downward adjustments. The bottom line is that some types of adjustments were way more effective than others. This indicates that to improve forecasting accuracy, it's crucial to make effective use of human adjustments (Fildes & Goodwin, 2007). However, people are not completely rational decision-makers, they are influenced by biases (Tversky & Kahneman, 1974). It is critical to get a deeper understanding of how underlying behavioral aspects influence adjustment characteristics and effectiveness, and how judgmental adjustments might improve forecasting accuracy. Therefore, this section focuses on the characteristics and timing of adjustments, and their effects on forecast accuracy.

2.3.1 Motivations for Judgmental Adjustments

Two steps can be distinguished in the process of judgmental adjustment of a statistical forecast. The first step consists of making a decision on whether an adjustment of the statistical forecast is required. If the decision is that an adjustment is required, the second step consists of deciding about the direction and size of the adjustment (Arvan et al., 2019; Lawrence et al., 2006). This section will focus on the first step of this process, the motivations of forecasters to adjust a statistical forecast. The direction and size of adjustments will be further discussed in Section 2.3.3 and Section 2.3.4, respectively.

In general, if there is contextual information that is not incorporated in the statistical forecast, an adjustment is required. In many industries, it is common for forecasters to use this contextual information, but also expertise, intuition, and/or analysis, as a basis for making the decision whether or not an adjustment is required (Fildes & Goodwin, 2007; Franses & Legerstee, 2009; Webby & O'Connor, 1996). However, there are some specific reasons or motivations for forecasters to make an adjustment and some of them will be discussed in this section.

Firstly, a reason for adjusting a system-generated forecast may be a desire that a forecaster has for a sense of ownership of a system that is experienced by the forecaster as a 'black box'. Because, when forecasting is left entirely to a statistical model, a sense of loss of control and ownership may be felt by forecasters (Goodwin, 2002). Especially the use of software packages for forecasting, a lack of training for forecasters, or limited transparency of in-house developed forecasting systems, may lead to forecasters making adjustments in order for them to perceive they have a better comprehension of the derivation of forecasts (Syntetos, Kholidasari & Naim, 2016). Another motivation for adjusting a system forecast comes from the 'illusion of control' effect. Here, the confidence people have in a forecast is actually higher if they have adjusted this forecast (Kottemann, Davis & Remus, 1994). So, this is a reason for them to adjust a forecast regardless of the adverse effects that this could have since they believe it improves forecasting performance anyway. This is also one of the reasons that unnecessary adjustments are often made to system-generated forecasts (Baecke et al., 2017). Another last thing that partly has to do with control of the forecasting process and the feeling of responsibility is the 'tinkering with data' effect, i.e. making small adjustments. However, the major reason for forecasters to tinker with the data is to simply show, to colleagues or managers for example, that they are actually working on their task (Baecke et al., 2017; Fildes et al., 2009). This ability to show their activities is therefore also one of the reasons for adjusting system forecasts.

As mentioned earlier, the processing of contextual information or information about special events in the forecast, which is not reflected in the forecasting model, is the primary way in which value can be added to the forecast by human judgment (Lawrence et al., 2006). This is therefore one of the motivations for forecasters to make an adjustment. However, certain types of events can play a bigger role than others in the decision-making process of whether or not to adjust a forecast. Fildes and Goodwin (2007) surveyed 149 forecasters to see how they utilized their judgment. One of the questions was directed at the reasons for adjusting. Almost 86% of the respondents declared that at least one of the special events, such as changes in regulations, weather or sporting events, that were given in the survey had a significant or extremely significant impact on how they utilized their judgment. This confirms that indeed, special events are a reason for forecasters to adjust system-generated forecasts. However, some reasons are more likely to trigger adjustments than others. The major reasons for forecasters to adjust a system-generated forecast were promotional activity and price changes. Competitor activities, on the other hand, were less likely to prompt forecasters to make an adjustment, perhaps because they are rarely aware of it in advance.

However, even if forecasters do not have additional information concerning special events, they make adjustments to the system-generated forecasts. One of the reasons for this is that forecasters may perceive noise in the time series, that is associated with random fluctuations, as patterns (Harvey, 1995; O'Connor, Remus & Griggs, 1993). This then leads to forecasters making, typically, small adjustments to the system-generated forecasts (Syntetos et al., 2016). This adjusting of forecasts based on noise in the data is also one of the causes of unnecessary adjustments (Baecke et al., 2017; Fildes et al., 2009). The level of noise is therefore an important contributor to the complexity of time series and, in general, forecasting tasks are therefore more complex if the data contains a lot of noise (Arvan et al., 2019).

There are many reasons for forecasters, valid or not, to adjust a system-generated forecast. In practice, this is reflected in the high frequencies of adjustments in multiple studies that were conducted on judgmental adjustments over time. By analyzing the survey results of 96 participants, who were responsible for sales forecasting in US companies, on their forecasting procedures, Sanders and Manrodt (1994) found that judgmentally adjusting system-generated forecasts was very common. Of the participants, 44.7% indicated that they always adjust the system-generated forecasts, which means that in those companies 100% of the system forecasts have been adjusted. Only 9.3% of the participants indicated that they never adjust the system forecasts. In the study of Fildes et al. (2009), referred to earlier, over 60,000 SKU level demand forecasts from four companies were analyzed. They found that for three out of four companies in their study, the number of system-generated forecasts that were adjusted was between 63% and 91%. In the study of (Syntetos et al., 2009), in about 74% of the cases, adjustments were made. Franses and Legerstee (2009) analyzed 30,517 forecasts for pharmaceutical products on SKU level and found that, overall, 89.5% of the system-generated forecasts were adjusted. In a study of Trapero et al. (2013), an analysis was conducted on 18,096 forecasts from a manufacturing company and here the vast majority (82%) of the system-generated forecasts was judgmentally adjusted. Finally, Van den Broeke et al. (2019) considered different time horizons in investigating the impact of judgmental adjustments. They found that for 7 out of 8 cases studied, at time horizon zero (the time horizon closest to the date of sale), the majority of forecasts was adjusted, in accordance with the above findings. In addition, they discovered that in general, as the time horizon shortened, so closer to the date of sale, the likelihood of a forecast being adjusted increased (Van den Broeke et al., 2019).

2.3.2 Accuracy of Judgmentally Adjusted Forecasts

Statistical forecasts have been claimed to be as good as or better than judgmental forecasts in several studies (Carbone, Andersen, Corriveau & Corson, 1983; Sanders, 1992; Webby, O'Connor & Lawrence, 2001). However, a gradual paradigm shift has occurred, based on further research, toward recognizing the possible value-adding function of human judgment in improving the accuracy of forecasts (Armstrong, 2006; Goodwin & Wright, 1993; Lawrence et al., 2006). Something important to note is that where there is criticism of human judgment due to its negative impact on decision quality, these results often have been obtained from experiments. And even though the designs of experiments may be complicated and hence realistic, frequently they are stylized with just quantitative information. In such cases, human judgment will always lose against statistical algorithms. In general, the results of empirical studies focusing on the quality of human judgment, conducted in a business context with access to trustworthy market intelligence, are often far more positive with regard to human ability (Fildes et al., 2009; Leitner & Leopold-Wildburger, 2011). Various studies that used company data showed that judgmental adjustments increase the accuracy of forecasts (Fildes et al., 2009; Syntetos et al., 2009).

Nevertheless, judgmental adjustments also have a downside. For judgments under uncertainty, it is suggested by empirical evidence that they are subject to different types of cognitive biases and that they are inherently non-optimal (Fildes & Goodwin, 2007; Tversky & Kahneman, 1974). Judgmental adjustments have been proven to be especially subject to these biases and inefficiencies (Fildes et al., 2009). Here, the concept of bias is referred to as a systematic pattern or a deviation in the judgment from the rational decision (Haselton, Nettle & Andrews, 2005). According to Tversky and Kahneman (1974), to reduce the complexity of forecasting tasks, and convert them to simpler judgmental operations, people use a limited number of heuristic principles. And while these heuristics can be quite useful, they can also lead to severe and systematic errors (biases). When it comes to decision-making, biases are more concerned with deviations in decision *outcomes*, and heuristics are more concerned with deviations in the *processes* underlying decision-making (Bendoly, Croson, Goncalves & Schultz, 2010).

Despite the biases that judgmental adjustments are subject to, it is widely supported that processing of reliable contextual information and information about special events that is not included in the statistical model, can, and is likely to, improve the accuracy of forecasts (Fildes et al., 2009; Goodwin, 2002; Lawrence et al., 2006; Webby & O'Connor, 1996). In terms of contextual information, there are several reasons to judgmentally adjust a system forecast. For example, people are better aware of price changes, holidays, promotions and advertising activities, government policy, activities of competitors, and changes in regulations (Perera et al., 2019). The fact that people are better aware of changes in regulations is especially important in the past two years, due to the COVID-19 crisis and the measures that were in place. The consequences of such measures can be estimated by managers, but due to a lack of previous data, applying a statistical technique to these kind of occurrences would be difficult (Fildes & Goodwin, 2007). This type of knowledge, contextual knowledge, is regarded as the most likely to contribute to the accuracy of forecasts in comparison to technical knowledge (Perera et al., 2019). Technical knowledge is concerned with the understanding of the techniques and methodologies utilized in the generation and analysis of statistical forecasts (Carbone et al., 1983; Sanders & Ritzman, 1992). However, although this type of knowledge is less likely to contribute to forecast accuracy, it may be advantageous. The behavior of forecasters can even differ based on this type of knowledge. For example, if the forecaster knows how the statistical forecast is derived and is therefore aware that the statistical method does not explicitly include certain information in the forecast, the forecaster will then probably make a larger adjustment to compensate for this (Perera et al., 2019).

An example of a study suggesting that knowledge about specific situations may be the most important factor in improving a forecast is the study of Mathews and Diamantopoulos (1986). In this empirical study, data on forecasts for 281 products of a UK healthcare company was analyzed to see what the consequences of human judgment are for the accuracy of forecasts. They found, even though there may be introduction of bias, that there was an overall improvement in the accuracy of forecasts by applying human judgment to quantitatively generated forecasts. In an empirical study by Syntetos et al. (2009), they analyzed intermittent demand forecast data from 138 SKUs of a large international pharmaceutical company's UK branch. They found, in accordance with Mathews and Diamantopoulos (1986), that it is advantageous to judgmentally adjust statistical forecasts based on additional knowledge available to forecasters. Finally, in the study of Fildes et al. (2009), the accuracy of forecasts was in general increased for 75% of the companies studied, by judgmentally adjusting the statistical forecasts based on information exclusively available to forecasters.

Processing contextual information in the operationally oriented forecasts is a motivation for judgmental adjustments in the context of this study. Based on the fact that the accuracy of forecasts is likely to be improved by judgmental adjustments that are based on contextual information that is exclusively available to forecasters, it is expected that in the context of this research the judgmental adjustments by analysts will contribute to a higher accuracy of the operationally oriented forecasts. The following hypothesis can therefore be formulated:

> ***Hypothesis 1:*** *Overall, judgmental adjustments increase the quality of an operationally oriented forecast. That means, the error of judgmentally adjusted forecasts is, on average, lower than that of algorithm-generated forecasts.*

2.3.3 Direction of Judgmental Adjustments

After the decision is made that an adjustment is needed, the second step in the process of judgmental adjustment is the decision on the direction and size of the adjustment (Arvan et al., 2019; Lawrence et al., 2006). This section will focus on the direction of adjustments. When an adjustment is made in an upward (downward) direction with respect to the statistical forecast, this is referred to as a positive (negative) adjustment. To generate accurate forecasts, it is critical that adjustments are in the right direction since especially wrong-sided adjustments have a significant negative impact on forecast accuracy (Fildes et al., 2009). According to several studies, forecasters are typically competent at determining whether an adjustment should be upwards or downwards. For example, McNees (1990) found that the majority (74%) of judgmental adjustments were in the right direction, by analyzing 841 judgmentally adjusted forecasts in a macroeconomic context. Only 26% of the adjustments were in the wrong direction. That means, the adjustment was upwards when a downward adjustment was required and the other way around, which will cause the accuracy to deteriorate anyway. The same was seen in the study of Mathews and Diamantopoulous (1990). Here those forecasts that were most in need of an adjustment were selected by the forecasters and, overall, adjusted in the right direction. Furthermore, in the study of Fildes et al. (2009), where data of three manufacturing companies and one retailer was analyzed, they found that, in general, the adjustments were made in the right direction. For the manufacturers, only 34.4% of the positive adjustments were in the wrong direction, and only 28.3% of the negative adjustments. For the retailer, however, slightly more than half of the positive adjustments were in the wrong direction (50.6%) but only 20.5% of the negative adjustments were in the wrong direction. Finally, Petropoulos, Fildes and Goodwin (2016) utilized a large multinational data set with statistical forecasts, adjustments, and actuals for pharmaceutical product demand to investigate the impact on subsequent judgmental adjustments after a big loss had taken place. In line with the previous studies, it was found that the majority (> 60%) of adjustments was in the right direction. Based on the findings of these studies, it is expected that the analysts in the context of this research will also be competent in determining the appropriate direction of an adjustment. The following hypothesis can therefore be formulated:

> ***Hypothesis 2a:*** *In general, analysts are competent in making adjustments in the right direction. That means, adjustments are in the right direction in more than* 50% *of the cases.*

Several studies on judgmental adjustments show that the effectiveness of judgmental adjustments is conditional on the characteristics of an adjustment (Van den Broeke et al., 2019; Fildes et al., 2009; Syntetos et al., 2009). One of the characteristics is the direction, and it is often seen that forecasts that are adjusted upwards are far less likely to enhance accuracy than those that are adjusted downwards. It is also often seen that the frequency with which upward adjustments are made is higher than that of downward adjustments. An example of this is the study of Fildes et al. (2009). In this study, of the 12,863 judgmental adjustments, 55% were upwards and 45% were downwards. And, it turned out that the downward adjustments were significantly more effective than the upward adjustments. Furthermore, the upward adjustments were also made more often in the wrong direction. Almost the same results were obtained in a study of Syntetos et al. (2009). In this study, they found that of the 3,659 judgmental adjustments, the majority (55.7%) were adjusted upwards. And besides, it was also the case here that the upwards adjustments were much less effective than the downward adjustments. Accordingly, in the study of Trapero et al. (2013), 72% of the judgmental adjustments were in an upward direction and only 28% were in a downward direction. Again here, the downward adjustments were beneficial for accuracy anyway and more effective than the upward adjustments. The upward adjustments only made a positive contribution under certain conditions, i.e. in non-promotional periods. Finally, in the study of Van den Broeke et al. (2019), they also found that in general, at time horizon zero, the adjustments were more frequently positive as opposed to negative. However, this is not the case for all horizons, their results show that adjustments are more frequently positive with a decreasing time horizon, i.e. the closer they are made to day of sales (Van den Broeke et al., 2019).

Fildes et al. (2009) found that the magnitude of bias in judgmental adjustments is influenced by the direction of an adjustment. Typically, the bias in downward adjustments is less than the bias in upward adjustments, and therefore they are inclined to have a better performance in terms of accuracy. Most of the time, the tendency of forecasters to adjust more upwards than downwards, and the fact that upward adjustments are relatively less effective in general, are explained by an optimism bias (Fildes et al., 2009; Syntetos et al., 2009; Trapero et al., 2013). If reliable information indicates that an upward adjustment is needed, the optimism bias may cause this adjustment to be too optimistic, i.e. excessively high. In addition, optimism bias may trigger an upward adjustment in the case where such information is not available (Syntetos et al., 2009). In the study of Fildes et al. (2009), this can be seen by the fact that the majority of upward adjustments were too optimistic. For one of the two groups of companies 66.4% of the upward adjustments were too optimistic and for the other group even 83%. This was partly because the upward adjustments were quite frequently in the wrong direction and partly because the adjustments were in the right direction but too large. Downward adjustments, on the other hand, were far less likely to be too optimistic, only 46% were too optimistic in this case, and they were also less frequently in the wrong direction.

However, there is another possible contributor to bias that may also lead to the fact that in many cases upward adjustments are made more frequently but are less effective, which is the asymmetric loss function. In forecasting, when the forecast differs from the actual value there is error, and although this is inevitable, it causes costs or losses. This is true in both over- and under-forecasting situations. However, the costs for over-forecasting and under-forecasting might be different and can therefore influence the forecasting process (Lawrence & O'Connor, 2005). In practice, it is seen that for companies the loss function is often asymmetrical. This means that errors of the same magnitude will have different losses for a company depending on whether this is a positive or negative forecast error (Goodwin, 1996). For the customer service oriented manufacturing company in the study of Goodwin (1996), this is the case. For this company, it is about two to three times more expensive to make a forecast that is too low in comparison with a forecast that is too high. This asymmetry was especially present because when a forecast was too low this necessitates over-time working, and perhaps more importantly, due to delays in the delivery of orders, there is a potential loss of customer goodwill. Excessive stocks as a consequence of over-forecasting, on the other hand, could easily be cleared (Goodwin, 1996). Accordingly, it is claimed that for a supply chain the cost of lost sales is higher than the cost of excessive stocks, for example, because of the risk of losing repeat customers (Leitner & Leopold-Wildburger, 2011; Perera et al., 2019; Tong, Feiler & Larrick, 2018). When there is an asymmetric loss function, where under-forecasting is more expensive than over-forecasting, forecasters are likely to be more inclined to *"err on the high side"* (Goodwin, 1996). However, in general, the shape of the loss function is likely to influence the forecasting process and the accuracy and bias resulting from it. The asymmetry of loss functions is therefore considered a possible source of bias in forecasting (Lawrence & O'Connor, 2005). The direction of this bias appears to be determined by the circumstances of the organization (Sanders & Manrodt, 1994).

As mentioned, a lot of times the loss function of companies is asymmetric. Acting upon this when making adjustments, would suggest a blur of the difference between a forecast and a decision according to Goodwin (1996). By Goodwin (1996), a *forecast* is considered as an estimation of what will happen in the future and a *decision* is considered as an estimate that should be acted upon in order to minimize loss. In this case, a *forecast* could be that 500 units might be sold next week. And if, for example, the costs for a stock out are higher than the costs for overstocking, a *decision* could be to have a stock of 550 units. However, since often there is acted upon the asymmetric loss function when making a forecast, this means many forecasts are actually decisions. Since humans are limited in their capacity for information processing and the fact that, arguably, making a decision is more complex than making a forecast, it is probable that some heuristic principles will be utilized. As a result, the forecast (or decision) may be biased and therefore not optimal. This confusion between forecasts and decisions has been suggested as one of the reasons why upward adjustments were so poorly applied by forecasters in the study of Fildes et al. (2009). It emerged that, for example, a forecast of 200 units might be increased to 250 units to lower the risk of a stock out. The problem with this situation was that the decisions, what they actually were, but labeled as 'forecasts', were subject to misinterpretation and that they could be adjusted even further upwards by others.

In the context of this research, for all operational domains, over-forecasting leads to too many employees being scheduled to work. In this case, the salary of the employees who were not actually needed can be regarded as a loss. In terms of under-forecasting, there is a difference between the domains. For the physical stores, under-forecasting the number of visitors leads to too few employees being scheduled with the risk that not all customers can be assisted where necessary. The possible consequence of this is lost sales and possibly even the risk of losing repeat customers. If due to under-forecasting, too few employees are scheduled for the warehouse and the delivery service, this can lead to the need to work overtime and delays in the delivery of orders, possibly with the consequence that customer satisfaction decreases. Finally, for the customer service, under-forecasting of the number of customer contacts, and therefore too few employees being scheduled, can result in a longer waiting time or even unanswered questions, with again the possible consequence of a decrease in customer satisfaction.

Given the main goal of Coolblue: 'making customers happy', it is expected that declining customer satisfaction or even losing repeat customers due to under-forecasting the workload, is experienced as more costly than scheduling too many employees. Assuming an asymmetric loss function where under-forecasting is more costly than over-forecasting, in combination with an optimism bias, for all operationally oriented forecasts, it is expected that analysts are more likely to adjust a forecast upwards than to adjust a forecast downwards. Therefore, the following hypothesis can be formulated:

> ***Hypothesis 2b:*** *Analysts are in general more likely to adjust a forecast in an upward direction as opposed to making an adjustment in a downward direction.*

As mentioned earlier, downward adjustments are in many cases more effective than upward adjustments. One possible explanation for the fact that downward adjustments are generally more effective, is that adjustments downward are only applied if there are indications that a downturn may arise (Baecke et al., 2017). Accordingly, Fildes et al. (2009) found that the information on which adjustments are based is more reliable in the case of negative adjustments compared with positive adjustments. And, typically, the bias for upward adjustments is higher than for downward adjustments, for example, because they are more subject to over-optimism. Furthermore, the asymmetric loss function that can affect forecasters can not only cause adjustments to be upwards relatively more often but can also cause excessively and unnecessarily upward adjustments, resulting in a worse performance of upward adjustments compared to downward adjustments (Syntetos et al., 2016). Something else worth mentioning is that upward adjustments can, in theory, generate an infinite error since these forecasts are not constrained in any way. Downward adjustments, on the other hand, are bounded by zero since negative forecasts do not exist. The error of downward adjustments is therefore also bounded. This could contribute to the fact that upward adjustments are more likely to have a larger error.

In conclusion, based on the findings from previous studies, and given that upward adjustments are more biased by over-optimism, in combination with an assumption of an asymmetric loss function with the possible confusion between forecasts and decisions, it is expected that the error after upward adjustments will be higher than that of downward adjustments in the context of this research. Therefore, the following hypothesis can be formulated:

> ***Hypothesis 2c:*** *Downward judgmental adjustments are in general more likely to decrease the forecasting error of operationally oriented forecasts than upward judgmental adjustments.*

2.3.4 Size of Judgmental Adjustments

After deciding on the direction of an adjustment, the next step of the judgmental adjustment process is to decide on the size of the adjustment (Arvan et al., 2019; Lawrence et al., 2006). As with the direction of an adjustment, it is suggested that there is a link between the size of an adjustment and the accuracy of forecasts. Typically, large adjustments are more effective than small adjustments for improving the accuracy of forecasts (Baecke et al., 2017; Van den Broeke et al., 2019). One example is the study of Diamantopoulos and Mathews (1989). In this study, a strong positive relationship was found between the size of an adjustment and the improvement in accuracy, i.e. the larger an adjustment, the larger the improvement in forecast accuracy. Another study in which mainly the relatively large adjustments contributed to the improvement in forecast accuracy, and in which the relatively smaller adjustments generally were not effective in improving the forecast accuracy, and often even deteriorated the accuracy, is that of Fildes et al. (2009). In addition, Syntetos et al. (2009) reported that large downward adjustments were especially effective in their context of intermittent demand. This was also seen in the study of Fildes et al. (2009), where the large downward adjustments brought about a significant improvement in accuracy for the manufacturers, however, this did not apply to the retailers in their study. For them, large negative adjustments generally showed a slight decline in accuracy. Finally, taking into account different time horizons, Van den Broeke et al. (2019) found that in general as the time horizon decreases, adjustments usually get larger. This may be due to the fact that more information becomes available as the time horizon decreases. The relatively smaller adjustments in a more faraway horizon might be due to the 'tinkering with the data' effect as proposed by Fildes et al. (2009).

There are several possible reasons for the fact that in general the relatively large adjustments are more effective than the small adjustments in terms of improving the accuracy of forecasts. First of all, large adjustments are most typically linked to reliable information about occurrences that will have large expected consequences that are not represented in the system forecast. Based on this, such adjustments are expected to be accompanied by significant accuracy improvements (Fildes et al., 2009). Furthermore, if forecasters make an adjustment based on information they consider to be unreliable, they are likely to reduce the size of the adjustment to hedge against the possible consequences of such adjustments. This is one of the reasons why small adjustments tend to be less effective (Fildes et al., 2009).

In Section 2.3.1, several motivations for adjustments were discussed. As mentioned, some of these motivations lead to unnecessary, typically small adjustments being made. An example is the 'illusion of control effect' causing small adjustments being made to the system-generated forecast without really having a good reason, which may result in a decrease in accuracy. Other examples showing that there is a tendency for small adjustments to be made for no good reason are the 'tinkering with the data' effect and the perception of noise in the data as a pattern. Because small adjustments are regularly made for no apparent reason, they frequently result in a decline in accuracy (Fildes et al., 2009).

In conclusion, since relatively large adjustments are generally based on reliable information and smaller adjustments are in general more likely to be based on unreliable information or made for no good reason, large adjustments are expected to be more likely to improve forecast accuracy than small adjustments. Therefore, the following hypothesis can be formulated:

> ***Hypothesis 3:*** *Large judgmental adjustments are generally more likely to decrease the forecasting error of operationally oriented forecasts than small judgmental adjustments.*

2.3.5 Timing of Judgmental Adjustments

The timing of a judgmental adjustment, in addition to the direction and the size of an adjustment, may be an important factor influencing forecast accuracy (Alvarado-Valencia et al., 2017; Van den Broeke et al., 2019). According to Nahmias and Olsen (2015), there is a negative relationship between the length of the forecast horizon and forecast accuracy, i.e. when the length of the forecast horizon increases, the accuracy of forecasts will decrease. Accordingly, Petropoulos et al. (2014) states that the forecast accuracy decreases as the forecast horizon becomes longer for regular/fast-moving data which is not intermittent. However, although this could be the case for statistical forecasting, not much research has been done into the influence of timing of judgmental adjustments on forecasting accuracy. In terms of forecast accuracy over time, it would be expected that the effectiveness of judgmental adjustments increases as more recent contextual information becomes available closer to the date of the sale. This is also important because accurate forecasting is critical close to the date of the sale since generally more operational commitments are made at that time, such as personnel capacity planning (Van den Broeke et al., 2019). Nevertheless, the few studies done on the timing of judgmental adjustments and their influence on accuracy suggest otherwise. This could be due to the fact that there is a possibility that over-reaction to new information, which is seen as a bias in judgmental adjustments, may hinder the benefit of new information becoming available over time, as suggested by Lawrence et al. (2000).

In the study of Franses and Legerstee (2011b), they investigated the impact of the forecast horizon on the quality of judgmentally adjusted forecasts. They utilized a large database for their analysis consisting of sales forecasts for pharmaceutical products on SKU-level. Although their results showed that in general adjusted forecasts were less accurate than the system forecasts in all horizons, the accuracy deterioration was the least in the most faraway horizon. And when forecasters did contribute significantly to the accuracy of the forecasts by judgmentally adjusting, this contribution became greater as the horizon get further away. Also striking and worth mentioning is that around the horizon which is most important for the company, in the case of this study 6 months ahead, the forecasts are the worst. It seems that forecasters exert too much impact on forecasts in horizons that are important to them and that the effectiveness of the adjustment decreases as a result (Franses & Legerstee, 2011b).

Another study considering different time horizons in investigating the impact of judgmental adjustments is that of Van den Broeke et al. (2019). They investigated not just the consequences of adjustments over time for the accuracy of forecasts, but also the change in characteristics of consecutive adjustments over different time horizons. They utilized a dataset with 307,232 forecasts over different time horizons from four different companies and expected that recent contextual information becoming available near the date of the sale would lead to an improvement in forecast accuracy. Their results partly confirm this but show that it depends, among other things, on whether the statistical forecast is updated or not. For two of the four companies in this study, their statistical forecasts are not updated over time. Forecast accuracy improvement tends to increase as the time horizon decreases for these companies. It was seen that for these companies the statistical forecasts were outperformed by the adjusted forecasts mainly in the time horizons closer to the date of the sale. In general, adjustments made further away from the date of the sale appear to have little effect or even may harm accuracy, contrary to the findings of Franses and Legerstee (2011b).

However, there are different findings for the other two companies in this study, which did update their statistical forecasts. Namely, closer to the date of the sale, the forecasting accuracy improvement decreased for these companies. So, in general, adjustments made closer to the date of the sale do not always seem to increase the average accuracy of forecasts, particularly not when updates are made to the statistical forecast over time. Van den Broeke et al. (2019) gives the same possible explanation as that of Franses and Legerstee (2011b), that is that it is possible that the forecasters are relying too heavily on their own judgment in comparison to the formal model. In addition, a possible explanation is given for the fact that mainly for the companies where the statistical forecast is updated, the adjustments close to the date of the sale, in general, do not improve forecast accuracy. That is, if a statistical forecast has been updated, it may already reflect the most up-to-date information, making the effort of judgmental adjustment unjustifiable (Van den Broeke et al., 2019).

In conclusion, closer to the date for which a forecast is made, more contextual information generally becomes available based on which an adjustment can be made. However, depending on how this information is processed, it can affect the accuracy of the forecasts. From existing literature, it appears that adjustments close to the forecast date are possibly subject to over-reaction to new information and forecasters relying too heavily on their own judgment in comparison to the model. Therefore, it is expected that the forecast error of adjusted forecasts increases with a decreasing time horizon. Based on this, the following hypothesis is formulated:

> ***Hypothesis 4a:*** *The forecast error of adjusted forecasts increases with a decreasing time horizon.*

In addition, given the fact that in the context of this study the statistical forecast, generated by the algorithm, is updated every day, and that it therefore may already reflect the most up-to-date information, it is expected that the added value that analysts can deliver to the forecast decreases over time. It is therefore expected that the forecast accuracy improvement relative to the algorithm-generated forecast declines with a decreasing time horizon. Based on this, the following hypothesis is formulated:

> ***Hypothesis 4b:*** *Given that the algorithm-generated forecast is updated over time, the forecast accuracy improvement relative to the algorithm-generated forecast decreases with a decreasing time horizon.*

3. Forecasting Process

In this chapter, the procedure for adjusting forecasts is described in Section 3.1. In addition, in Section 3.2, a numerical example of the forecast calculations is given.

3.1 Adjustment Procedure

Within the forecasting procedure for the operational forecasts of Coolblue, first, a statistical forecast is generated by the machine learning algorithm of the data science team. This statistical forecast will be referred to as the *Raw Forecast (RF)* from now on. There is a dashboard that shows the data analysts what features are included in this raw forecast to give some insight into how this forecast is constructed. The data analysts of team forecasting can make adjustments to the raw forecast based on contextual knowledge they possess that has not been processed in the raw forecast. This then results in an *Adjusted Forecast (AF)*, which is communicated to the stakeholders, and serves as input for their planning process.

Analysts can make adjustments by entering a percentage or absolute adjustment or by entering an override in a Google Sheet. A fictional example of this adjustment sheet is shown in Table 3.1. It is recorded which *key* it concerns, which indicates for which specific forecast this adjustment is entered. For example, the key 3_146 indicates that this forecast is for Belgium (3) Parcel (146). In addition, a start date is defined, which indicates the first date for which the adjustment applies, and an end date can be entered when an adjustment applies for several days. Furthermore, a justification and a category can be attached to an adjustment to give the reason for the adjustment and to categorize the adjustment. Finally, the justifier, i.e. the analyst who makes the adjustment, is stored in this sheet, and the creation date is specified, i.e. the date on which the adjustment is made.

Table 3.1: Fictional Example of the Adjustment Sheet

key	**start_date**	**end_date**	**percentage**	**absolute**	**override**	**justification**	**category**	**justifier**	**creation_date**
1_146	28-09-2021	05-10-2021	-5%			Long term assumptions	adjust level	jan.janssen	10-02-2021
1_159	02-07-2022			+175		Warm weather	other	jan.janssen	25-06-2022
3_146	27-08-2022				334	Back to school	campaign	jan.janssen	22-08-2022

Every morning around 5 A.M. the forecasting infrastructure, called *BluePearl*, is run and a dump file of this Google Sheet is written to BigQuery. This is a data warehouse where all the data is stored. Besides that the BluePearl is automatically run at 5 A.M., it can also be run manually at any other time. In Figure 3.1, it is shown how adjustments are processed. After the BluePearl run, there is the *raw forecast* (RF) of that day. The *total adjustment* is a combination of all adjustments that were made previously for a certain key and date. The *Adjusted Forecast Morning (AFM)* is the forecast that is available every morning, which consists of the *raw forecast* of that day, with all previously made adjustments (*total adjustment*) processed in it. If an adjustment is the first one for a particular key and date, there are no previous adjustments and therefore no total adjustment. In that case, the AFM is equal to the RF. If analysts decide that the AFM is not at the right level and that an (additional) *adjustment* is required, they enter the adjustment into the adjustment sheet, which results in the *adjusted forecast end of day* (AF).

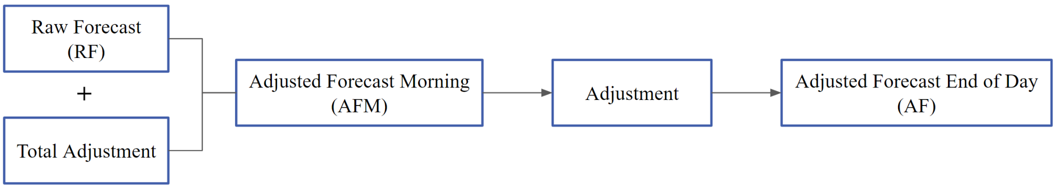

Figure 3.1: Forecasting Process

3.2 Example Forecast Calculation

To clarify how the adjusted forecast end of day is calculated after adjustment(s), Table 3.2 provides a numerical example. The numbers in this example are fictional and the adjustments are aimed at a specific key and date. The very first adjustment for this forecast was made on 10-02-2021, referred to as the *creation date*. On that day the *raw forecast* (RF) was 100. Because the adjustment on 10-02-2021 is the very first adjustment, there are no previous adjustments, and therefore the AFM is equal to the RF, which is 100. If a percentage adjustment is made, the percentage is always expressed over the raw forecast (RF) and added to the AFM. In this case, the *adjustment* was + 50% and therefore, the AF at the end of the day became 150.

The row in gray is not an adjustment but is added to simply show how the process works with the daily updated raw forecast. On 11-02-2021, the day after the first adjustment, the raw forecast is updated and changed from 100 to 104. The total adjustment, a combination of all previously made adjustments, is + 50%. The AFM on 11-02-2021 is now not '100 + 50%' but '104 + 50%' leading to an AFM of 156. Since no adjustment was made on 11-02-2021, the AF end of day is also 156. So, despite no adjustments being made, the AF end of day has changed compared to the previous day. This shows that there can be two reasons for a change in the AF end of day forecast that is in the system and which is ultimately communicated to stakeholders. This forecast can change due to a change in the raw forecast, which is updated daily, or it can change due to an active adjustment of an analyst.

Table 3.2: Numeric Example Forecasting Process

Creation Date	Raw Forecast	Total Adjustment	AF Morning	Adjustment	Calculation	AF End of Day
10-02-2021	100	-	100	+ 50%	100 + 0.5 * 100	150
11-02-2021	104	+ 50%	156	-	-	156
15-02-2021	110	+ 50%	165	+ 20	165 + 20	185
22-02-2021	110	+ 50%, + 20	185	157	157	157

On 15-02-2021, the raw forecast is 110, given the total adjustment at that time (+ 50%), the AFM is 165. If absolute adjustments are made, these are directly processed in the AFM. In the example, the + 20 will be directly added to the 165 (AFM), which leads to an AF end of day of 185. On 22-02-2021 the raw forecast is still 110. The total adjustment is now '+ 50%, + 20' which leads to an AFM of 185. On that day, an override of 157 is entered. This is from then on a hard coded adjusted forecast (AF) and stays the adjusted forecast without any dependency on the raw forecast or previous adjustments.

3.2.1 Updated Raw Forecast

Since the raw forecast is generated every day, it could be that there is instability in this forecast, i.e. that the forecast varies from day to day. This is not desirable, especially not in the last three weeks. By then the stakeholders already started to make a planning. If there is a lot of instability in the raw forecast in these weeks, the stakeholders see a different forecast every time and they should continuously upscale or downscale capacity. To prevent this possible instability in the last three weeks, the raw forecast is not updated anymore by then, which means that the raw forecast remains the same in the last three weeks.

This means that if a percentage or absolute adjustment is made in the last three weeks, the resulting adjusted forecast (AF) is no longer subject to changes in the raw forecast, and that this is the forecast that is communicated to stakeholders. However, if such an adjustment is made before the last three weeks, the resulting adjusted forecast is still subject to the changes that take place in the raw forecast.

4. Methodology

In this chapter, the methods used for the research are discussed. The main methods are hypothesis testing and simulation, for Research Questions 1 and 2, respectively. To test the hypotheses, first, data is collected. The data extraction method as well as the required data transformations are discussed in Section 4.1. The initial data set and variables are described in Section 4.2. Furthermore, the chosen error measure is explained in Section 4.3, and the data cleaning and preparation method is described in Section 4.4. Finally, in Section 4.5, the approach for testing hypotheses, the exploratory research, and the simulation of new adjustment procedures are discussed.

4.1 Data Extraction

In this section, it is discussed how the data was extracted and which data transformations were required to retrieve the necessary variables from the database. The data is stored in BigQuery, and extracted by writing queries using Structured Query Language (SQL). To keep the queries manageable, the data is extracted separately for every domain and merged later. The data transformations that have been performed are described in Section 4.1.1.

4.1.1 Data Transformations

Creation Date Adjustment
The date on which an adjustment is made is inserted in the Google Sheet by analysts when they are entering an adjustment. However, this column is not included in the connection between Google Sheets and BigQuery, and therefore, the date on which an adjustment is created is not directly available. A workaround is used to obtain the creation date of an adjustment. Every day a dump file of the Google Sheet is written to BigQuery. This dump file is a snapshot of all the data in the Google Sheet at a point in time. The moment a snapshot is taken is recorded in an *insert timestamp* variable. To retrieve the creation date of an adjustment, it is checked when it was the first time a certain adjustment appeared in the snapshot table in BigQuery. Assuming that the BluePearl runs the next morning, the insert timestamp was subtracted by 1 day to get the creation date of an adjustment.

Adjustments Made For Several Days Splitted
Adjustments can be entered in the Google Sheet for a specific day but also for several days. When an adjustment is entered for several days, an end date is also entered in the Google Sheet in addition to the start date. However, for each of the days in such a series of adjustments, the raw forecast will most likely be different, and the target will also be different for each day. To make these adjustments suitable for analysis in this study, all series of adjustments have been split into separate rows, one for each date in the series. In this way, the raw forecast and the target can be matched to the adjustment for a specific day.

Adjustments Made On One Day Merged
According to the analysts' thinking process, all adjustments made on a specific day for the same date and key are considered a final decision for that day, and should also be treated as a single decision. For example, if an analyst makes a series adjustment for a particular week by applying +10% for that entire week, but for a particular day in that week, the analyst wants to make an additional increase or decrease in the forecast, this results in two separate adjustments. First a series adjustment for the entire week, then an adjustment for a single day in that week. To ensure that these two separate adjustments made on the same day are treated as a single decision, they are merged.

Adjusted Forecast End of The Day
The adjusted forecast can be retrieved from BigQuery and is available after the BluePearl run. However, if the BluePearl is run the next morning, the adjustments made will be calculated over the raw forecast (RF) of the next morning. Analysts base their adjustments on the values of the RF and AFM of the day of adjusting. Therefore, in order to see the result of their adjustments, an *adjusted forecast end of day* (AF) is calculated. If an override was entered, this value is simply taken as the adjusted forecast end of day. To retrieve the adjusted forecast at the end of the day (AF) in case there was a percentage or an absolute adjustment, Equation 4.1 is used.

$$\begin{aligned}\text{AdjustedForecastEndofDay}_d = {} & AdjustedForecastMorning_d \\ & + PercentageAdjustment * RawForecast_d \\ & + AbsoluteAdjustment\end{aligned} \tag{4.1}$$

4.1.2 Data Choices

Target / Actual
In the forecasting process of Coolblue, a distinction is made between actuals and targets. The *actual* indicate how much work has been done on a given day. The *target* is adjusted for the workload that was actually there on a given day. For example, on a particular day, 100 products are ordered. However, not all orders can be processed in the warehouse. Due to a shortage of capacity, only 80 orders can be processed. The orders that could not be processed (20 orders) are moved to the next day. The *actual* would in this case be the amount of work that has been processed, i.e. 80 orders. The *target* will indicate the workload that was actually there if all orders were to be processed, i.e. 100 orders. The operational forecasts focus on estimating the real workload for the various domains, hence the targets are the most important in this scenario and will be used in the analysis as the 'actuals'. The targets are for this reason also the input for the raw forecasts generated by the data science team.

4.2 Data Sets

In Table 4.1, the number of complete triplets per domain and the total number of complete triplets is shown. A complete triplet means that the raw forecast, the adjusted forecast, and the target are available. Several constraints have been incorporated in the queries for this data. First, the data runs from 20-05-2020 to 20-05-2022, in terms of days for which a forecast has been made. In this way, there is exactly two years of data. In addition, the forecasting team started to use BluePearl at the beginning of 2020. In this initial period, a lot of testing was done with making adjustments. To prevent adjustments that are made during the implementation phase of this new system to be in the data set, a constraint for the creation date of an adjustment is included such that adjustments that are created before May 1st, 2020 are excluded from the data set. Furthermore, observations for which the creation date of an adjustment is after the date for which the forecast is made are excluded since this would mean that adjustments were made for the past. Also, series adjustments for which the end date is earlier than the start date have been removed from the data set. In Section 4.2.1, for all four domains, some additional information is given about the data set.

Table 4.1: Available data per domain

Domain	**Data Set**	**Complete Triplets**
Customer Service	stg_cus_voice_co_st	14,166
Delivery	stg_cbd_delivery_or_st	12,251
Warehouse	stg_spp_outbound_po_st	34,729
Stores	stg_sto_walk_in_vr_st	55,379
Total		**116,525**

4.2.1 Additional Information Data Sets

Customer Service - Contacts Offered (CO)
For the forecasts in the customer service domain, a distinction is made in country. Furthermore, a distinction is made between voice (calls) and non-voice (emails, social media). Data science only provides a raw forecast for The Netherlands and Belgium, and only for voice contacts. The analysis is therefore focused on adjustments made to forecasts on the voice channels for The Netherlands and Belgium. These voice channels are further subdivided into *advice* and *service* contacts. It is at this level that the adjustments are made in this domain.

Coolblue Delivery - Orders Delivered (OD)
For the forecasts in the delivery domain, a distinction is again made in country. There is no data science forecast for Germany in this domain either. The focus of the analysis is therefore on the adjustments to forecasts for The Netherlands and Belgium. In addition to the country, a distinction is made between two workload types, referred to as *networks*: 1M doorstep and 2M delivery. Here, 1M and 2M stand for 1 man and 2 men, respectively. Adjustments in this domain are made at the country and network level.

Warehouse - Products Ordered (PO)
For the forecasts in the domain warehouse, again only the adjustments to forecasts for The Netherlands and Belgium are considered. Furthermore, a distinction is made between parcel, XL products, white goods, and products to be picked up at a store. It is also at these levels that adjustments are made.

Stores - Visitors Registered (VR)
For the stores domain, the data science team generates a raw forecast for all individual stores. This forecast is made per store for three *streams*: advice, pick-up, and service. Adjustments for store forecasts are made on both stream and total level, i.e. for a certain store an adjustment can be made for just *advice* but can also be made for the total amount of customers expected in that store.

4.2.2 Initial Variables

In Table 4.2, the variables that are relevant for analysis are shown. Some of the variables were extracted directly from the database, others were obtained through the data transformations.

Table 4.2: Variables initial data set

Variable	Description	Type	Example
Domain	Domain to which the data belongs	String	stores
Creation Date Adjustment	Date on which an adjustment is made	Date	2022-04-21
Start Date	Day for which a forecast is made	Date	2022-02-15
Percentage Sum Day	Sum of all percentage adjustments made on the same day, for the same Key and Start Date	Float	-0.2
Absolute Sum Day	Sum of all absolute adjustments made on the same day, for the same Key and Start Date	Float	15.0
Override Final	Latest inserted override on a particular day	Float	1035.52518262412
Raw Forecast	Statistical forecast generated by the algorithm of the data science team	Float	479.46860258008871
Adjusted Forecast	Value at the end of the day after processing the adjustments made on a particular day	Float	586.00932952485516
Target	Real workload for Products Ordered, Orders Delivered, Visitors Registered or Contacts Offered	Float	431.0

4.3 Error Measure

It is crucial to choose an appropriate error measure for evaluating the accuracy of judgmental adjustments (Davydenko & Fildes, 2013; Fildes & Goodwin, 2007). Many error measures are available, however, each method of measuring the error has its advantages and disadvantages (Goodwin & Lawton, 1999; Perera et al., 2019). One of the most widely used error measures is the Mean Absolute Error (MAE). For the MAE, first, the forecast error (e_i) is calculated by taking the difference between the actual value (A_i) and the forecast (F_i), and subsequently, the average of all the absolute forecast errors is taken (Hyndman & Athanasopoulos, 2018). The MAE is given by the following formula:

$$\text{MAE} = \frac{1}{n} \times \sum_{i=1}^{n} |e_i| \tag{4.2}$$

Where,

$$e_i = A_i - F_i \tag{4.3}$$

Advantages of the MAE are that calculation and interpretation are relatively simple. However, the MAE is a scale-dependent measure and is therefore only useful when all data sets that are used for the analysis of the performance of different forecasting methods, are on the same scale (Shcherbakov et al., 2013; Hyndman & Athanasopoulos, 2018). In this research, the MAE is not appropriate since the error will be compared across data sets from several domains with different scales. An error measure that is scale-independent, is the Mean Absolute Percentage Error (MAPE). Among the scale-independent measures, the MAPE is one of the most utilized in both academic research and in industry (Hyndman & Athanasopoulos, 2018; Perera et al., 2019). In addition to the fact that this measure is most frequently used, it is also an error measure that is intuitive and easy to explain and interpret (Sanders & Manrodt, 2003). And, perhaps more importantly, the MAPE is the currently used error measure at Coolblue. The MAPE is given by the following formula:

$$\text{MAPE} = \frac{1}{n} \times \sum_{i=1}^{n} \frac{\mid \text{e}_i \mid}{\text{A}_i} \times 100 \tag{4.4}$$

Of course, this error measure has drawbacks as well. One of them is that the results obtained with this error measure can be significantly influenced by outliers. In addition to an outlier analysis, this study will therefore use the Median Absolute Percentage Error (MdAPE) because this error measure is more robust in terms of outliers (Shcherbakov et al., 2013). The MdAPE is calculated by taking the median of the Absolute Percentage Error (APE) and is given by the following formula:

$$\text{MdAPE} = \underset{i=1,n}{\text{median}} \left(\frac{\mid \text{e}_i \mid}{\text{A}_i} \times 100 \right) \tag{4.5}$$

A drawback of percentage based error measures, such as the MdAPE, is that this measure is undefined if the actuals are zero ($A_i = 0$) (Hyndman & Koehler, 2006). In this case, division by zero would take place, which is mathematically impossible. However, in the context of this research, where forecasts are made on an aggregated level, the chances of having actuals equal to zero are minimal. The occurrences where actuals are zero, which indicates that the workload was zero, are most likely holidays. Most probably no adjustments are needed for these days because the raw forecast should already take these holidays into account.

In addition to the fact that actuals of zero can be a problem for this error measure, close to zero actual values are also an issue. When the actual values are close to zero, this results in extremely high error percentages (Davydenko & Fildes, 2013). However, again, since this study is focusing at forecasts on an aggregated level, close to zero values are hardly expected. Another property of percentage based error measures that is worth mentioning is the asymmetry of this type of error measures. Negative errors ($F_i > A_i$) are penalized more severely than positive ones ($F_i < A_i$) (Hyndman & Athanasopoulos, 2018; Shcherbakov et al., 2013). This is due to the fact that for too low forecasts, the percentage error cannot surpass 100%. For forecasts that are excessively high, on the other hand, there is no limit for the percentage error. This should be taken into account when interpreting the results in this study.

4.4 Data Cleaning and Preparation

For data cleaning and preparation of the data set, all observations for which the target, the raw forecast (RF), the adjusted forecast (AF) and/or the adjusted forecast morning (AFM) were less than zero were removed from the data set. This was done because no negative values are possible for the forecast and the target, and they are therefore considered incorrect. Furthermore, when for an observation the adjusted forecast morning (AFM) was equal to the adjusted forecast end of day (AF), it was also removed from the data set. This means that in the end no adjustment was made that day. This is caused by several adjustments being made on a day that cancel each other out or by adjustments of 0% by means of applying a shortcut of copying series of adjustments by analysts.

4.4.1 Added Variables

Additional variables have been created in order to be able to analyze the data and to obtain the desired insights. Firstly, variables have been created with the absolute percentage errors of the forecasts, so that the performance of (adjusted) forecasts can be expressed. In addition, many dummy variables were created, for example to indicate the adjustment type, to indicate whether an adjustment was upward or downward, and whether or not an adjustment was made in the right direction. Furthermore, timing variables have been created to see how many days, weeks or months before the forecast date an adjustment has been made.

By ranking over and partitioning by the key and the forecast date, and order by the creation date, a variable was created that indicates the number of adjustment for a particular key and date. Moreover, a variable was created to indicate whether an adjustment was Covid-related or not. A piece of query has been written in which all indications for Covid-related adjustments have been filtered based on category and/or justification. These indications for classifying an adjustment as Covid-related have been established after meetings with all analysts. Examples of indications are; ‘corona’, ‘thuiswerken’, ‘lockdown’, and ‘measures’. Finally, a variable has been created for the relative size of an adjustment. For this purpose, all adjustments were first converted to a percentage change. The median size of adjustments was then calculated per domain, see Table 4.3. Subsequently, Equation 4.6 was used to determine the relative size of an adjustment per domain, i.e. if an adjustment was relatively small (0) or relatively large (1).

Table 4.3: Median Size of Adjustments per Data Set

Data Set	Median Size of Adjustments
Total	10.83%
CO	12.19%
OD	8.82%
PO	9.92%
VR	11.76%

$$\text{Relative Size } = \begin{cases} \text{IF adjustment size} \leq \text{median size THEN 0} \\ \text{IF adjustment size} > \text{median size THEN 1} \end{cases} \tag{4.6}$$

4.4.2 Outlier Analysis

An outlier is a data point that significantly differs from the other data points in a dataset (Hair, Black, Babin & Anderson, 2014). The statistical power can be reduced by increased variability of the data if outliers remain in the dataset. In this way, they have a disproportionate impact on statistical analysis. Therefore, an outlier analysis is performed to remove them from the dataset. The outliers are based on the raw forecast error and the adjusted forecast error. Since the data is not normally distributed, and since there are some very extreme outliers, the Interquartile Range (IQR) is used to detect and remove outliers. This method does not require a particular distribution of the data and is dependent on the median instead of the mean of the data, and therefore more robust against outlier influence with regard to the calculations. For the IQR, the first (Q1) and the third (Q3) quartile are calculated after which the first quartile is subtracted from the third quartile:

$$\begin{aligned} Q_1 &= (n+1) * \frac{1}{4} \\ Q_3 &= (n+1) * \frac{3}{4} \\ IQR &= Q_3 - Q_1 \end{aligned} \tag{4.7}$$

Subsequently, the following formula is used for detecting outliers (Montgomery & Runger, 2018):

$$\text{Outlier } = \begin{cases} X < Q1 - 1.5^{*}IQR \\ OR \\ X > Q3 + 1.5^{*}IQR \end{cases} \tag{4.8}$$

If for a specific observation both the raw forecast error and the adjusted forecast error were detected as an outlier, this observation was removed from the data set. If only one of the errors was detected as an outlier in a particular observation, this observation remained in the data set. After data cleaning and removing outliers, the total data set consists of 111,852 triplets.

4.4.3 Train set and Test set

The total data set is divided into a train data set and a test data set. This has been done so that new forecasting procedures can be trained on the training data set and the performance of these models can be evaluated on the test data set. To ensure the same time periods are analyzed for each domain, a time-based split of 80%/20% is created based on the forecast dates. That is, the first year and a half, from 20-05-2020 to 20-12-2021, is used for training, and the last half year, from 21-12-2021 to 20-05-2022, is used for testing a new adjustment procedure. In Table 4.4, the exact amount of data for the train and test data set can be found per domain.

Table 4.4: Amount of Train and Test Data per Domain

Domain	Training Data	Test Data
CO	9,873 (71.21%)	3,992 (28.79%)
OD	6,569 (55.02%)	5,371 (44.98%)
PO	25,155 (74.88%)	8,440 (25.12%)
VR	33,968 (64.76%)	18,484 (35.24%)

4.5 Data Analysis

4.5.1 Hypothesis Testing

The hypotheses generated in Chapter 2 are tested using the training data set. In addition, exploratory research is conducted, see Section 4.5.2. The program RStudio version 4.2.1 is used to perform the statistical tests. QQ-plots showed that the data is not normally distributed. Therefore, non-parametric tests are used. For non-parametric tests, there are no requirements regarding the distribution of the data (Montgomery & Runger, 2018). In Table 4.5, it is shown, per hypothesis, which statistical tests are used. In Appendix A, a description of these statistical tests is given.

Table 4.5: Statistical Tests

	H1	H2a	H2b	H2c	H3	H4a	H4b
Statistical Tests							
Wilcoxon Signed-Rank test	x			x	x		
Kruskal-Wallis test						x	x
Pairwise Wilcox test						x	x
Chi-square goodness-of-fit test	x	x	x		x		
2-Sample test for equality of proportions		x	x	x	x		
Regression		x	x		x	x	x
Effect Size							
Wilcoxon effect size r	x			x	x		
Cramer's v	x	x	x		x		

4.5.2 Exploratory Research

Besides the main topics that are investigated by means of hypothesis testing, exploratory research is done. This exploratory part is focused on topics for which little or no literature exists. Meetings with three of the analysts were arranged to get input on interesting topics for the exploratory part of the research. Based on these meetings, it is determined which topics could provide valuable insights and which therefore fall within the scope of the research. The exploratory topics are described below.

Covid and Non-Covid Related Adjustments
To investigate whether there is a difference in the characteristics and performance of adjustments that were directly linked to a period with disruptions, in this case the COVID-19 pandemic, and the adjustments that were not directly linked to disruptions, a distinction has been made between these two types of adjustments. Insights are obtained on this topic, however, conclusions regarding the hypotheses were drawn based on Non-Covid data to be able to give recommendations for a normal period, i.e. a period without disruptions.

Type of Adjustment - Absolute, Percentage, Override
The type of adjustment was also examined on an exploratory basis. The type of adjustment refers to whether it was an absolute adjustment, a percentage adjustment or an override. To date, there is no literature on the way of adjusting and its effects on forecasting accuracy. In this study it has been investigated whether there is a difference in the characteristics and performance of these different types of adjustments.

Effect of Subsequent Adjustments
The characteristics and performance of subsequent adjustments are an exploratory topic since minimal research has been done on this. To investigate the effects of subsequent adjustments in terms of error increase or decrease, the isolated effect must be checked. If, for example, an adjustment is the third one for a particular key and date, see the +10% adjustment in Figure 4.1, and the performance of the AF is compared with that of the RF, this does not necessarily say anything about this specific adjustment in isolation, but more about what the combination of those three adjustments has caused in terms of error increase or decrease. In order to be able to investigate the isolated effects of subsequent adjustments, the performance of the adjusted forecast (AF) is therefore compared with the adjusted forecast morning (AFM).

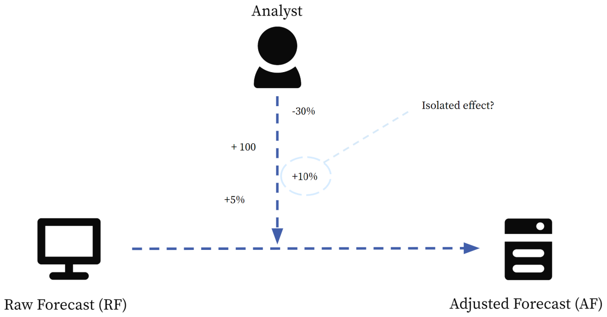

Figure 4.1: Isolated Effect of Adjustments

4.5.3 Improve Adjustment Procedure

Based on the insights obtained from the first research question, new forecasting procedures are proposed. To examine the performance of these new procedures, they are tested by means of simulation on the test data set. The program RStudio version 4.2.1 was used again for these simulations. The statistical test used to see if any improvements are significant is the Wilcoxon signed-rank test for paired samples. In addition, the associated Wilcoxon r effect size was used to check how large the effect size was.

5. Results

To test the previously formulated hypotheses, statistical tests have been performed. This chapter describes the results that have emerged from the statistical tests. The outcome of each hypothesis will be given along with additional results from exploratory topics. As mentioned in Section 4.5.1, the training data is used to test the hypotheses. For each hypothesis, an overview of the main result will be given for the total training data set (N = 75,565), and for the subdivision of the total training data set in Covid data (N = 6,047) and Non-Covid data (N = 69,518). The hypotheses are accepted or rejected based on the Non-Covid data, because simulation models are built and final recommendations are given for a normal period.

5.1 Descriptive Statistics

In Table 5.1, the average number of adjustments that is made for a particular date and key is given for the total data set and per domain. It should be noted that this is the average number of adjustments for those dates and keys for which at least one adjustment was made. The maximum number of adjustments made for a given date and key is 44.

Table 5.1: Average Number of Adjustments per Key and Date

Data Set	Avg. No. of Adjustments
Total	4.66
CO	5.82
OD	6.72
PO	7.84
VR	3.46

In addition, in Table 5.2, the the number of adjustments per adjustment type is given. By merging the adjustments made on one day for the same date and key, adjustments have been created that are a combination of a percentage adjustment and an absolute adjustment, this is referred to as *combination* adjustments.

Table 5.2: Number of Adjustments per Adjustment Type

Adjustment Type	Amount of adjustments (%) *Total*	Amount of adjustments (%) *Non-Covid*	Amount of adjustments (%) *Covid*
Percentage	62,520 (82.74%)	57,098 (82.13%)	5,422 (89.66%)
Absolute	5,725 (7.58%)	5,481 (7.88%)	244 (4.04%)
Override	5,873 (7.77%)	5,734 (8.25%)	139 (2.3%)
Combination	1,447 (1.91%)	1,205 (1.73%)	242 (4%)

5.2 Accuracy of adjusted forecasts

Hypothesis 1: *Overall, judgmental adjustments increase the quality of an operationally oriented forecast. That means, the error of judgmentally adjusted forecasts is, on average, lower than that of algorithm-generated forecasts.* ✔

For hypothesis 1, it was examined whether judgmental adjustments increase the accuracy of an operationally oriented forecast. To check whether an adjustment resulted in an improvement of the forecast, it was tested for all adjustments, whether after an adjustment the adjusted forecast (AF) had a lower error than the raw forecast (RF). In addition, it has been investigated on an exploratory basis whether subsequent judgmental adjustments, i.e. not the first adjustment for a certain date and key, improved the operationally oriented forecast compared to the adjusted forecast morning (AFM). This AFM is available at the beginning of the day and includes all previously made adjustments, calculated over the raw forecast of that day, see Chapter 3.

For all adjustments in the data set, the Absolute Percentage Error (APE) of the RF and the AF was calculated. Subsequently, the median was taken. In Figure 5.1, the MdAPE of the RF and the AF are shown per data set. The Wilcoxon Signed-Rank test for paired samples was applied to see whether the difference in medians was significant. As can be seen, both for the Covid data and the Non-Covid data, the MdAPE of the AF is lower than the MdAPE of the RF. This indicates that, generally, after an adjustment, the forecast error of the AF was lower than that of the RF. In Table 5.3, it can be seen that the difference in MdAPE is statistically significant in both cases. This means that, in general, judgemental adjustments increase the forecast accuracy, and therefore, based on the Non-Covid data, hypothesis 1 is accepted.

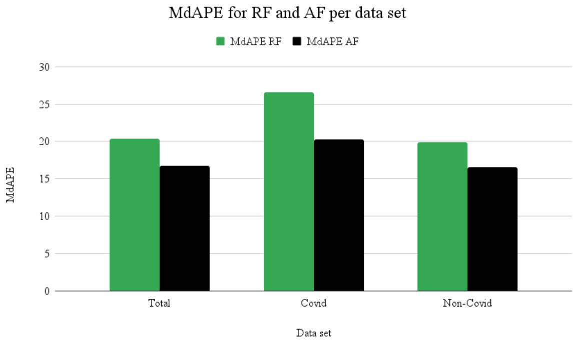

Figure 5.1: MdAPE RF and AF per data set

Table 5.3: MdAPE RF vs. AF per data set

	Total	Covid	Non-Covid
MdAPE RF	20.26	26.47	19.79
MdAPE AF	16.68	20.19	16.43
Test Statistic	V = 1,614,825,705	V = 11,031,323	V = 1,357,661,007
p-value	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.14	0.24	0.13

Covid vs. Non-Covid

It is striking that the MdAPE of the RF for the Covid data (26.47) is significantly higher than for the Non-Covid data (19.79), see Table 5.4. In addition, the MdAPE of the AF for the Covid data (20.19) is also significantly higher than for the Non-Covid data (16.43). This means that, in general, the performance of both the raw forecast (RF) and the adjusted forecast (AF) was worse for the Covid data compared with the Non-Covid data. However, the median difference in %p (*percentage point*) between the APE of the RF and the AF, relative to the RF, is significantly higher for the Covid data (-6.21%p) than for the Non-Covid data (-2.57%p), W = 189,316,742, p-value < 0.001***, $r = 0.16$. This implies that, in general, the judgmental adjustments made by analysts had a greater positive impact for the Covid data than for Non-Covid data.

Table 5.4: MdAPE RF and AF Covid vs. Non-Covid data

	RF	AF
MdAPE Covid	26.47	20.19
MdAPE Non-Covid	19.79	16.43
Test Statistic	W = 248,923,553	W = 238,711,501
p-value	<0.001***	<0.001***
Wilcoxon r	0.31	0.23

Subsequent Adjustments

In Figure 5.2, the MdAPE of the AFM and the MdAPE of the newly adjusted forecast (AF) are shown per data set. This is only done for adjustments where the number of the adjustment is greater than 1, because only then there are previously made adjustments. The sample sizes for adjustments with No. > 1 are as follows: Total N = 58,874; Covid N = 3,903; Non-Covid N = 54,971. For the Covid data, there is no significant difference in MdAPE, see Table 5.5. For the Non-Covid data, the MdAPE of the AF is higher than the MdAPE of the AFM. This would indicate that, in general, subsequent adjustments increase the forecast error compared to the previous adjusted forecast. However, the Wilcoxon effect size r is below the threshold of 0.1, and therefore this effect can be considered negligible. The remainder of the analysis for this hypothesis focuses on Non-Covid data.

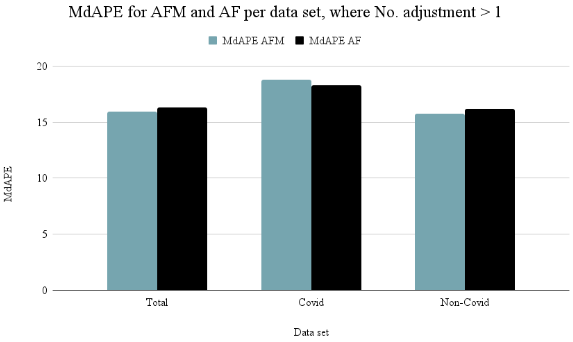

Figure 5.2: MdAPE AFM and AF per data set

Table 5.5: MdAPE AFM vs. AF per data set where No. adjustment >1

	Total	Covid	Non-Covid
MdAPE AFM	15.86	18.71	15.66
MdAPE AF	16.23	18.21	16.11
Test Statistic	V = 825,321,720	V = 3,923,027	V = 714,494,140
p-value	<0.001***	0.106	<0.001***
Wilcoxon r	0.04	-	0.05

First Adjustment

It has been shown that generally after an adjustment the forecast error of the AF is lower than that of the RF. To see whether the very first adjustment, made to the pure raw forecast, generally results in an improvement compared to that raw forecast, the MdAPE of the RF is compared with the MdAPE of the AF. It appears that after the first adjustment that was made for a particular date and key, the AF MdAPE is 17.7 and the RF MdAPE is 21.39, V = 61,283,927, p-value < 0.001***, $r = 0.15$. That means, in general, the AF is outperforming the RF after the first adjustment. The median difference in %p between the APE of the RF and the AF for first adjustments, relative to the RF, is -3.49%p. In terms of frequency, the first adjustment is outperforming the raw forecast in 56.31% of the cases.

Frequencies

In Table 5.6, it is shown how frequent the AF was better, worse or equal to the RF and the AFM. It can be seen that after the majority of the adjustments, the forecast error of the AF was lower than that of the RF (55.91%), $\chi^2(1, N = 69,518) = 970.54$, p < 0.001***, Cramer's v = 0.12. This contributes to the support for hypothesis 1. In addition, in slightly less than half of the cases (49.2%), subsequent adjustments improved the forecast compared to the AFM.

Table 5.6: Performance of the AF in percentages compared to the RF and the AFM

	RF	AFM
AF better	38,866 (55.91%)	27,044 (49.2%)
AF worse	29,673 (42.68%)	27,916 (50.78%)
AF equal	979 (1.41%)	11 (0.02%)

Number of Adjustment

In Figure 5.3 the MdAPEs of the RF, the AF and the AFM are shown per number of adjustment. The maximum number of adjustments made for a given date and key is 44. However, the sample size for high adjustment numbers is very small and this makes interpretation of the graph difficult. That is why the graph only shows data with a sample size (N) $\geq$ 100. This was the case up to and including adjustment number 19. This covers 98.16% of all adjustments and will now apply to any chart where the MdAPE is plotted by number of adjustment.

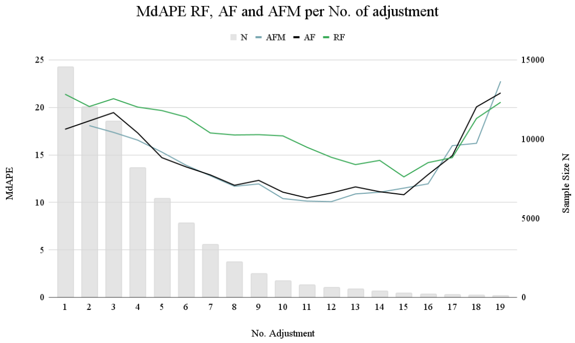

Figure 5.3: MdAPE RF, AF and AFM per No. of adjustment

In Figure 5.3, it can be seen that until the number of adjustment is 17, the MdAPE of the AF is lower than the MdAPE of the RF. The difference in MdAPE between the AFM and the AF is less clear. For adjustments numbers 2, 3 and 4, the MdAPE of the AFM is lower than the MdAPE of the AF, indicating that subsequent adjustments generally did not improve the previous adjusted forecast. After the fourth adjustment, they alternate. An overall downward trend of the MdAPE is visible for both the RF, the AF and the AFM with an increasing number of adjustment. This downward trend applies until approximately 15 adjustments for the RF. For the AF, the increase of the MdAPE starts around 11 adjustments.

Adjustment Types

To find out whether there are differences in performance for different types of adjustments, statistical tests have been performed. In Figure 5.4, the MdAPE of the RF and the AF are shown for the three main adjustment types, i.e. percentage (Pct.), absolute (Abs.) and override. It can be seen that after both percentage and override adjustments, the MdAPE of the AF is lower than the MdAPE of the RF. The differences in MdAPE for these adjustment types are significant, see Table 5.7. After absolute adjustments, the MdAPE of the AF is significantly higher than the MdAPE of the RF. This might imply that absolute adjustments generally do not contribute to improving the forecast accuracy relative to the RF. However, to be able to draw conclusions, the isolated effect of the different types of adjustments must be examined, which will be done in the next paragraph.

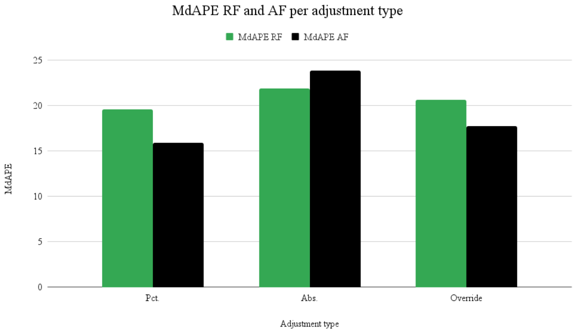

Figure 5.4: MdAPE RF, AF per adjustment type for all adjustments

Table 5.7: MdAPE RF vs. AF per adjustment type

	Pct. Adjustment	Abs. Adjustment	Override
MdAPE RF	19.52	21.81	20.55
MdAPE AF	15.83	23.75	17.67
Test Statistic	V = 942,280,614	V = 5,903,539	V = 9,227,969
p-value	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.17	0.19	0.11

The isolated effect of a particular type of adjustment refers to what this adjustment actually caused in terms of error increase or decrease rather than the final situation after a series of adjustments. To this end, the effect of the very first adjustment in relation to the RF (Figure 5.5) and the effect of subsequent adjustments in relation to the AFM (Figure 5.6) was examined. As can be seen in Table 5.8, the difference in MdAPE between the RF and the AF for first absolute and first override adjustments is not significant. For percentage adjustments, the MdAPE of the AF is significantly lower than the MdAPE of the RF, indicating that first percentage adjustments generally decrease the forecast error relative to the RF. The median difference between the APE of the RF and the AF for first percentage adjustments, relative to the RF, is -4.6%p. In terms of how frequent the AF APE was lower than the RF APE for first adjustments of a particular adjustment type, this was the case for percentage adjustments after 57.63% of the adjustments, for absolute adjustments after 49.06%, and for override adjustments after 51.73%.

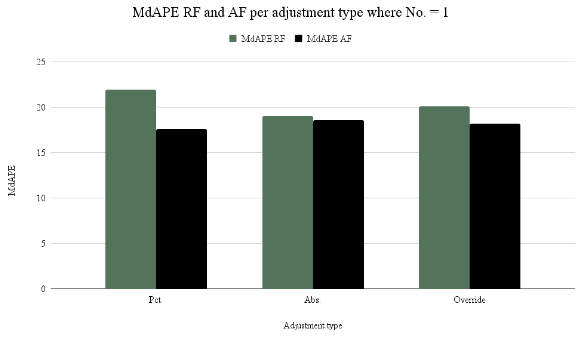

Figure 5.5: MdAPE RF, AF per adjustment type

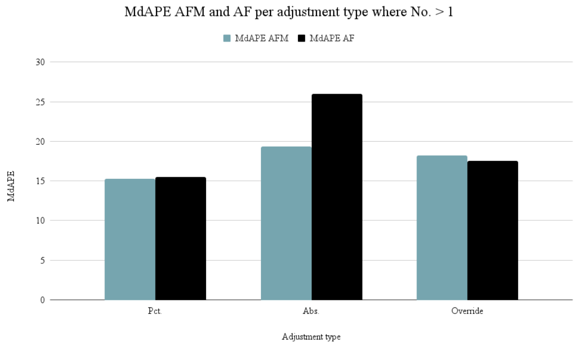

Figure 5.6: MdAPE AFM, AF per adjustment type

Table 5.8: MdAPE RF vs. AF per adjustment type for first adjustment

	Pct. Adjustment	Abs. Adjustment	Override
MdAPE RF	21.86	18.97	20
MdAPE AF	17.5	18.55	18.1
Test Statistic	V = 41,802,442	V = 246,516	V = 668,779
p-value	<0.001***	0.27	0.57
Wilcoxon r	0.18	-	-

In Figure 5.6, the MdAPE of the AFM is plotted against the MdAPE of the newly adjusted forecast (AF). For percentage adjustments, the MdAPE of the AF is slightly higher than the MdAPE of the AFM. However, the Wilcoxon effect size r is below the threshold indicating that the difference in medians is not statistically significant and therefore negligible (Table 5.9). For absolute adjustments, the MdAPE of the AF is significantly higher than the MdAPE of the AFM, and the effect size is medium. This means that subsequent absolute adjustments generally increase the forecast error compared to the already adjusted forecast. The median %p difference in APE relative to the AFM for absolute adjustments is +4.86%p. For subsequent override adjustments, the MdAPE of the AF is slightly lower than the MdAPE of the AFM, however, the effect size is again below the threshold and the effect is therefore negligible. In terms of frequency, the AFM was improved by subsequent adjustments for percentage adjustments in 50.48% of the cases, for absolute adjustments in 33.93% of the cases, and for override adjustments in 52.77% of the cases.

Table 5.9: MdAPE AFM vs. AF per adjustment type

	Pct. Adjustment	Abs. Adjustment	Override
MdAPE AFM	15.22	19.25	18.18
MdAPE AF	15.45	25.88	17.46
Test Statistic	V = 503,943,461	V = 2,603,270	V = 4580344
p-value	0.003**	<0.001***	<0.001***
Wilcoxon r	0.01	0.41	0.08

Domains

Finally, in Figure 5.7, the MdAPE of the RF and the AF are shown per domain. For all domains, the MdAPE of the RF is higher than the MdAPE of the AF. For CO (N = 9,478), OD (N = 6,404) and PO (N = 22,972), this difference in MdAPE is statistically significant, see Table 5.10. However, for CO the effect size does not reach the threshold and is therefore negligible. For VR (N = 30,664), the difference between the MdAPE of the RF and the AF is not significant. For OD the median difference in APE between the RF and the AF is -6.6%p relative to the RF. For PO this is -4.46%p.

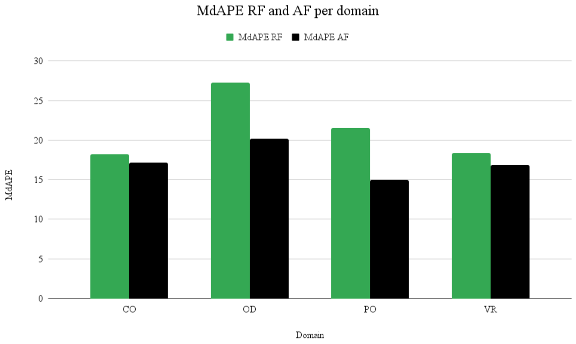

Figure 5.7: MdAPE RF and AF per domain

Table 5.10: MdAPE RF vs. AF per domain

	CO	OD	PO	VR
MdAPE RF	18.14	27.21	21.45	18.32
MdAPE AF	17.1	20.14	14.91	16.75
Test Statistic	V =23,072,413	V = 14,433,490	V = 172,484,907	V = 228,338,899
p-value	0.002**	<0.001***	<0.001***	0.179
Wilcoxon r	0.03	0.36	0.3	-

When investigating how subsequent adjustments perform compared to the previously adjusted forecast for the various domains, it appears that for both CO and OD the MdAPE of the AF is lower than that of the AFM, see Figure 5.8. However, for both the effect size threshold is not reached, see Table 5.11. For PO, there is no significant difference. For VR, the MdAPE of the AF is higher than the MdAPE of the AFM, and this difference is significant. The median difference in APE between the AF and the AFM for VR is +2.44%p relative to the AFM.

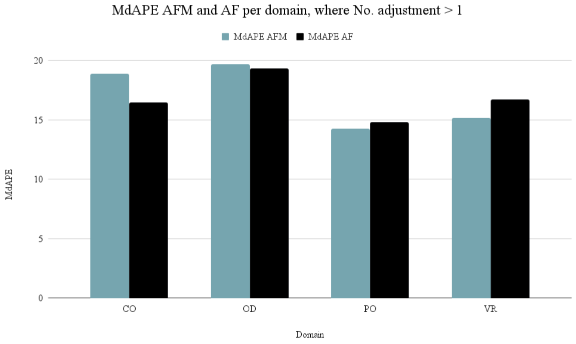

Figure 5.8: MdAPE AFM and AF per domain

Table 5.11: MdAPE AFM vs. AF per domain

	CO	OD	PO	VR
MdAPE AFM	18.84	19.62	14.22	15.14
MdAPE AF	16.43	19.27	14.74	16.69
Test Statistic	V = 15,587,829	V = 8,016,097	V = 104,185,043	V = 94,863,411
p-value	<0.001***	<0.001***	0.141	<0.001***
Wilcoxon r	0.08	0.06	-	0.15

In Table 5.12, for all domains it is shown what the percentage is of adjustments after which the APE of the AF was lower than the APE of the RF. In addition, the percentage of subsequent adjustments that improved the AFM in terms of APE is shown per domain. Striking is that in all domains after more than 50% of the adjustments the APE of the AF is lower than the APE of the RF. Furthermore, except for VR, in all domains at least 50% of the subsequent adjustments contribute to lowering the APE compared to the AFM.

Table 5.12: Performance of adjustments relative to the RF and AFM per domain

	CO	OD	PO	VR
APE AF<RF	50.31%	64.69%	62.6%	50.79%
APE AF<AFM	54.97%	50.87%	50.03%	45.93%

5.3 Direction of judgmental adjustments

5.3.1 Right vs. wrong direction of adjustment

Hypothesis 2a: *In general, analysts are competent in making adjustments in the right direction. That means, adjustments are in the right direction in more than 50% of the cases.* ✔

Hypothesis 2a examined whether analysts are capable of choosing the right direction for an adjustment. For generating accurate forecasts it is critical that adjustments are in the right direction, i.e. upward when required and downward when required, since especially wrong-sided adjustments have a significant negative impact on forecast accuracy.

In Figure 5.9, the percentage of adjustments that were in the right and wrong direction per data set are shown. As can be seen, for both the Covid and the Non-Covid data set, the majority of adjustments (>50%) were in the right direction. In Table 5.13, the exact amounts and percentages are shown. Chi-square goodness-of-fit tests have been applied to see if there is a significant difference from a 50/50 distribution of right and wrong directed adjustments. The results show that this is the case for all data sets. Based on the Non-Covid data, where 62.09% of the adjustments were in the right direction, hypothesis 2a is accepted. Striking is that the proportion of adjustments that was in the right direction, is significantly higher in the Covid data (76.85%) than in the Non-Covid data (62.09%), $\chi^2(1, N = 75,565) = 520.82$, p < 0.001***. The remainder of the analysis for this hypothesis focuses on Non-Covid data.

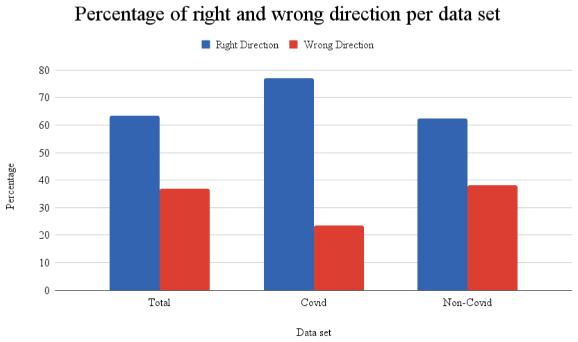

Figure 5.9: Percentage of right and wrong direction per data set

Table 5.13: Percentage of right and wrong direction per data set

	Total	**Covid**	**Non-Covid**
Right Direction	47,810 (63.27%)	4,647 (76.85%)	43,163 (62.09%)
Wrong Direction	27,755 (36.73%)	1,400 (23.15%)	26,355 (37.91%)
Test statistic	$\chi^2(1) = 5,322.6$	$\chi^2(1) = 1,743.5$	$\chi^2(1) = 4,063.8$
p-value	<0.001***	<0.001***	<0.001***
Cramers v	0.27	0.54	0.24

First and Subsequent Adjustments

The first adjustment that is made for a particular date and key, on the pure raw forecast, is in 9,987 (68.65%) cases in the right direction and in 4,560 (31.35%) cases in the wrong direction. This distribution of right and wrong directed adjustments is significantly different from a 50/50 distribution, $\chi^2(1, N = 14,547) = 2,024.6$, p < 0.001***, Cramers v: 0.37. For subsequent adjustments, the distribution of right (60.35%) and wrong (39.65%) directed adjustments is also significantly different from a 50/50 distribution, $\chi^2(1, N = 54,971) = 2,356.3$, p < 0.001***, Cramers v: 0.21. However, the proportion of adjustments in the right direction is significantly higher for first adjustments (68.65%) than for subsequent adjustments (60.35%), $\chi^2(1, N = 69,518) = 336.43$, p < 0.001***. This means that analysts were relatively more capable of choosing the right direction for an adjustment when it was made to the pure raw forecast than when it was made to an already adjusted forecast.

Number of Adjustment and Timing

In Figure 5.10, the percentage of adjustments that were made in the right direction is shown per number of adjustment. It is also noticeable here, as mentioned in the previous paragraph, that the first adjustment is relatively often made in the right direction. Overall, a moderate downward trend is visible in the percentage of adjustments in the right direction with an increasing number of adjustment (-1.112*x + 68.004, $p < 0.001$***, $R^2 = 0.67$). Only when an adjustment is the 17th for a particular date and key, the percentage of adjustments made in the right direction is less than 50%, however, it should be taken into account that the sample size is relatively small for such large numbers of adjustments.

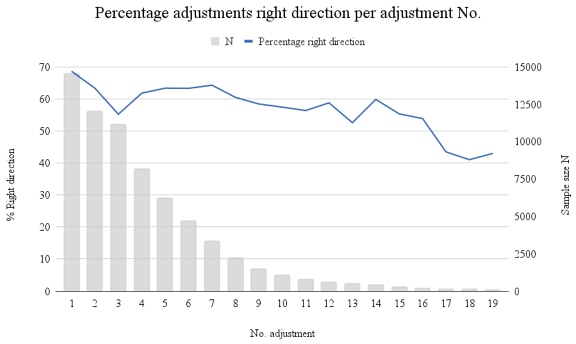

Figure 5.10: % Adjustments right direction per No.

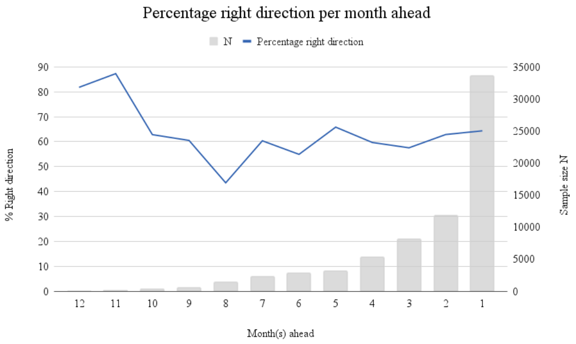

Figure 5.11: % Right direction per month ahead

In Figure 5.11, the percentage of adjustments that were in the right direction is plotted per month ahead. Over the year, no clear trend is visible. However, when considering the 13-week time horizon, in which the majority of all adjustments are made (77.43%), a moderate upward trend of the percentage of adjustments that were made in the right direction can be seen with a decreasing number of weeks in advance (-0.7449*x + 66.5402, $p = 0.002$**, $R^2 = 0.58$). This implies that analysts are relatively better at determining the right direction of an adjustment as the forecast date approaches.

Adjustment Types

In Figure 5.12, the percentage of adjustments that were in the right and wrong direction per adjustment type are shown. As can be seen, for both the percentage and override adjustments, the majority of adjustments was in the right direction. For absolute adjustments, on the other hand, adjustments were more often made in the wrong direction. This may be one of the reasons why this type of adjustment has a worse performance compared to the other adjustment types, as seen in Section 5.2. As can be seen in Table 5.14, for all adjustment types the distribution of right and wrong directed adjustments was significantly different from a 50/50 distribution. However, for absolute adjustments the Cramers v effect size is below the 0.1 threshold value, meaning that this effect can be seen as negligible.

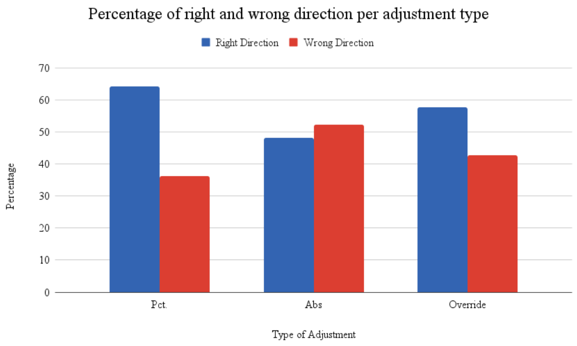

Figure 5.12: Percentage of right and wrong direction per adjustment type

Table 5.14: Percentage of right and wrong direction per adjustment type

	Pct. Adjustment	Abs. Adjustment	Override
Right Direction	36,565 (64.04%)	2,627 (47.93%)	3,295 (57.46%)
Wrong Direction	20,533 (35.96%)	2,854 (52.07%)	2,439 (42.54%)
Test statistic	$\chi^2(1) = 4501.5$	$\chi^2(1) = 9.4014$	$\chi^2(1) = 127.79$
p-value	<0.001***	0.002**	<0.001***
Cramers v	0.28	0.04	0.15

When examining the wrongly directed absolute adjustments (N = 2,854) in more detail, it appears that in 2,273 (79.64%) cases an upward adjustment was made while a downward adjustment was needed. In 581 (20.36%) cases, a downward adjustment was made while an upward adjustment was needed. This indicates that the percentage of wrongly directed absolute adjustments is mostly due to making upward adjustments while downward adjustments were required.

Upward and Downward Adjustments

When examining all upward and downward adjustments, and whether they were in the right direction, it is found that both for upward and downward adjustments the majority of adjustments (> 50%) was made in the right direction, see Table 5.15. However, downward adjustments were significant more often in the right direction (67.83%) compared to upward adjustments (56.09%), $\chi^2(1, N = 69,518) = 1016$, p < 0.001***.

Table 5.15: Right and wrong direction upwards and downward adjustments

	Upwards	Downwards
Right Direction	19,061 (56.09%)	24,102 (67.83%)
Wrong Direction	14,922 (43.91%)	11,433 (32.17%)
Test statistic	$\chi^2(1) = 504.11$	$\chi^2(1) = 4516.8$
p-value	<0.001***	<0.001***
Cramers v	0.12	0.36

Domains

Finally, the finding that in general the majority of adjustments is adjusted in the right direction, is robust across all domains, see Figure 5.13. For all domains, the distribution of the right and wrong directed adjustments is significantly different from a 50/50 distribution, see Table 5.16.

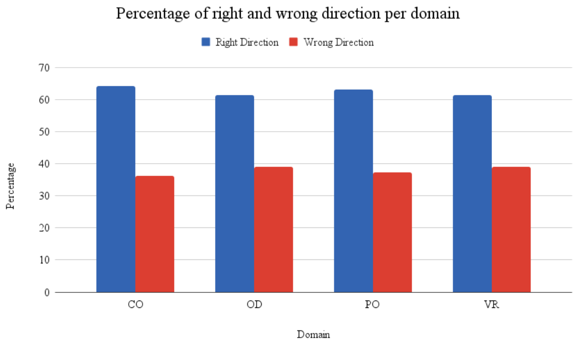

Figure 5.13: Percentage of right and wrong direction per domain

Table 5.16: Right and wrong direction per domain

	CO	OD	PO	VR
Right Direction	6,072 (64.06%)	3,918 (61.18%)	14,435 (62.84%)	18,738 (61.11%)
Wrong Direction	3,406 (35.94%)	2,486 (38.82%)	8,537 (37.16%)	11,926 (38.89%)
Test statistic	$\chi^2(1) = 749.9$	$\chi^2(1) = 320.21$	$\chi^2(1) = 1514.3$	$\chi^2(1) = 1513.3$
p-value	<0.001***	<0.001***	<0.001***	<0.001***
Cramers v	0.28	0.22	0.26	0.22

5.3.2 Frequency of upward and downward adjustments

Hypothesis 2b: *Analysts are in general more likely to adjust a forecast in an upward direction as opposed to making an adjustment in a downward direction.* ✖

Hypothesis 2b tested whether analysts are more inclined to make upward adjustments compared to downward adjustments. For this purpose, the percentage of upward and downward adjustments was examined. In Figure 5.14, the percentage of upward and downward adjustments is shown per data set. For the Covid data, the proportion of upward adjustments is significantly larger than the proportion of downward adjustments, and this distribution is significantly different from a 50/50 distribution, see Table 5.17. For the Non-Covid data, on the other hand, significantly more than 50% of the adjustments was downwards. This would indicate that analysts were more likely to make a downward adjustment compared to making an upward adjustment. However, the Cramers v effect size is below the 0.1 threshold value, which means that this effect can be seen as negligible. Since acceptance and rejection of hypotheses are based on the Non-Covid data, hypothesis 2b can be rejected. While it is not convincing that analysts are more inclined to make downward adjustments, it is in any case not true that analysts are generally more inclined to make upward adjustments. The remainder of the analysis for this hypothesis focuses on Non-Covid data.

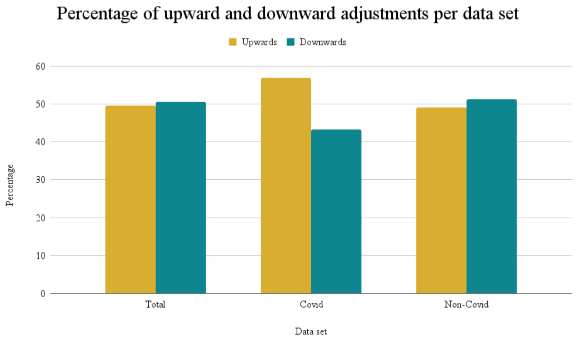

Figure 5.14: Percentage of upward and downward adjustments per data set

Table 5.17: Percentage of upward and downward adjustments per data set

	Total	Covid	Non-Covid
Upwards	37,419 (49.52%)	3,436 (56.82%)	33,983 (48.88%)
Downwards	38,146 (50.48%)	2,611 (43.18%)	35,535 (51.12%)
Test statistic	$\chi^2(1) = 6.9944$	$\chi^2(1) = 112.56$	$\chi^2(1) = 34.649$
p-value	0.008**	<0.001***	<0.001***
Cramers v	0.01	0.14	0.02

Of all 69,518 adjustments in the Non-Covid data set, 30,490 (43.86%) were required to be upward, and 33,983 (48.88%) upward adjustments were actually made. In 39,002 cases (56.1%), a downward adjustment was required, and 35,535 (51.12%) downward adjustments were made. In 26 cases (0.04%), there was no adjustment needed at all. This shows that in terms of what was required, analysts were more inclined to adjust upward. When checking potential bias in the raw forecast, it appears that the RF is slightly more prone to over-forecast (51.47%), than to under-forecast (48.53%). This indicates that relative to the RF, more downward adjustments were required than upward adjustments.

First and Subsequent Adjustments

To find out whether the behavior of analysts differs when making an adjustment to the pure raw forecast compared to making a subsequent adjustment to a previously adjusted forecast, it is examined whether there is a difference in the percentage of upward and downward adjustments for those two kinds of adjustments. It turns out that for first adjustments, the proportion of upward adjustments (56.33%) is significantly higher than the proportion of downward adjustments (43.67%). For subsequent adjustments, it is the other way around, significantly more downward adjustments (53.09%) are made compared to upward adjustments (46.91%), see Table 5.18. However, the Cramers v effect size is below the 0.1 threshold value for the subsequent adjustments, and therefore this effect is negligible. Nevertheless, striking is that analysts are generally more inclined to make an adjustment upward when they make an adjustment to the pure raw forecast, in contrast to the results for subsequent adjustments.

Table 5.18: % upward and downward adjustments for first and subsequent adjustments

	First Adjustment	Subsequent Adjustments
Upwards	8,194 (56.33%)	25,789 (46.91%)
Downwards	6,353 (43.67%)	29,182 (53.09%)
Test statistic	$\chi^2(1) = 232.99$	$\chi^2(1) = 209.43$
p-value	<0.001***	<0.001***
Cramers v	0.13	0.06

Number of Adjustment and Timing

In Figure 5.15, the percentage of upward adjustments is plotted per number of adjustment. It can be seen that, for first adjustments, the percentage of upward adjustments is relatively high (56.33%), as also mentioned in the previous paragraph. Overall, there is no significant trend. However, from adjustment number 2 till adjustment number 12, a moderate upward trend of the percentage of upward adjustments is visible with an increasing number of the adjustment (0.938*x + 42.637, $p = 0.001$, $R^2 = 0.73$), where x is the number of the adjustment. From when it is the second adjustment for a certain key and date up to and including the 7th, the percentage of upward adjustments is below 50%, indicating that relatively more downward adjustments were made. From the time that the adjustment number is 8, the percentage of upward adjustments is higher than the percentage of downward adjustments until the number of the adjustment is 15. However, it must be taken into account that the sample size decreases significantly as the number of the adjustment increases.

In Figure 5.16, the percentage of upward adjustments is plotted per month ahead. Both 11 and 12 months ahead, no upward adjustments were made at all. Only for adjustments made 6, 7 and 8 months ahead the percentage of upwards adjustments is larger than 50%, i.e. relatively more upward than downward adjustments were made. From 5 months in advance the percentage of upward adjustments is below 50%. Overall, there is no significant trend over the 12 months time horizon. For the 13-week time horizon, there is also no significant trend in the percentage of upward (and downward) adjustments.

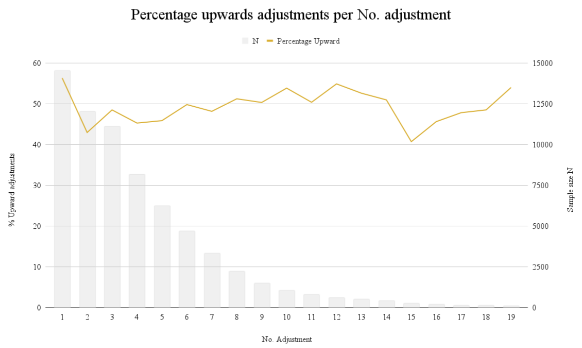

Figure 5.15: % Upward adjustments per No.

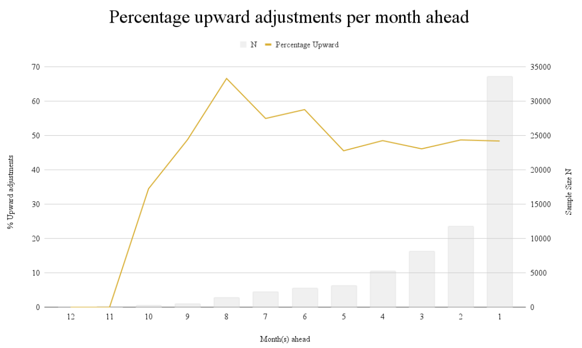

Figure 5.16: % Upward adjustments per month ahead

Adjustment Types

To see whether there is a difference in the percentage of upward and downward adjustments for different adjustment types, Figure 5.17 is shown. For both percentage and override adjustments, the proportion of downward adjustments is significantly higher than the proportion of upward adjustments, see Table 5.19. However, the Cramers v effect size is again below the 0.1 threshold value for both types of adjustments, indicating that the effect is neglectable. For absolute adjustments, the proportion of upward adjustments is significantly larger than the proportion of downward adjustments, and the Cramers v effect size can be classified as large.

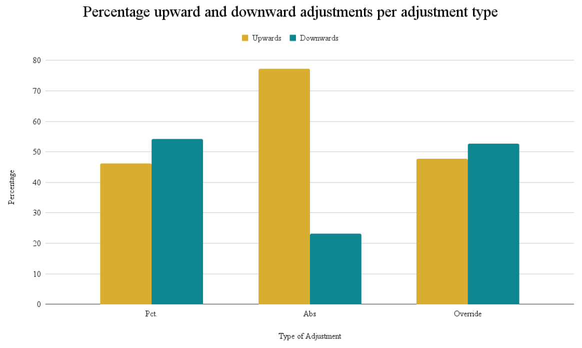

Figure 5.17: Percentage upward and downward adjustments per adjustment type

Table 5.19: Percentage upward and downward adjustments per adjustment type

	Pct. Adjustment	Abs. Adjustment	Override
Upwards	26,277 (46.02%)	4,225 (77.08%)	2,724 (47.51%)
Downwards	30,821 (53.98%)	1,256 (22.92%)	3,010 (52.49%)
Test statistic	$\chi^2(1) = 361.62$	$\chi^2(1) = 1608.3$	$\chi^2(1) = 14.265$
p-value	<0.001***	<0.001***	<0.001***
Cramers v	0.08	0.54	0.05

Domains

Finally, the results for the various domains differ in terms of the percentage of upward and downward adjustments, see Figure 5.18. For both CO and OD, the distribution is significantly different from a 50/50 distribution in the sense that more downward than upwards adjustments were made. However, the Cramers v effect size is below the 0.1 threshold value for OD. For PO the distribution does not differ from a 50/50 distribution, which indicates that approximately as much upward as downward adjustments have been made within this domain. For VR, relatively more adjustments were made upwards, however, again, the Cramers v effect size is below the 0.1 threshold value here, see Table 5.20.

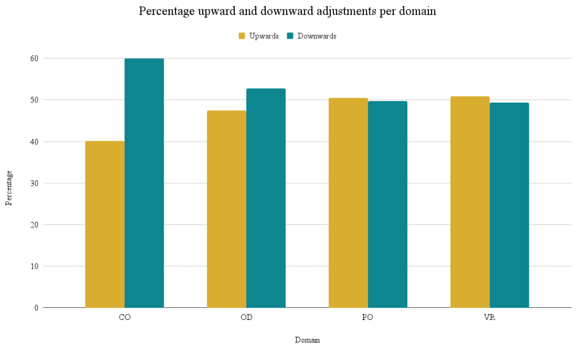

Figure 5.18: Percentage upward and downward adjustments per domain

Table 5.20: Percentage upward and downward adjustments per domain

	CO	OD	PO	VR
Upwards	3,795 (40.04%)	3,032 (47.35%)	11,578 (50.4%)	15,578 (50.8%)
Downwards	5,683 (59.96%)	3,372 (52.65%)	11,394 (49.6%)	15,086 (49.2%)
Test statistic	$\chi^2(1) = 376.09$	$\chi^2(1) = 18.051$	$\chi^2(1) = 1.4738$	$\chi^2(1) = 7.8941$
p-value	<0.001***	<0.001***	0.225	0.005**
Cramers v	0.2	0.05	-	0.02

5.3.3 Performance of upward and downward adjustments

Hypothesis 2c: *Downward judgmental adjustments are in general more likely to decrease the forecasting error of operationally oriented forecasts than upward judgmental adjustments.* ✔

For hypothesis 2c, it was investigated whether downward adjustments are more inclined to lower the error of the operational forecasts than upward adjustments. For all upward and downward adjustments, it was tested whether after an adjustment the adjusted forecast (AF) had a lower error than the forecast generated by the algorithm (RF). In addition, to check the isolated effect of upward and downward adjustments, it has been investigated whether and how much upward and downward adjustments, if they were the very first one for a particular key and date, improved the forecast compared to the RF, and it has been investigated whether and how much upward and downward adjustments, if they were subsequent adjustments, improved the forecast compared to the already adjusted forecast (AFM).

In Figure 5.19, it can be seen that, for both the Covid and the Non-Covid data, and after both upward and downward adjustments, the MdAPE of the AF is lower than the MdAPE of the RF. In Table 5.21, it can be seen that, except for upward adjustments in the Covid data set, this difference in MdAPE is significant. However, for upward adjustments in the Non-Covid data set, the Wilcoxon r effect size is below the threshold value of 0.1 and therefore, this effect can be considered negligible. For downward adjustments, on the other hand, the Wilcoxon r effect size can be classified as large and small for Covid and Non-Covid data respectively. This is already an indication that downward adjustments are in general more likely to decrease the forecasting error than upward adjustments. However, the isolated effect of upward and downward adjustments must be checked to be able to draw conclusions.

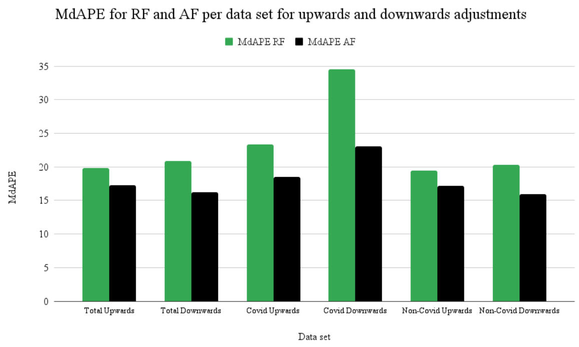

Figure 5.19: MdAPE RF and AF upward and downward for all adjustments

Table 5.21: MdAPE of RF and AF per data set for upwards and downwards adjustments

	Total Upwards	Total Downwards	Covid Upwards	Covid Downwards	Non-Covid Upwards	Non-Covid Downwards
MdAPE RF	19.71	20.82	23.25	34.45	19.37	20.2
MdAPE AF	17.22	16.18	18.42	22.95	17.08	15.81
Test Statistic	V = 353,892,088	V = 454,069,978	V = 2,638,435	V = 2,683,958	V = 295,223,377	V = 385,361,651
p-value	<0.001***	<0.001***	0.323	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.03	0.25	-	0.52	0.03	0.23

To investigate the isolated effect of upward and downward adjustments, the performance of first adjustments relative to the RF is shown in Figure 5.20 and Table 5.22. Here, it can be seen that for first upward adjustments, for the Covid data, there is no significant difference between the MdAPE of the RF and the AF, and that for the Non-Covid data the effect size is below the threshold value. For first downward adjustments, the MdAPE of the AF is significantly lower than the MdAPE of the RF for both the Covid and the Non-Covid data. This means that generally, with regard to first adjustments, downward adjustments outperform upward adjustments in the sense that they are more likely to decrease the forecast error. In terms of magnitude, for first downward adjustments in the Non-Covid data, the median difference in APE relative to the RF is -4.85%p.

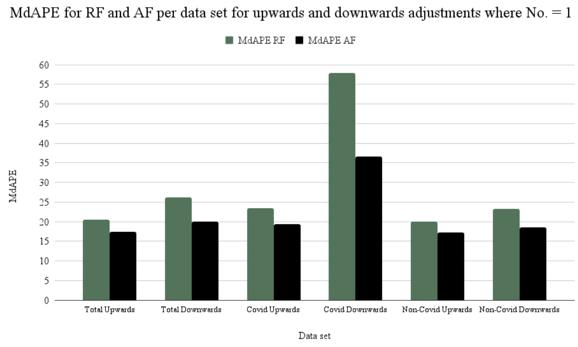

Figure 5.20: MdAPE RF, AF for up and down

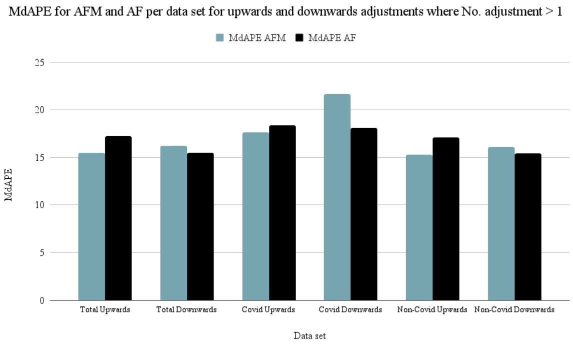

Figure 5.21: MdAPE AFM, AF for up and down

Table 5.22: MdAPE for RF and AF per data set for first upwards and downwards adjustments

	Total Upwards	Total Downwards	Covid Upwards	Covid Downwards	Non-Covid Upwards	Non-Covid Downwards
MdAPE RF	20.35	26.08	23.19	57.75	19.91	23.81
MdAPE AF	17.32	19.86	19.14	36.51	17.14	18.37
Test Statistic	V = 21,589,646	V = 20,165,431	V = 201,756	V = 699,443	V = 17,620,804	V = 13,100,622
p-value	<0.001***	<0.001***	0.38	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.05	0.34	-	0.66	0.05	0.26

To investigate whether there is a difference between subsequent upward and downward adjustments in terms of improving a previously adjusted forecast, the difference in MdAPE between the AFM and the AF was checked for upward and downward adjustments. As can be seen in Figure 5.21, for both the Covid and the Non-Covid data set, the MdAPE of the AF is higher than the MdAPE of the AFM for upward adjustments. For downward adjustments, the MdAPE of the AF is lower than the MdAPE of the AFM. These differences in MdAPE are all significant, as can be seen in Table 5.23. In terms of magnitude for the Non-Covid data, for subsequent adjustments, the median difference in APE is +2.89%p relative to the AFM for upward adjustments and -2.17%p for downward adjustments. This difference is significant, W = 456,068,663, p < 0.001***, $r = 0.18$. This means that generally, with regard to subsequent adjustments, downward adjustments are more likely to decrease the forecasting error than upward adjustments. In fact, it turns out that in general subsequent upward adjustments are detrimental to forecast accuracy. Since both for first adjustments and for subsequent adjustments the downward adjustments outperform the upward adjustments in terms of lowering the forecast error, hypothesis 2c is accepted. The remainder of the analysis for this hypothesis focuses on Non-Covid data.

Table 5.23: MdAPE of AFM, AF per data set for adjustments up and down, No. >1

	Total Upwards	Total Downwards	Covid Upwards	Covid Downwards	Non-Covid Upwards	Non-Covid Downwards
MdAPE AFM	15.44	16.17	17.56	21.56	15.24	16.03
MdAPE AF	17.17	15.45	18.31	18.01	17.05	15.34
Test Statistic	V = 148,693,245	V = 267,581,713	V = 1,417,276	V = 601,804	V = 120,056,430	V = 242,791,340
p-value	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.22	0.13	0.11	0.28	0.24	0.12

When expressing the results for the Non-Covid data in percentages (Table 5.24), it is seen that, taking into account all adjustments, the percentage of downward adjustments after which the APE of the AF was lower than that of the RF (58,26%), is significant larger than the percentage of upward adjustments (53,44%), $\chi^2(1, N = 69,518) = 163.47$, p < 0.001***. When focusing on the isolated effect of upward and downward adjustments, it can be seen that of all first downward adjustments 59.42% decreased the APE, where 53.91% of all first upward adjustments decreased the APE relative to the RF. This difference in proportions is significant, $\chi^2(1, N = 14,547) = 44.025$, p < 0.001***. In addition, for subsequent adjustments, the percentage of downward adjustments that improved the APE relative to the AFM (55,4%) is significantly larger than the percentage of upward adjustments (42,17%), $\chi^2(1, N = 54,971) = 958.37$, p < 0.001***. This is additional support for hypothesis 2c. Downward adjustments are indeed more likely to decrease the forecast error, both relative to the RF and to the AFM, compared to upward adjustments.

Table 5.24: Performance of the AF in percentages for upward and downward adjustments

	RF (All)		RF (First)		AFM	
	Upward	**Downward**	**Upward**	**Downward**	**Upward**	**Downward**
AF better	18,162 (53.44%)	20,704 (58.26%)	4,417 (53.91%)	3,775 (59.42%)	10,876 (42.17%)	16,168 (55.4%)
AF worse	15,532 (45.71%)	14,141 (39.79%)	3,729 (45.51%)	2,557 (40.25%)	14,905 (57.8%)	13,011 (44.59%)
AF equal	289 (0.85%)	690 (1.94%)	48 (0.59%)	21 (0.33%)	8 (0.03%)	3 (0.01%)

Adjustment Types

In Figure 5.22, it can be seen that, taking into account all adjustments, after both percentage and override adjustments, the AF has a lower MdAPE than the RF for both upward and downward adjustments. However, for upward adjustments the Wilcoxon r effect size is below the threshold value of 0.1 for both types of adjustments, see Table 5.25. After absolute downward adjustments, the difference is not significant. After absolute upward adjustments, the AF has a higher MdAPE than the RF. This is already an indication that the overall effect seen with upward and downward adjustments is also reflected in the various types of adjustments. To check the isolated effect of upward and downward adjustments per adjustment type, the first adjustment and subsequent adjustments were examined relative to the RF and the AFM, respectively.

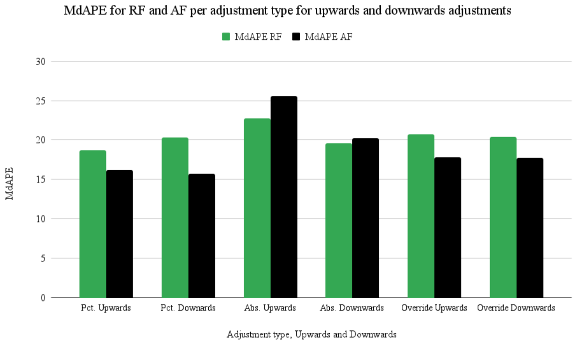

Figure 5.22: MdAPE RF, AF upward and downward per adjustment type for all adjustments

Table 5.25: MdAPE of RF and AF per adjustment type for upwards and downwards

	Pct. Upwards	Pct. Downwards	Abs. Upwards	Abs. Downwards	Override Upwards	Override Downwards
MdAPE RF	18.64	20.23	22.64	19.53	20.67	20.28
MdAPE AF	16.12	15.62	25.46	20.14	17.73	17.63
Test Statistic	V = 183,803,529	V = 292,690,761	V = 3,179,981	V = 415,949	V = 2,031,591	V = 2,603,281
p-value	<0.001***	<0.001***	<0.001***	0.098	<0.001***	<0.001***
Wilcoxon r	0.08	0.25	0.25	-	0.08	0.13

When focusing on the first adjustments, see Figure 5.23 and Table 5.26, it can be seen that only for downward percentage adjustments there is a significant difference in MdAPE between the RF and the AF that reaches the threshold value for the effect size. This indicates that first downward percentage adjustments, in general, decrease the forecast error relative to the RF. All other types of first adjustments do not result in a significant difference in MdAPE.

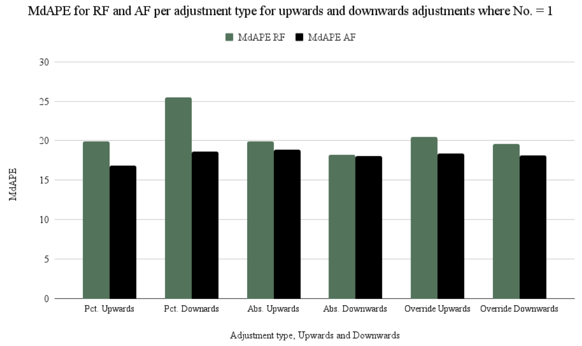

Figure 5.23: MdAPE RF, AF per adjustment type

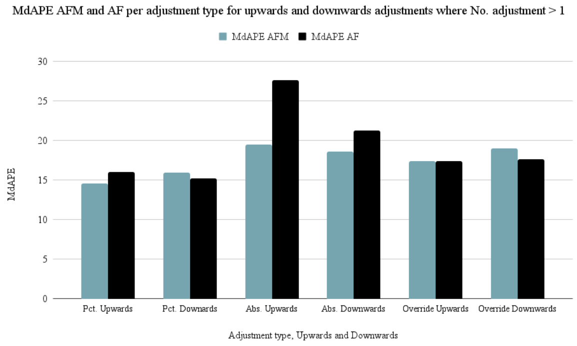

Figure 5.24: MdAPE AFM, AF per adjustment type

Table 5.26: MdAPE for RF, AF per adjustment type for first up- and downwards adjustments

	Pct. Upwards	Pct. Downwards	Abs. Upwards	Abs. Downwards	Override Upwards	Override Downwards
MdAPE RF	19.87	25.38	19.83	18.11	20.37	19.51
MdAPE AF	16.77	18.53	18.8	17.97	18.28	18.06
Test Statistic	V = 11,673,306	V = 9,204,359	V = 89,402	V = 39,041	V = 209,208	V = 129,714
p-value	<0.001***	<0.001***	0.12	0.88	0.44	0.94
Wilcoxon r	0.07	0.32	-	-	-	-

In terms of improving the accuracy of a forecast that has already been adjusted, it appears that for percentage adjustments, the MdAPE of the AF is higher than the MdAPE of the AFM for upward adjustments, where for downward adjustments, the MdAPE of the AF is lower than the MdAPE of the AFM, see Figure 5.24. These differences are significant, see Table 5.27. This means that, in general, subsequent upward percentage adjustments increase the error relative to the previous adjusted forecast, and that subsequent downward percentage adjustments generally decrease the error. For subsequent absolute adjustments, the MdAPE of the AF is higher than that of the AFM for both upward and downward adjustments. For absolute downward adjustments the Wilcoxon *r* effect size is below the threshold value of 0.1 and thus negligible. Subsequent absolute upward adjustments, on the other hand, are especially detrimental in terms of increasing the forecast error when taking into account the effect size, which is nearly large. Finally, for override adjustments, for both the upward and downward adjustments, MdAPE of the AF is lower than that of the AFM.

Table 5.27: MdAPE of AFM and AF per adjustment type for upward and downward, No. >1

	Pct. Upwards	Pct. Downwards	Abs. Upwards	Abs. Downwards	Override Upwards	Override Downwards
MdAPE AFM	14.46	15.82	19.41	18.46	17.31	18.92
MdAPE AF	15.92	15.09	27.54	21.15	17.28	17.53
Test Statistic	V = 74,579,614	V = 186,078,378	V = 1,421,609	V = 172,040	V = 699,161	V = 1,676,641
p-value	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.2	0.12	0.49	0.07	0.13	0.24

To provide more insight into the performance of upward and downward adjustments per adjustment type, the frequencies are given with which they outperformed the RF and the AFM. In Table 5.28, the percentages, per adjustment type, of upward and downward adjustments after which the AF had a lower APE than the RF are shown for all adjustments and for first adjustments. In addition, the frequencies with which subsequent upward and downward adjustments improved the AFM are shown per adjustment type. Here, it can be seen that for all adjustment types, subsequent upward adjustments improve the forecast in less than 50% of the cases.

Table 5.28: Performance AF relative to RF, AFM in % for up and down per adjustment type

	RF (All)	RF (First)	AFM
Pct. Up	54.50%	54.83%	44.4%
Pct. Down	58.65%	61.18%	55.15%
Abs. Up	43.79%	48.39%	31.23%
Abs. Down	51.75%	50.13%	45.19%
Override Up	55.76%	51.28%	42.62%
Override Down	56.38%	52.29%	60.86%

Domains

When examining the performance of upward and downward adjustments for the different domains, it can be seen that, except for upward adjustments for VR, after both upward and downward adjustments the AF has a lower MdAPE than the RF, see Figure 5.25. For upward adjustments for VR, the AF MdAPE is higher than the RF MdAPE. The differences in MdAPE are significant for all domains, for both upward and downward adjustments, see Table 5.29.

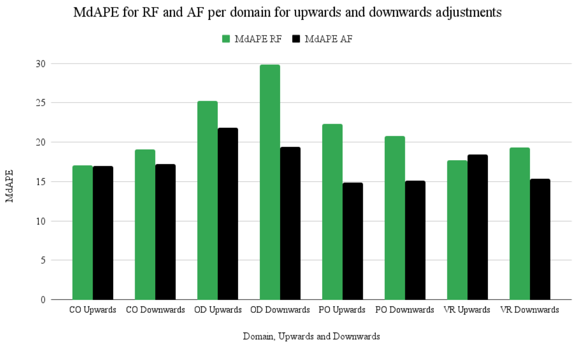

Figure 5.25: MdAPE RF, AF per domain for all adjustments

Table 5.29: MdAPE of RF and AF per domain for upwards and downwards adjustments

	CO Upwards	CO Downwards	OD Upwards	OD Downwards	PO Upwards	PO Downwards	VR Upwards	VR Downwards
MdAPE RF	16.98	19.01	25.14	29.76	22.21	20.7	17.64	19.23
MdAPE AF	16.86	17.17	21.72	19.32	14.81	15.01	18.35	15.25
Test Statistic	V = 3,087,261	V = 9,283,624	V = 2,647,798	V = 4,623,602	V = 44,959,883	V = 41,328,576	V = 50,211,980	V = 63,932,248
p-value	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.12	0.14	0.13	0.55	0.31	0.29	0.13	0.16

To investigate the isolated effects of upward and downward adjustments for the different domains, the performance of first adjustments relative to the RF is shown in Figure 5.26 and Table 5.30. It can be seen that, in general, first downward adjustments decrease the MdAPE relative to the RF in every domain. First upward adjustments in VR do not result in a significant difference in MdAPE, and first upwards adjustments in CO cause an increase in the MdAPE relative to the RF. First upward adjustments in OD and PO generally decrease the forecast error compared to the RF.

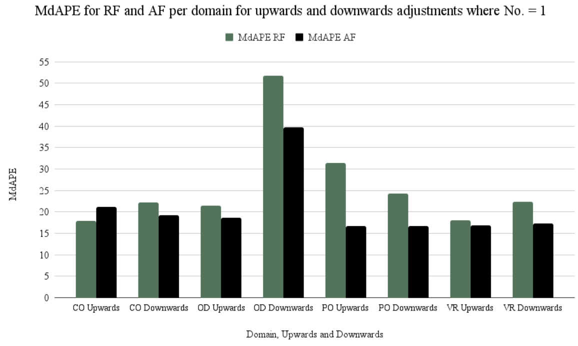

Figure 5.26: MdAPE RF, AF per domain

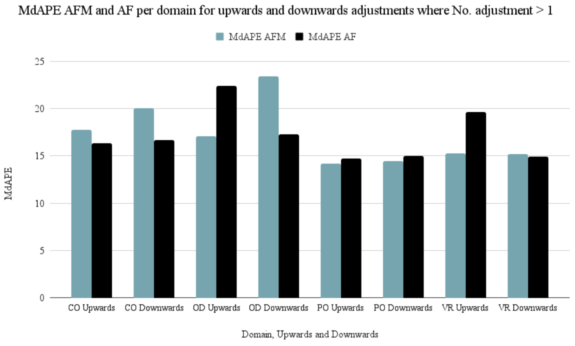

Figure 5.27: MdAPE AFM, AF per domain

Table 5.30: MdAPE for RF and AF per domain for first upwards and downwards adjustments

	CO Upwards	CO Downwards	OD Upwards	OD Downwards	PO Upwards	PO Downwards	VR Upwards	VR Downwards
MdAPE RF	17.73	22.08	21.38	51.63	31.26	24.2	17.9	22.25
MdAPE AF	20.97	19.1	18.54	39.59	16.62	16.63	16.76	17.1
Test Statistic	V = 72,086	V = 487,614	V = 32,063	V = 123,366	V = 656,238	V = 521,999	V = 8,376,118	V = 3,411,497
p-value	<0.001***	<0.001***	0.01*	<0.001***	<0.001***	<0.001***	0.31	<0.001***
Wilcoxon r	0.36	0.25	0.14	0.67	0.66	0.36	-	0.18

When comparing the MdAPE of the AFM and the AF for the upward and downward adjustments within the domains, it can be seen in Figure 5.27, that for upward adjustments the MdAPE of the AF is only lower than the MdAPE of the AFM for CO. However, the Wilcoxon *r* effect size is below the threshold value of 0.1 and therefore negligible. Upward adjustments in all other domains did not result in a significant difference in MdAPE (PO) or in an increase of the forecasting error compared to the previous adjusted forecast (OD and VR). Downward adjustments, on the other hand, resulted in no significant difference (PO and VR) or in a decrease of the forecast error (CO and OD), see Table 5.31.

Table 5.31: MdAPE of AFM and AF per domain for upward and downward, No. >1

	CO Upwards	**CO Downwards**	**OD Upwards**	**OD Downwards**	**PO Upwards**	**PO Downwards**	**VR Upwards**	**VR Downwards**
MdAPE AFM	17.65	19.96	16.99	23.32	14.11	14.35	15.15	15.14
MdAPE AF	16.25	16.59	22.34	17.21	14.63	14.88	19.55	14.83
Test Statistic	V = 2,217,063	V = 6,050,936	V = 771,824	V = 3,268,248	V = 25,257,103	V = 26,823,868	V = 12,284,113	V = 37,078,209
p-value	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	0.002**	<0.001***	<0.001***
Wilcoxon r	0.06	0.19	0.48	0.53	0.05	0.03	0.41	0.07

In Table 5.32, the percentages of upward and downward adjustments after which the AF had a lower APE than the RF for all and for first adjustments, and the frequency with which the AFM was improved in terms of APE for upward and downward adjustments are shown per domain. It can be seen that, except for PO, first upward adjustments improve the forecast in <50% of the cases. Furthermore, especially for OD and VR the frequency with which subsequent upward adjustments have improved the forecast is relatively low.

Table 5.32: Performance AF relative to RF, AFM in % for up and down per domain

	RF (All)	**RF (First)**	**AFM**
CO Up	47.06%	39.89%	54.09%
CO Down	52.47%	55.42%	55.59%
OD Up	56.37%	46.67%	26.04%
OD Down	72.18%	83.87%	73.92%
PO Up	64.72%	85.83%	49.79%
PO Down	60.45%	56.79%	50.27%
VR Up	46.05%	49.43%	34.63%
VR Down	55.68%	58.01%	55.3%

5.4 The size of adjustments

Hypothesis 3: *Large judgmental adjustments are generally more likely to decrease the forecasting error of operationally oriented forecasts than small judgmental adjustments.* ***Undecided***

For hypothesis 3, it was examined whether relatively large adjustments are more inclined to reduce the forecast error than relatively small adjustments. For all relatively large and small adjustments it was tested whether after an adjustment the AF had a lower error than the RF. To check the isolated effect of large and small adjustments, it has been investigated whether and how much large and small adjustments, if they were the very first one for a particular key and date, improved the forecast compared to the RF, and it has been investigated whether and how much large and small adjustments, if they were subsequent adjustments, improved the forecast compared to the AFM.

In Figure 5.28, it can be seen that for both Covid and Non-Covid data, after both relatively large and relatively small adjustments, the MdAPE of the AF is lower than the MdAPE of the RF. These differences in MdAPE are all statistically significant, see Table 5.33. However, for relatively large adjustments in the Non-Covid data the Wilcoxon *r* effect size is below the threshold value of 0.1 and therefore negligible. This means that, for the Non-Covid data, overall, after relatively small adjustments the performance compared to the RF is better than after relatively large adjustments. However, the isolated effects of large and small adjustments must be checked to be able to draw conclusions.

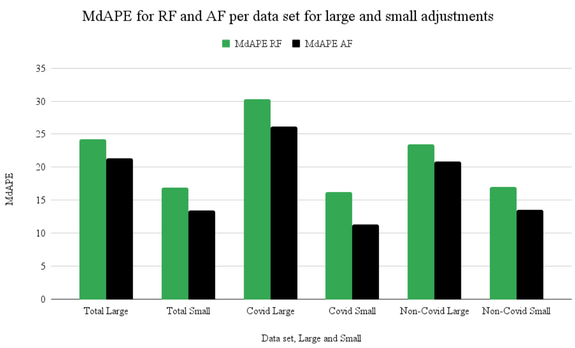

Figure 5.28: MdAPE RF, AF for Large and Small

Table 5.33: MdAPE of RF and AF per data set for large and small adjustments

	Total Large	Total Small	Covid Large	Covid Small	Non-Covid Large	Non-Covid Small
MdAPE RF	24.15	16.85	30.24	16.17	23.34	16.88
MdAPE AF	21.23	13.32	26.05	11.21	20.8	13.41
Test Statistic	V = 387,064,769	V = 424,446,040	V = 6,193,774	V = 694,691	V = 294,416,144	V = 390,811,645
p-value	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.07	0.23	0.23	0.27	0.05	0.23

To investigate the isolated effect of relatively large and small adjustments, the performance of first adjustments relative to the RF is shown in Figure 5.29 and Table 5.34. It can be seen that for the Covid data, first small and first large adjustments decrease the MdAPE relative to the RF, where the effect size is largest for relatively large adjustments. For the Non-Covid data, there is no significant difference in MdAPE after relatively small adjustments, and relatively large adjustments generally decrease the forecast error relative to the RF. This means that generally, with regard to first adjustments in the Non-Covid data, large adjustments outperform small adjustments regarding decreasing the forecast error. In terms of magnitude, for first large adjustments in the Non-Covid data, the median difference in APE relative to the RF is -8.85%p.

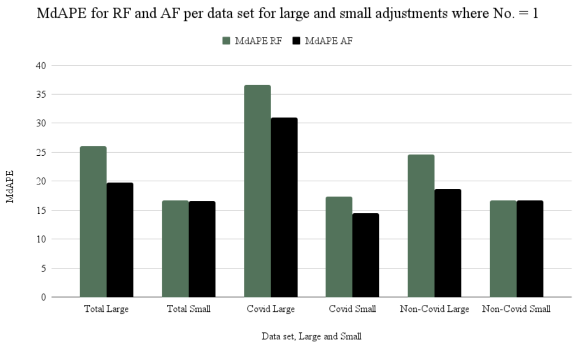

Figure 5.29: MdAPE RF, AF for Large and Small

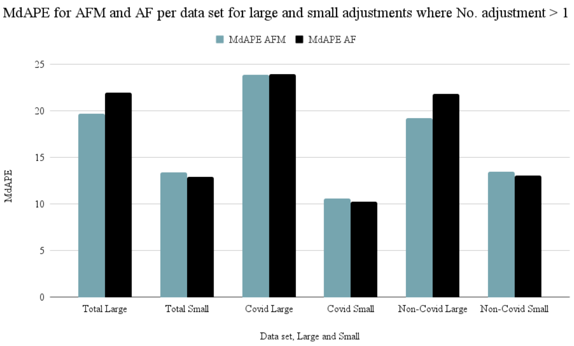

Figure 5.30: MdAPE AFM, AF for Large and Small

Table 5.34: MdAPE for RF and AF per data set for first large and small adjustments

	Total Large	Total Small	Covid Large	Covid Small	Non-Covid Large	Non-Covid Small
MdAPE RF	25.88	16.55	36.45	17.27	24.49	16.53
MdAPE AF	19.67	16.43	30.93	14.42	18.54	16.54
Test Statistic	V = 36,821,092	V = 8,796,481	V = 1,240,513	V = 34,918	V = 24,364,650	V = 7,719,985
p-value	<0.001***	0.02*	<0.001***	0.02*	<0.001***	0.09
Wilcoxon r	0.23	0.03	0.46	0.13	0.18	-

To investigate whether there is a difference between subsequent large and small adjustments in terms of improving a previously adjusted forecast, the difference in MdAPE between the AFM and the AF was checked for large and small adjustments. In Figure 5.30, it can be seen that for relatively large adjustments in the Non-Covid data, the MdAPE of the AF is higher than the MdAPE of the AFM, this difference is significant, see Table 5.35. This means that for the Non-Covid data, subsequent large adjustments generally increase the forecast error. In terms of magnitude, the median difference in APE relative to the AFM is +5.87%p for these relatively large subsequent adjustments. The difference in MdAPE for relatively large adjustments in the Covid data is not significant. For relatively small adjustments, the MdAPE of the AF is lower than the MdAPE of the AFM in both the Covid and Non-Covid data. However, the effect sizes are both below the threshold value and therefore negligible. Since for the Non-Covid data the MdAPE of the AF for relatively small adjustments is negligible lower than the MdAPE of the AFM, and that for relatively large adjustments the MdAPE of the AF is significantly higher than the MdAPE of the AFM, this means that generally, with regard to subsequent adjustments, small adjustments are more likely to decrease the forecasting error than large adjustments. The remainder of the analysis for this hypothesis focuses on Non-Covid data.

Table 5.35: MdAPE of AFM and AF per data set for large and small adjustments, No. >1

	Total Large	Total Small	Covid Large	Covid Small	Non-Covid Large	Non-Covid Small
MdAPE AFM	19.57	13.3	23.79	10.48	19.15	13.38
MdAPE AF	21.89	12.85	23.81	10.14	21.7	12.94
Test Statistic	V = 158,869,776	V = 267,038,480	V = 1,958,945	V = 341,385	V = 125,035,510	V = 248,371,329
p-value	<0.001***	<0.001***	0.851	0.002**	<0.001***	<0.001***
Wilcoxon r	0.14	0.07	-	0.09	0.16	0.07

When expressing the results for the Non-Covid data in percentages (Table 5.36), it can be seen that, taking into account all adjustments, the proportion of adjustments after which the APE of the AF is lower than the APE of the RF is significantly higher for relatively small adjustments (58.25%) compared to relatively large adjustments (53.43%), $\chi^2(1, N = 69,518) = 163.5$, $p < 0.001$***. When focusing on the isolated effect of large and small adjustments, it can be seen that of all first large adjustments 60.46% decreased the APE, where 49.53% of all first small adjustments decreased the APE relative to the RF. This difference in proportions is significant, $\chi^2(1, N = 14,547) = 165.75$, $p < 0.001$***. The proportion of subsequent adjustments decreasing the APE of the AFM is larger for relatively small adjustments (52.47%) compared to relatively large adjustments (45.18%), $\chi^2(1, N = 54,971) = 288.88$, $p < 0.001$***. Based on all findings above, hypothesis 3 can neither be accepted nor rejected. When it is the first adjustment for a particular key and date, relatively large adjustments are more likely to decrease the forecast error. However, when it is a subsequent adjustment, relatively small adjustments are more likely to decrease the forecast error.

Table 5.36: Performance of the AF in percentages for large and small adjustments

	RF (All)		RF (First)		AFM	
	Large	**Small**	**Large**	**Small**	**Large**	**Small**
AF better	18,029 (53.43%)	20,837 (58.25%)	5,461 (60.46%)	2,731 (49.53%)	11,166 (45.18%)	15,878 (52.47%)
AF worse	15,412 (45.67%)	14,261 (39.87%)	3,533 (39.11%)	2,753 (49.93%)	13,544 (54.81%)	14,372 (47.5%)
AF equal	304 (0.9%)	675 (1.89%)	39 (0.43%)	30 (49.93%)	2 (0.01%)	9 (0.03%)

First and Subsequent Adjustments

In order to find out whether analyst behavior differs in terms of the size of an adjustment for first adjustments to the pure raw forecast compared to subsequent adjustments, the median size was compared. The median size of an adjustment across the entire Non-Covid training data set is 10.43%. The median size of an adjustment that was the first one for a certain date and key is 15%. Finally, the median size of subsequent adjustments is 9.75%. The difference in median size for first and subsequent adjustments is significant, W = 481,852,937, $p < 0.001$***, $r = 0.32$. This means that analysts generally make larger adjustments to the pure raw forecast than to an already adjusted forecast. In Figure 5.31, the median size of an adjustment is plotted per number of adjustment. It can be seen that with an increasing number of adjustment, the size of an adjustment decreases (-0.5603*x + 12.5345, $p < 0.001$*** $R^2 = 0.85$).

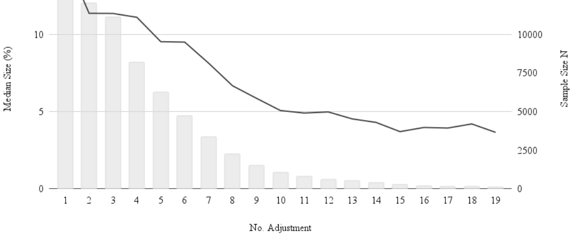

Figure 5.31: Median size of adjustments per No. adjustment

Adjustment Types

It was also investigated whether there is a difference in performance of relatively large and small adjustments for the different adjustment types. First, the ratio of relatively large and small adjustments per adjustment type will be discussed. For percentage adjustments, in 50.65% of the cases the adjustment was relatively large and for 49.35% of the cases relatively small. For absolute adjustments, relatively more adjustments were large (59.73%) than small (40.27%). And for override adjustments, the vast majority of adjustments was relatively small (81.53%), and only a small part of the adjustments was relatively large (18.47%). When focusing on the MdAPE of the AF and the RF for the various adjustment types, it can be seen that, when significance and the threshold value for the effect size are taken into account, after relatively small percentage and override adjustments the AF MdAPE is lower than the RF MdAPE. In addition, after relatively large absolute adjustments, the MdAPE of the AF is significantly higher than the MdAPE of the RF, see Figure 5.32 and Table 5.37. For all other adjustments, the difference in MdAPE was not significant or the effect size threshold value was not reached. To check the isolated effect of large and small adjustments per adjustment type, the first adjustment and subsequent adjustments are examined relative to the RF and the AFM, respectively.

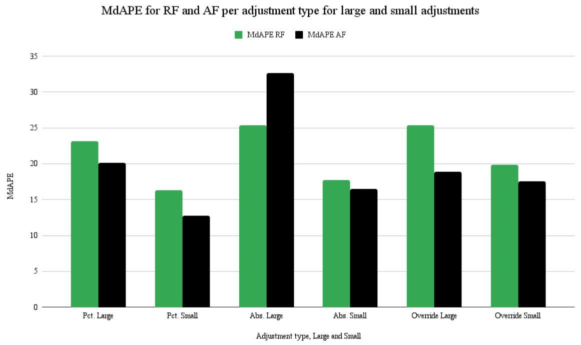

Figure 5.32: MdAPE RF, AF large and small per adjustment type for all adjustments

Table 5.37: MdAPE of RF and AF per adjustment type for large and small adjustments

	Pct. Large	Pct. Small	Abs. Large	Abs. Small	Override Large	Override Small
MdAPE RF	23.06	16.25	25.28	17.66	25.28	19.76
MdAPE AF	20.08	12.68	32.59	16.36	18.83	17.45
Test Statistic	V = 226,309,906	V = 247,364,555	V = 1,708,677	V = 1,265,978	V = 308,430	V = 6,139,033
p-value	<0.001***	<0.001***	<0.001***	0.111	0.005**	<0.001***
Wilcoxon r	0.09	0.26	0.31	-	0.09	0.11

When focusing on first adjustments, see Figure 5.33 and Table 5.38, it can be seen that only for large percentage and large override adjustments, the MdAPE of the AF is significantly lower than the MdAPE of the RF, and that the threshold value for the effect size is reached. All other types of first adjustments do not result in a significant difference in MdAPE or did not reach the effect size threshold of 0.1. This indicates that, in general, regarding first adjustments, only the large percentage and large override adjustments significantly decreased the forecast error relative to the RF.

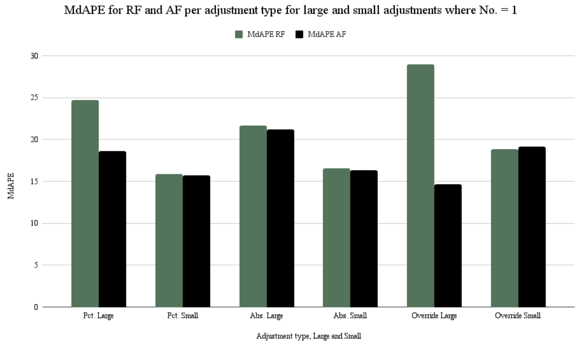

Figure 5.33: MdAPE RF, AF per adjustment type

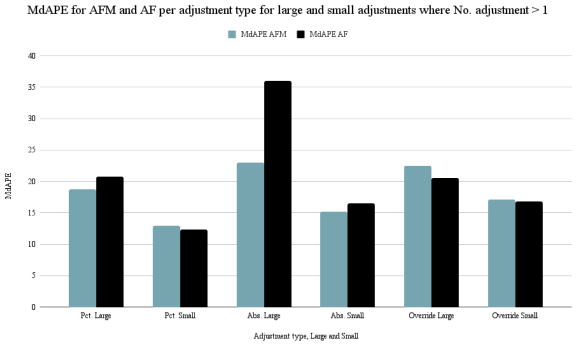

Figure 5.34: MdAPE AFM, AF per adjustment type

Table 5.38: MdAPE for RF and AF per adjustment type for first large and small adjustments

	Pct. Large	Pct. Small	Abs. Large	Abs. Small	Override Large	Override Small
MdAPE RF	24.65	15.8	21.56	16.47	28.87	18.76
MdAPE AF	18.53	15.67	21.13	16.23	14.57	19.08
Test Statistic	V = 20,662,837	V = 3,379,910	V = 56,001	V = 66,665	V = 14,742	V = 463,614
p-value	<0.001***	<0.001***	0.03*	0.77	<0.001***	0.004**
Wilcoxon r	0.19	0.07	0.09	-	0.46	0.08

With regard to subsequent adjustments, it can be seen that absolute adjustments, regardless of the relative size, have a higher MdAPE of the AF compared with the AFM, see Figure 5.34. The difference in MdAPE is significant, and it is striking that the Wilcoxon effect size for large absolute adjustments is large, see Table 5.39. Indicating that, subsequent large absolute adjustments, in general, significantly increase the forecast error relative to the previous adjusted forecast. With regard to percentage adjustments, the MdAPE of the AF for relatively small adjustments is lower than the MdAPE of the AFM, while for relatively large adjustments the MdAPE of the AF is higher than that of the AFM. This indicates that for subsequent percentage adjustments, generally the relatively large ones increase the forecast error while relatively small ones decrease the forecast error, which is in line with the general results. For override adjustments, the difference in MdAPE is not significant for relatively large adjustments and the Wilcoxon r effect size threshold value is not reached for relatively small adjustments.

Table 5.39: MdAPE of AFM and AF per adjustment type for large and small, No. >1

	Pct. Large	Pct. Small	Abs. Large	Abs. Small	Override Large	Override Small
MdAPE AFM	18.68	12.9	22.95	15.07	22.36	16.98
MdAPE AF	20.72	12.27	35.92	16.39	20.44	16.67
Test Statistic	V = 93,213,554	V = 168,942,355	V = 745,483	V = 527,726	V = 192,085	V = 2,906,779
p-value	<0.001***	<0.001***	<0.001***	<0.001***	0.37	<0.001***
Wilcoxon r	0.11	0.1	0.53	0.23	-	0.09

To provide more insight into the performance of relatively large and small adjustments per adjustment type, in Table 5.40, the percentages of large and small adjustments after which the AF had a lower APE than the RF per adjustment type are shown for all adjustments and for first adjustments. In addition, the percentage of large and small subsequent adjustments that improved the AFM in terms of the APE are shown per adjustment type. It can be seen that for first adjustments, irrespective of the adjustment type, small adjustments are successful in <50% of the cases. With regard to subsequent adjustments, large percentage adjustments, and absolute adjustments irrespective of their size, are relatively often ineffective, with absolute large adjustments being the worst.

Table 5.40: Performance AF relative to RF, AFM in % for large and small per adjustment type

	RF (All)	RF (First)	AFM
Pct. Large	54.6%	60.92%	46.91%
Pct. Small	58.94%	49.99%	53.48%
Abs. Large	41.11%	48.50%	29.61%
Abs. Small	52.29%	49.61%	41%
Override Large	57.98%	77.04%	52.38%
Override Small	55.66%	48.25%	52.88%

Large, Small, Upward and Downward Adjustments

The combination of relatively large and small adjustments and upward and downward adjustments has also been examined. First of all, the median size of upward adjustments is slightly larger (10.68%) than the median size of downward adjustments (10.2%), W = 637,154,606, p-value < 0.001***, $r = 0.07$. However, the Wilcoxon r effect size is below the threshold value of 0.1 and therefore this difference can be considered negligible. In terms of frequency, for upward adjustments, there is no significant difference in the amount of relatively large and small adjustments, 50.08% of the adjustments were large and 49.92% were small. For downward adjustments, significantly more adjustments were small (52.93%), compared to large (47.07%), $\chi^2(1, N = 35, 535) = 121.63$, p < 0.001***, Cramer's $v = 0.06$. However, the Cramer's v effect size is below the threshold value of 0.1, and therefore this difference is again negligible.

When taking into account all adjustments (Figure 5.35), it appears that after all combinations of large and small and upward and downward adjustments, the MdAPE of the AF is lower than the MdAPE of the RF. These differences are all significant, however, the Wilcoxon effect size r is below the threshold value for large upwards adjustments, and therefore this difference is negligible, see Table 5.41. This could mean that large upwards adjustments are generally ineffective, but the isolated effect should be checked.

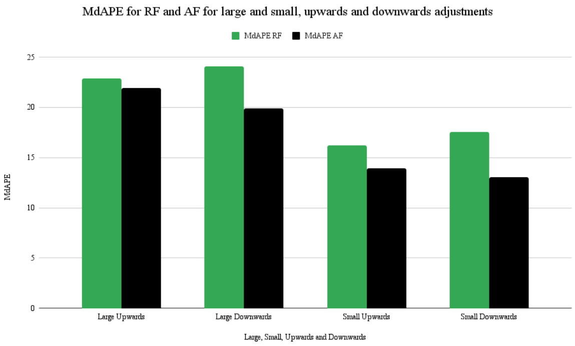

Figure 5.35: MdAPE RF and AF large and small for upward and downward for all adjustments

Table 5.41: MdAPE of RF and AF for large and small, upwards and downwards adjustments

	Large Upwards	Large Downwards	Small Upwards	Small Downwards
MdAPE RF	22.83	24.03	16.17	17.48
MdAPE AF	21.89	19.84	13.88	13
Test Statistic	V = 65,391,816	V = 81,952,792	V = 82,251,375	V = 114,058,745
p-value	<0.001***	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.08	0.18	0.15	0.3

To investigate the isolated effect of the combinations of upward, downward, large and small adjustments, the performance of first adjustments relative to the RF is shown in Figure 5.36 and Table 5.42. It can be seen that the Wilcoxon effect size r is below the threshold for first large upwards adjustments. First small upward adjustments, in general, increase the forecast error relative to the RF. First large and small downward adjustments generally decrease the error relative to the RF, where the effect size is the largest for first large downward adjustments.

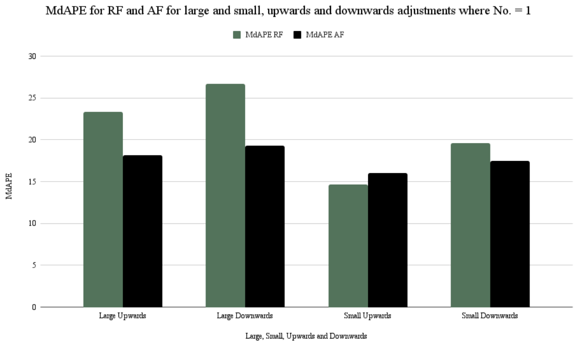

Figure 5.36: MdAPE RF, AF large, small, up, down

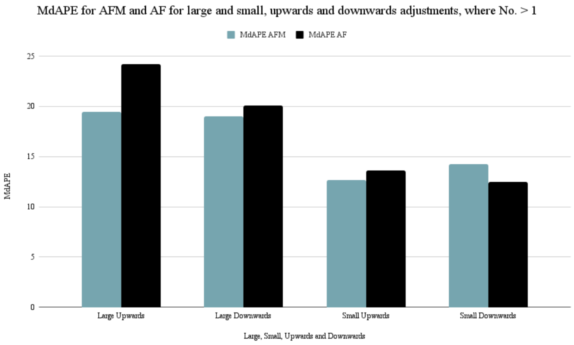

Figure 5.37: MdAPE AFM, AF large, small, up, down

Table 5.42: MdAPE for RF, AF for first large and small, up- and downwards adjustments

	Large Upwards	Large Downwards	Small Upwards	Small Downwards
MdAPE RF	23.3	26.65	14.61	19.51
MdAPE AF	18.11	19.21	15.96	17.4
Test Statistic	V = 7,340,088	V = 4,933,215	V = 1,853,848	V = 1,961,383
p-value	<0.001***	<0.001***	<0.001***	<0.001***
Wilcoxon r	0.08	0.32	0.13	0.19

With regard to the performance of subsequent adjustments (Figure 5.37), it can be seen that large and small upward adjustments, generally increased the forecast error relative to the AFM, see Table 5.43. The effect size for small upward adjustments is small and for large upwards adjustments medium. Subsequent small downward adjustments, on the other hand, in general, decreased the forecast error relative to the AFM. For subsequent large downward adjustments the effect size is below the threshold value and therefore negligible.

Table 5.43: MdAPE of AFM and AF for large and small, upward and downward, No. >1

	Large Upwards	Large Downwards	Small Upwards	Small Downwards
MdAPE AFM	19.36	18.93	12.6	14.21
MdAPE AF	24.16	20	13.54	12.43
Test Statistic	V = 20,669,133	V = 43,024,156	V = 35,175,571	V = 91,364,457
p-value	<0.001***	0.002**	<0.001***	<0.001***
Wilcoxon r	0.35	0.03	0.24	0.33

In Table 5.44, the frequency with which the APE of the AF was lower than the APE of the RF for all adjustments and for first adjustments, after all combinations, is shown. In addition, the frequency with which the different kind of adjustments improved the AFM is shown. It can be seen that small upward adjustments are effective in <50% of the cases, irrespective if they were the first or a subsequent adjustment. Furthermore, for subsequent adjustments, in terms of the frequency with which an adjustment improved the forecast accuracy, small downward adjustments are most effective and large upward adjustments are least effective.

Table 5.44: Performance AF relative to RF, AFM in % for large, small, upward, downward

	RF (All)	**RF (First)**	**AFM**
Large Upward	51.05%	59.38%	41.8%
Small Upward	55.85%	44.30%	42.49%
Large Downward	55.85%	61.93%	48.28%
Small Downward	60.41%	55.65%	61.06%

Domains

Finally, the performance of large and small adjustments was examined for the various domains. In Figure 5.38, the MdAPE of the RF and the AF are shown per domain after all relatively large and small adjustments. When taking into account significance and the Wilcoxon effect size threshold, see Table 5.45, it appears that for both OD and PO, after both large and small adjustments, the MdAPE of the AF is significantly lower than the MdAPE of the RF. It should be noted that the effect sizes are larger for the relatively small adjustments compared to the relatively large ones. For both CO and VR the differences are not significant or negligible for both large and small adjustments. However, to be able to draw conclusions, the isolated effects of large and small adjustments must be checked per domain.

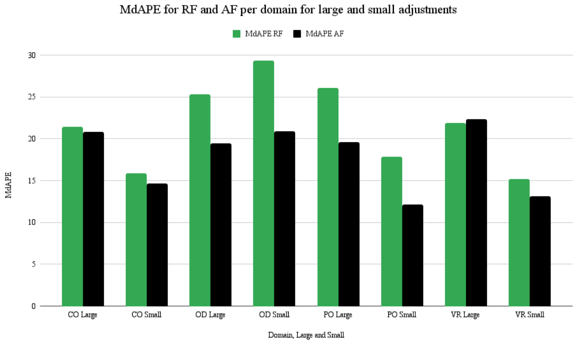

Figure 5.38: MdAPE RF and AF per domain for large and small for all adjustments

Table 5.45: MdAPE of RF and AF per domain for large and small adjustments

	CO Large	**CO Small**	**OD Large**	**OD Small**	**PO Large**	**PO Small**	**VR Large**	**VR Small**
MdAPE RF	21.33	15.81	25.23	29.25	26.04	17.76	21.85	15.12
MdAPE AF	20.74	14.57	19.39	20.84	19.55	12.07	22.25	13.09
Test Statistic	V = 5,817,915	V = 5,725,737	V = 2,434,308	V = 5,027,310	V = 34,547,180	V = 53,088,840	V = 54,003,597	V = 60,500,871
p-value	0.665	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
Wilcoxon r	-	0.07	0.21	0.48	0.21	0.39	0.06	0.09

When examining the first adjustments relative to the RF, it can be seen in Figure 5.39 and Table 5.46, that for OD both first large and small adjustments in general decreased the MdAPE relative to the RF. For PO, first large adjustments decreased the forecast error compared to the RF. For all other first adjustments, there was no significant difference in MdAPE or the threshold value for the effect size was not reached.

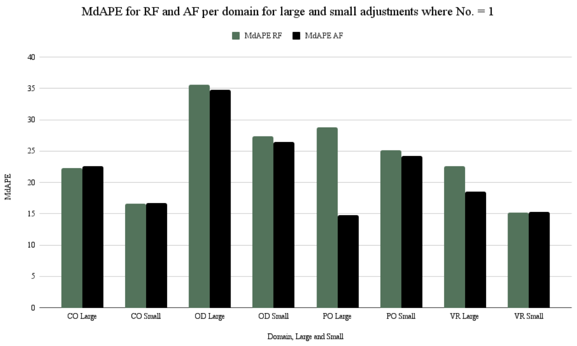

Figure 5.39: MdAPE RF, AF per domain, up/down

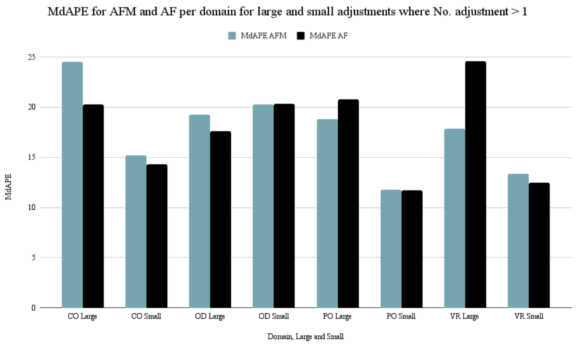

Figure 5.40: MdAPE AFM, AF per domain, up/down

Table 5.46: MdAPE for RF and AF per domain for first large and small adjustments

	CO Large	CO Small	OD Large	OD Small	PO Large	PO Small	VR Large	VR Small
MdAPE RF	22.18	16.49	35.46	27.3	28.66	25	22.47	15.1
MdAPE AF	22.45	16.56	34.72	26.38	14.71	24.09	18.44	15.16
Test Statistic	V = 431,574	V = 96,487	V = 106,026	V = 47,438	V = 1,425,928	V = 96,110	V = 7,610,626	V = 3,773,112
p-value	0.65	0.77	<0.001***	<0.001***	<0.001***	0.04*	<0.001***	0.41
Wilcoxon r	-	-	0.29	0.44	0.58	0.09	0.08	-

In Figure 5.40, the MdAPE of the AF and the AFM are shown for subsequent adjustments. With regard to relatively large and small subsequent adjustments per domain, it can be seen that there are only significant effects with an effect size above the threshold for relatively small adjustments in CO and relatively large adjustments in VR. In all other cases there is either no significant difference in MdAPE or a negligible difference compared to the AFM, see Table 5.47. After relatively small adjustments for CO, the MdAPE of the AF was lower than the MdAPE of the AFM, indicating that, in general, the forecast error was decreased. After relatively large adjustments in VR, the MdAPE of the AF was higher than the MdAPE of the AFM, which means that generally relatively large subsequent adjustments in VR increase the forecast error.

Table 5.47: MdAPE of AFM and AF per domain for large and small adjustments, No. >1

	CO Large	CO Small	OD Large	OD Small	PO Large	PO Small	VR Large	VR Small
MdAPE AFM	24.44	15.17	19.17	20.21	18.73	11.74	17.82	13.33
MdAPE AF	20.22	14.23	17.57	20.31	20.71	11.63	24.53	12.41
Test Statistic	V = 3,229,497	V = 4,645,824	V = 1,288,647	V = 2,898,423	V = 17,595,248	V = 36,026,314	V = 16,393,242	V = 34,905,486
p-value	0.146	<0.001***	0.151	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
Wilcoxon r	-	0.14	-	0.08	0.08	0.04	0.31	0.08

In Table 5.48, the percentages of large and small adjustments after which the AF had a lower APE than the RF for all adjustments and for first adjustments is shown per domain. In addition, for every domain the frequency with which the AFM was improved by large and small adjustments is given. It can be seen that, with regard to first adjustments, the general tendency that small adjustments generally do not improve forecast is seen in three out of the four domains, except for OD. Moreover, especially for PO and VR, the frequency with which subsequent large adjustments improved the forecast accuracy is relatively low, with the lowest percentage for large subsequent adjustments in VR.

Table 5.48: Performance AF relative to RF, AFM in % for large and small per domain

	RF (All)	RF (First)	AFM
CO Large	49.61%	50,23%	53.84%
CO Small	51.04%	48,8%	55.98%
OD Large	58.7%	66,43%	52.17%
OD Small	69.34%	70,62%	49.97%
PO Large	59.93%	80,38%	48.91%
PO Small	64.91%	43,15%	50.87%
VR Large	49.18%	55,42%	37.37%
VR Small	52.42%	48,71%	53.61%

5.5 The influence of timing

Hypothesis 4a: *The forecast error of adjusted forecasts increases with a decreasing time horizon.* ✖

Hypothesis 4b: *Given that the algorithm-generated forecast is updated over time, the forecast accuracy improvement relative to the algorithm-generated forecast decreases with a decreasing time horizon.* ✖

For hypothesis 4a and hypothesis 4b, the performance of judgmental adjustments over time has been investigated. In Table 5.49, it is shown how many adjustments have been made per month ahead. As can be seen, almost half of all adjustments (48.33%) were made within one month before the forecast date. In total, 93.24% of all adjustments were made within 6 months before the forecast date.

Table 5.49: Timing of adjustments

Month	N	% of adjustments	% Cumulative
1	33,595	48.33	48.33
2	11,825	17.01	65.34
3	8,166	11.75	77.08
4	5,269	7.58	84.66
5	3,185	4.58	89.24
6	2,777	3.99	93.24
7	2,242	3.23	96.46
8	1,442	2.07	98.54
9	557	0.80	99.34
10	304	0.44	99.78
11	134	0.19	99.97
12	22	0.03	100.00

1 Year Time Horizon

In Figure 5.41, the MdAPE of the RF and AF are plotted per month ahead. As seen, and confirmed by a Kruskal-Wallis test, there is a significant difference between the MdAPEs per month ahead for the AF, $\chi^2(11, N = 69, 518) = 2256.6$, p < 0.001***. From 12 months to 8 months in advance, the MdAPE of the AF follows an upward trend (-4.375*x + 60.899, p = 0.02*, $R^2 = 0.82$). However, it should be noted that the sample size is relatively small during this period. From 8 months in advance, there is a significant downward trend present with a decreasing number of months in advance (1.65*x + 13.28, p < 0.001***, $R^2 = 0.87$). Since 98.54% of all adjustments were made in the period of 8 to 1 month in advance, this is a first indication that the error of the adjusted forecast is, in general, decreasing with a decreasing time horizon, and that therefore hypothesis 4a could be rejected. What can also be seen is that from 3 months up to and including 1 month in advance, when 77.08% of all adjustments were made, the difference between the RF and the AF increases, indicating an increasing forecast accuracy improvement, contradicting hypothesis 4b.

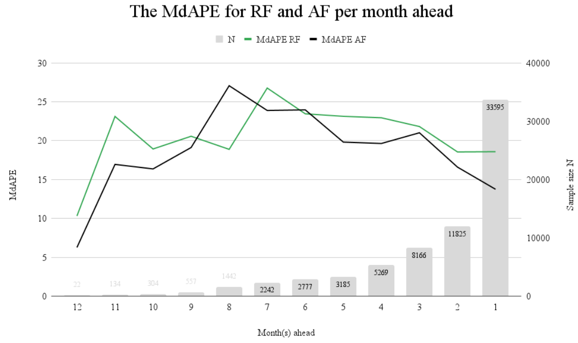

Figure 5.41: MdAPE RF, AF per month ahead

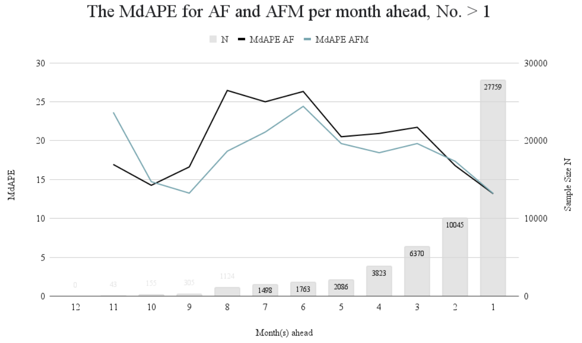

Figure 5.42: MdAPE AFM, AF per month ahead

It can be seen that the RF has a similar kind of pattern in terms of the MdAPE over time. Also for the RF there are significant differences between the MdAPEs per month ahead, $\chi^2(11, N = 69, 518) = 894.17$, p < 0.001***. For the RF it can be seen that from 7 months ahead there is a significant downward trend of the MdAPE with decreasing months ahead (1.274*x + 17.074, p = 0.001**, $R^2 = 0.88$). As seen, and confirmed by a pairwise comparison, there is no significant difference between the MdAPE of 1 and 2 months ahead. This could be due to the fact that the RF is not updated anymore the last 3 weeks before the forecast date.

In Figure 5.42, the MdAPE of the AF and the AFM are plotted against the month ahead to see, over time, whether or not a previously adjusted forecast is, in general, improved by a subsequent adjustment. It is notable that except for 11 and 10 months ahead, where the sample size is relatively small, only subsequent adjustments made within two months before the forecast date generally provide an improvement compared to the AFM.

13 Weeks Time Horizon

Since 77.43% of all adjustments were made in the last quarter before the forecast date, and since this is a commonly used time horizon within the company, there is zoomed in on adjustments made within 13 weeks ahead and later. In Figure 5.43, it can be seen that within the 13-weeks ahead time horizon, most adjustments were made 3 weeks ahead. Because stakeholders start making a planning 3 weeks in advance, this is generally an important moment to ensure that there is a decent forecast in the system. In the second and last week before the forecast date, a relatively large amount is still being adjusted. A Kruskal-Wallis test shows that there is a significant difference between the MdAPEs per week ahead for the AF over the 13-weeks time horizon, $\chi^2(12, N = 53,830) = 1208.4$, p < 0.001***. Striking is that from 8 weeks in advance, analysts are generally capable of improving the forecast compared to the RF. From 6 weeks in advance, this difference in MdAPE between the RF and the AF is significant, V = 2,287,478, p-value < 0.001***, $r = 0.12$. Overall, it can be seen that the MdAPE of the AF is decreasing with a decreasing time horizon. A linear trend line is plotted for the AF (0.8188*x + 11.6424, p < 0.001***, $R^2 = 0.82$), which shows that this trend is significant over the 13-week time horizon. Therefore, hypothesis 4a is rejected.

For the RF, there is also a significant difference between the MdAPEs per week ahead, $\chi^2(12, N = 53,830) = 301.24$, p < 0.001***. And also for the RF, there is a significant downward linear trend with a decreasing time horizon (0.4286*x + 16.7392, p < 0.001***, $R^2 = 0.69$). Since the slope of the trend line is greater for the MdAPE of the AF, this results in an increasing difference between the MdAPE of the AF and the MdAPE of the RF over time. Despite the fact that the statistical forecast is not updated anymore in the last 3 weeks, a growing difference between the MdAPE of the RF and the AF has been visible in the weeks before that, which means that there is an increasing forecast accuracy improvement with a decreasing time horizon, and therefore, hypothesis 4b is rejected.

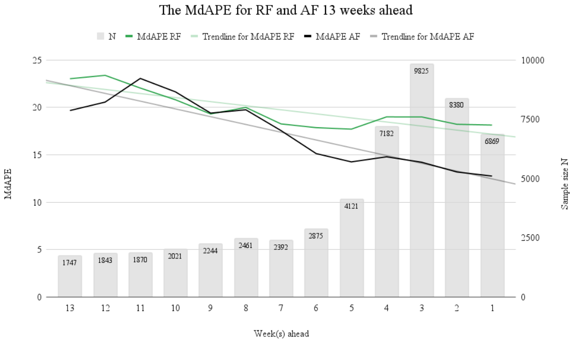

Figure 5.43: MdAPE RF, AF 13 weeks ahead

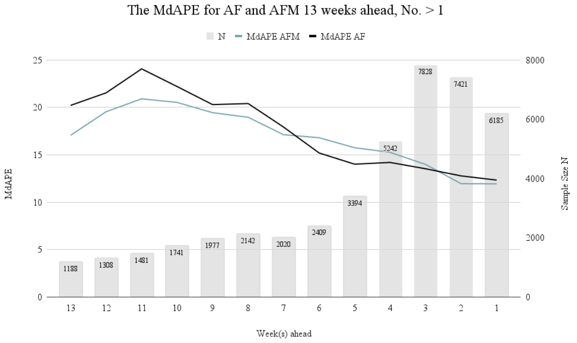

Figure 5.44: MdAPE AF, AFM 13 weeks ahead

1 Month Time Horizon

In Figure 5.44, the MdAPE of the AF and the AFM are plotted over the 13 weeks ahead time horizon to examine the performance of subsequent adjustments per week ahead. It can be seen here that, if subsequent adjustments are made within 13 to 7 weeks before the forecast date, this generally leads to a higher MdAPE compared to the AFM. Adjustments made within 6 to 3 weeks ahead generally lead to an improvement of the already adjusted forecast. In the final two weeks, the MdAPE of the AF is again higher than the MdAPE of the AFM, indicating that, in general, the forecast accuracy is not improved in the final two weeks relative to the already adjusted forecast.

Figure 5.45 has zoomed in on the daily level for the adjustments made in the last month before the forecast date. In the last month, 48.33% of all adjustments were made. A Kruskal-Wallis test shows that there is a significant difference between the MdAPEs per day in the final month for the AF, $\chi^2(29, N = 33,595) = 116.71$, p < 0.001***. With decreasing days ahead the MdAPE of the AF is slightly decreasing (0.06533*x + 12.78923, p = .001**). Striking is that there is a significant increase in MdAPE from day 6 to day 1.

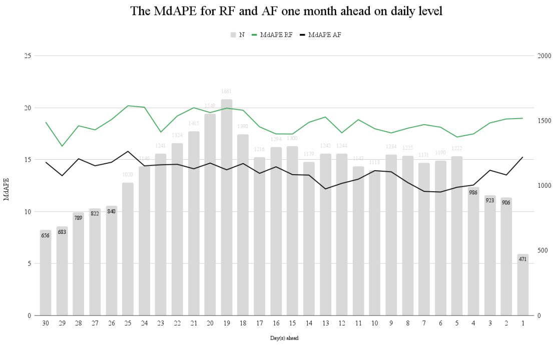

Figure 5.45: MdAPE RF and AF one month ahead

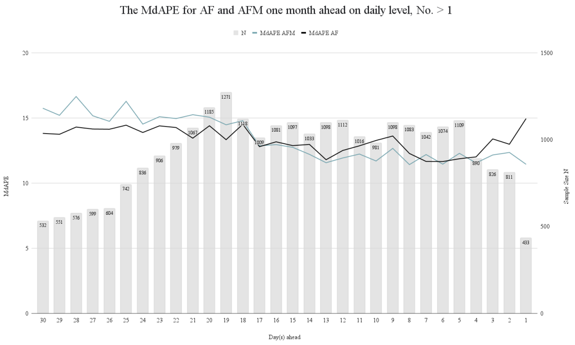

Figure 5.46: MdAPE AF and AFM one month ahead

In Figure 5.46, it can be seen that at the beginning of the last month, the AFM is generally still being improved by subsequent adjustments. After that, except for 7 and 5 days ahead, the MdAPE of the AF is higher than the MdAPE of the AFM. It can also be seen here that in the last 4 days the difference between the MdAPE of the AF and the AFM is significantly greater with the peak being the last day. This indicates that not only is the forecast accuracy, in general, not improving in the last few days, but also that the deterioration of the forecast accuracy is increasing in magnitude over the last 4 days.

Contacts Offered

In Figure 5.47, for CO (Contacts Offered), the MdAPE of the RF and the AF is plotted per week ahead for the 13-week time horizon. There is a significant difference between the MdAPEs per week ahead for the AF, $\chi^2(12, N = 7,929) = 849.17$, p < 0.001***. In general, there is a decreasing trend of the AF MdAPE with a decreasing number of weeks ahead (1.20*x + 12.37, p = .004**). From 8 weeks ahead, the MdAPE of the AF decreases sharply. However, only from 5 weeks ahead, the MdAPE of the AF is lower than the MdAPE of the RF. This indicates that, in general, the adjustments made 5 weeks in advance or later improved the forecast accuracy compared to the RF, which is in contrast to the adjustments made earlier. Regarding subsequent adjustments, over the time horizon of 13 weeks, it can be seen that only from 6 weeks in advance the AF has a lower MdAPE than the AFM, indicating that the already adjusted forecast was generally improved by subsequent adjustments made 6 weeks in advance or later, see Figure 5.48.

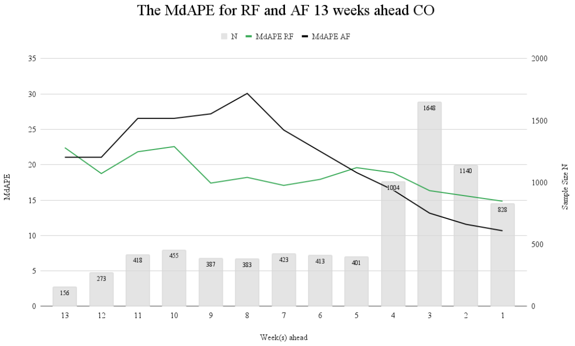

Figure 5.47: RF AF 13 weeks ahead CO

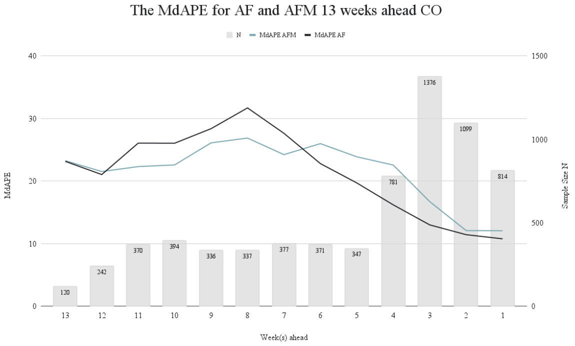

Figure 5.48: AF AFM 13 weeks ahead CO

Orders Delivered

For OD (Orders Delivered), there is also a significant difference in MdAPEs for the AF over the 13 week time horizon, $\chi^2(12, N = 5,135) = 394.45$, $p < 0.001$***. With a decreasing number of weeks ahead, the MdAPE of the AF is decreasing (1.592*x + 12.930, $p < 0.001$***). For OD, the MdAPE of the AF is regardless of the number of weeks ahead lower than the MdAPE of the RF, see Figure 5.49. However, the difference between the MdAPE of the RF and the AF becomes smaller with a decreasing time horizon up to 3 weeks in advance, then the RF MdAPE remains at the same level and the AF MdAPE continues to decrease as the forecast date gets closer. The performance of the AF relative to the AFM alternate over the time horizon of 13 weeks. From 5 weeks in advance, the MdAPE of the AF is lower than the MdAPE of the AFM, see Figure 5.50. This indicates that the already adjusted forecast was generally improved by subsequent adjustments made 5 weeks in advance or later.

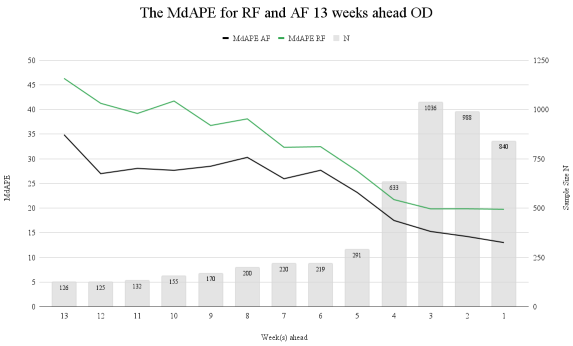

Figure 5.49: RF AF 13 weeks ahead OD

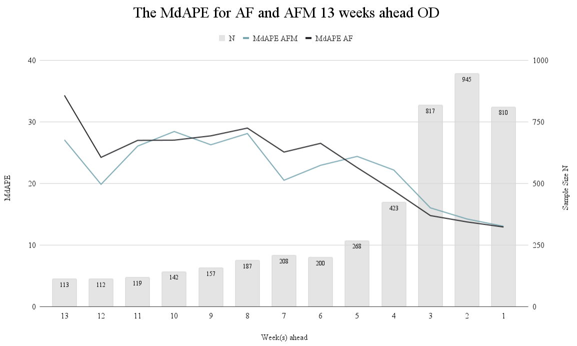

Figure 5.50: AF AFM 13 weeks ahead OD

Products Ordered

Comparable to OD, the MdAPE of the AF is regardless of the number of weeks ahead lower than the MdAPE of the RF for PO (Products Ordered), see Figure 5.51. Over the time horizon of 13 weeks, there is a significant difference in MdAPEs of the AF for PO, $\chi^2(12, N = 15,731) = 368.95$, $p < 0.001$***. With a decreasing number of weeks ahead, the MdAPE of the AF is decreasing (0.7578*x + 9.8909, $p < 0.001$***). This also applies to the MdAPE of the RF (0.9338*x + 15.4957, $p < 0.001$***), which follows a similar pattern. Regarding subsequent adjustments for PO, from 13 to 10 weeks in advance, the accuracy of the forecast is generally deteriorated due to subsequent adjustments. From 9 weeks to 3 weeks in advance, the AFM is generally slightly improved by subsequent adjustments. However, in the last two weeks the AFM is the one with the lowest MdAPE, see Figure 5.52.

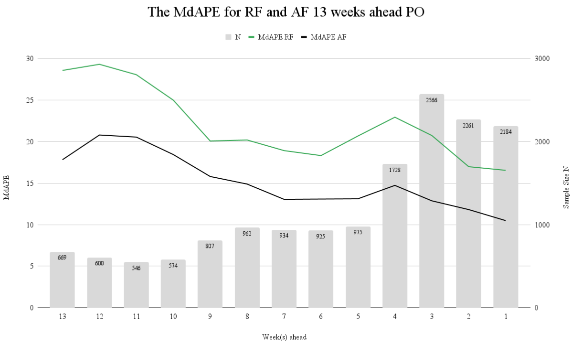

Figure 5.51: RF AF 13 weeks ahead PO

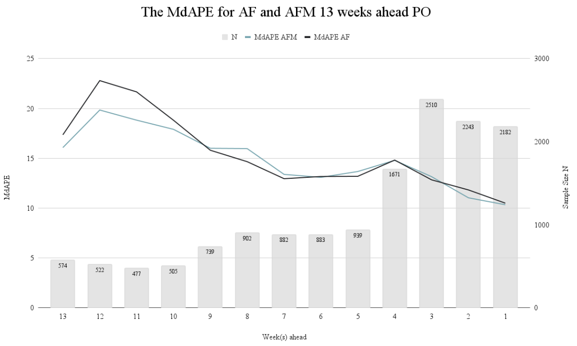

Figure 5.52: AF AFM 13 weeks ahead PO

Visitors Registered
Finally, for VR (Visitors Registered), there is also a significant difference between MdAPEs of the AF over the 13 week time horizon, $\chi^2(12, N = 25,035) = 378.88$, $p < 0.001$***. With a decreasing number of weeks ahead, the MdAPE of the AF is generally decreasing (0.6173*x + 12.9657, p = 0.002**). Especially from 7 to 6 weeks in advance there is a significant decrease in MdAPE for the AF. From 5 weeks ahead, the MdAPE of the AF and the RF rises slightly as the time horizon decreases, see Figure 5.53. Only when adjustments were made 6 weeks ahead or later, the MdAPE of the AF was in lower than the MdAPE of the RF. With regard to improving an already adjusted forecast, it can be seen that for VR, from 13 to 7 weeks in advance, there is generally no improvement in forecast accuracy relative to the AFM. For 6 and 5 weeks in advance, the AF MdAPE is lower than the MdAPE of the AFM, indicating that, in general, the forecast was improved by subsequent adjustments. Adjustments made 4 weeks in advance or later, in general, deteriorate the accuracy again, see Figure 5.54.

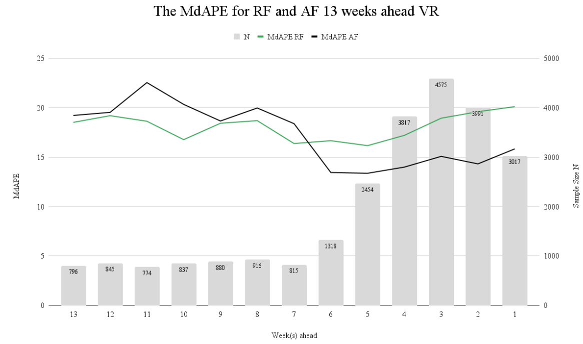

Figure 5.53: RF AF 13 weeks ahead VR

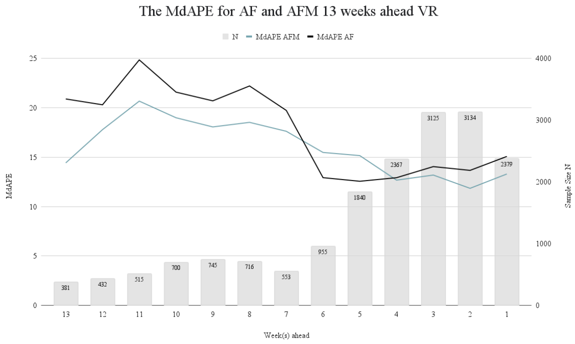

Figure 5.54: AF AFM 13 weeks ahead VR

5.6 Main Findings

The analysis results demonstrate that, overall, judgemental adjustments improve the quality of operationally oriented forecasts. Striking is that this, in contrast to percentage adjustments and overrides, generally does not apply to absolute adjustments. In addition, in approximately half of the cases a subsequent adjustment reduces the forecast error. Furthermore, analysts prove to be competent at adjusting forecasts in the right direction, and there was an equal ratio of upward and downward adjustments. Moreover, downward judgmental adjustments were, in general, more likely to decrease the forecasting error of operationally oriented forecasts than upward judgmental adjustments. Especially subsequent upward adjustments were found to be generally detrimental. Relatively large judgmental adjustments were generally more likely to decrease the forecasting error of operationally oriented forecasts than relatively small judgmental adjustments when it came to the first adjustment for a particular key and date. For subsequent adjustments, relatively small adjustments were more likely to decrease the forecast error, and relatively large adjustments, in general, increased the error in this case. Finally, it turned out that the forecast error of adjusted forecasts decreased with a decreasing time horizon, and that the forecast accuracy improvement increased with a decreasing time horizon. The findings from this chapter will be used in the next chapter to see if the accuracy of adjustments can be improved by applying several new forecasting procedures.

6. Improve adjustment procedure

This chapter is focused on designing an adjustment procedure that takes into account the behavioral tendencies of the analysts to improve the accuracy of the operationally oriented forecasts. The findings from the analysis in Chapter 5 serve as input for this new adjustment procedure. In Section 6.1, the improvement strategies used are discussed. In Section 6.2, improvement models are designed in which the improvement strategies are applied. The models are trained on the Non-Covid training data set, and tested on the Non-Covid test data set, such that recommendations can be given for a normal period, i.e. a period without disruptions.

6.1 Improvement Strategies

Coolblue's adjustment procedure may be enhanced to raise the accuracy of its operationally oriented forecasts based on the results of which judgmental adjustments increase or decrease the forecast accuracy. There are several possible strategies, with differing levels of complexity, to improve the accuracy of adjustments. To ensure that the improvements for the adjustment procedure are targeted at those adjustments that are generally detrimental, the new models purely focus on the adjustments with a relatively moderate to poor performance. Therefore, the models are targeted at those adjustment types that had the highest effect sizes in terms of increasing the forecast error, and the lowest percentage of effective adjustments. In Sections 6.1.1 and 6.1.2, the improvement strategies, that are later applied in the models, are discussed.

6.1.1 Improvement Strategy A

One of the most simple, yet successful methods to reduce bias and inefficiencies, is to average the statistical forecast with the human judgment, as proposed by Blattberg and Hoch (1990). In the Blattberg-Hoch method, the statistical forecast and the judgmental forecast are given equal weight. According to Blattberg and Hoch (1990), the reason an approach as simple as averaging is so successful is that statistical models and humans have different but complimentary strengths and weaknesses. This theory of Blattberg and Hoch (1990), in which the skills of statistical models and humans are combined, is the first improvement strategy that will be applied. For those adjustments that are generally detrimental, the average is taken from the raw forecast (RF) and the adjusted forecast (AF), to investigate whether this can increase the forecast accuracy. As a result of combining these two forecasts, and thereby reducing the total difference between the RF and AF, this strategy dampens all adjustments made so far for a certain key and date. This improvement strategy will from now on be referred to as *improvement strategy A*.

6.1.2 Improvement Strategy B

In the context of this study, the statistical forecast, referred to as the raw forecast (RF) is a point of reference from which the analysts make adjustments, and therefore, the judgmental forecast is not independent. According to Fildes et al. (2009), when the statistical forecast and the judgmental forecast are not independent, instead of averaging the statistical forecast with the judgmental forecast, the judgmental adjustments should be dampened to reduce bias and inefficiency. This is therefore the second improvement strategy that is applied. By dampening the adjustments that are generally detrimental and thereby reducing their impact on forecast accuracy, and making use of the adjustments that are generally beneficial in terms of reducing the forecast error, it is assumed that the overall performance of adjustments can be increased. For this improvement strategy, the AFM is averaged with the AF. This strategy ensures that only the most recently made adjustment for a certain key and date is dampened, and is therefore able to fully focus on dampening only the generally detrimental adjustments. This improvement strategy will from now on be referred to as *improvement strategy B*.

It will be examined which of the two improvement strategies described above works best. Next to using these two improvement strategies, with two averaging methods as a basis, there will be experiments with different weights assigned to the forecasts to investigate whether completely eliminating particular types of adjustments is advantageous or not. In Section 6.2, the improvement models, in which the two improvement strategies are applied, are discussed.

6.2 Improvement Models

In this section, three models with two variants each have been designed based on the results of the analysis. These all focus on a different type of adjustment that could be considered relatively ineffective because it generally increased the error and or improved the forecast in less than 50% of cases. In Section 6.2.4, the performance of all models is evaluated against the current adjusted forecast (AF), since this is the forecast that is now communicated to stakeholders.

6.2.1 Model 1 - Subsequent Upward Adjustments

The results of the analysis showed that subsequent upward adjustments are generally detrimental to the forecast accuracy. The median difference in APE is +2.89%p relative to the AFM, and only 42.17% of the subsequent upward adjustments were effective in reducing the forecast error. In order to investigate whether the performance of these generally harmful adjustments can be improved, the two improvement strategies described in the previous section are applied. In model 1A, the RF and the AF are averaged (Equation 6.1), and in model 1B, a damping effect is applied by averaging the AFM and the AF (Equation 6.2). In order to find out whether tackling these subsequent upward adjustments is effective in itself, the AF is still taken for all other adjustments.

$$\textbf{Model 1A} = \begin{cases} \text{IF upward AND No.} > 1 \text{ THEN } \frac{RF+AF}{2} \\ \text{ELSE AF} \end{cases} \tag{6.1}$$

$$\textbf{Model 1B} = \begin{cases} \text{IF upward AND No.} > 1 \text{ THEN } \frac{AFM+AF}{2} \\ \text{ELSE AF} \end{cases} \tag{6.2}$$

6.2.2 Model 2 - Subsequent Large Adjustments

The second model focuses on subsequent large adjustments. This type of adjustment was generally not conducive to the accuracy of the forecasts. In terms of magnitude, the median difference in APE relative to the AFM is +5.87%p for relatively large subsequent adjustments. In addition, only 45.18% of these adjustments were effective. Again, it is checked whether the generally detrimental impact of these types of adjustments can be reduced by applying the two improvement strategies. In Equation 6.3, improvement strategy A is used, and in Equation 6.4, improvement strategy B is applied. The AF is still taken for all other adjustments in order to determine whether dealing with these subsequent upward adjustments is effective in itself.

$$\textbf{Model 2A} = \begin{cases} \text{IF large AND No.} > 1 \text{ THEN } \frac{RF+AF}{2} \\ \text{ELSE AF} \end{cases} \tag{6.3}$$

$$\textbf{Model 2B} = \begin{cases} \text{IF large AND No.} > 1 \text{ THEN } \frac{AFM+AF}{2} \\ \text{ELSE AF} \end{cases} \tag{6.4}$$

6.2.3 Model 3 - Absolute Adjustments

The last model focuses on absolute adjustments. It was seen that regardless of whether they were first or subsequent, they generally did not contribute to lowering the forecast error or even increase it significantly. Absolute adjustments were effective in less than 50% of the cases and subsequent absolute adjustments even in only 33.93% of the cases. For these subsequent absolute adjustments, the median difference in APE relative to the AFM was +4.86%p. Again, improvement strategy A (Equation 6.5) and improvement strategy B (Equation 6.6) are applied. In order to find out whether tackling these absolute adjustments is effective in itself, the AF is again taken for all other adjustments.

$$\textbf{Model 3A} = \begin{cases} \text{IF absolute THEN } \frac{RF+AF}{2} \\ \text{ELSE AF} \end{cases} \tag{6.5}$$

$$\textbf{Model 3B} = \begin{cases} \text{IF absolute THEN } \frac{AFM+AF}{2} \\ \text{ELSE AF} \end{cases} \tag{6.6}$$

6.2.4 Performance Models 1, 2, 3

In Table 6.1, the MdAPE after applying all six improvement models is given. Compared to the MdAPE of the current AF (16.43), all models show an overall decrease in the MdAPE. In addition, the median %p difference in APE relative to the AF, for the adjustment types addressed in the various models, is given. For all adjustment types addressed in the models, a decrease in forecast error is seen. Since all models do lead to an improvement compared to the current AF, in Section 6.2.5, a model is designed in which all three types of adjustments are taken into account.

Table 6.1: MdAPE and median %p difference relative to the AF for adjustment types addressed in the models

Model	MdAPE	%p difference
1A	15.94	-1.20
1B	15.88	-2.44
2A	15.36	-3.64
2B	15.66	-6.67
3A	16.22	-5.83
3B	16.30	-3.76

6.2.5 Model 4 - Combination Model

The fourth model is a combination of models 1, 2 and 3. Since adjustments can be a combination of upward, large and absolute, in one model only improvement strategy A is used (Equation 6.7), and in the other model only improvement strategy B is used (Equation 6.8). In Section 6.2.6, the performance of the combination model will be discussed.

$$\textbf{CombiModel A} = \begin{cases} \text{IF (upward AND No. } > 1\text{) OR} \\ \quad \text{(large AND No. } > 1\text{) OR} \\ \quad \text{absolute THEN } \frac{RF+AF}{2} \\ \text{ELSE AF} \end{cases} \tag{6.7}$$

$$\textbf{CombiModel B} = \begin{cases} \text{IF (upward AND No. } > 1\text{) OR} \\ \quad \text{(large AND No. } > 1\text{) OR} \\ \quad \text{absolute THEN } \frac{AFM+AF}{2} \\ \text{ELSE AF} \end{cases} \tag{6.8}$$

6.2.6 Performance Combination Model

When the combination models are applied to the Non-Covid training data, CombiModel A, where the RF was averaged with the AF, leads to an MdAPE of 15.31 and CombiModel B, where the generally harmful adjustments were dampened by averaging the AF and the AFM, leads to an MdAPE of 15.47. When comparing the performance of all 8 models, it appears that CombiModel A leads to the lowest MdAPE, and is therefore the best performing model.

To see whether the difference in medians between the current AF (16.43) and CombiModel A (15.31) is significant, the Wilcoxon Signed-Rank test for paired samples is applied. This test shows that the difference is significant, V = 686,600,644, p-value < 0.001***, $r = 0.14$. When checking the median %p difference in APE between the current AF and CombiModel A, for those adjustment types addressed in this model, it is found that this is -1.65%p.

Now that this new model leads to a significant improvement in the training data, it must be tested whether this is also the case in the test data. In the Non-Covid test data, the MdAPE of the AF is 17.11, and the MdAPE of CombiModel A is 16.56. The difference in MdAPE is significant, V = 184,963,593, p-value < 0.001***, $r = 0.1$. This means that CombiModel A, also in the test data, generally produces a lower error than the AF. When taking all adjustment types into account that are addressed in CombiModel A, it is seen that the median %p difference between the APE of the AF and the APE of the CombiModel A forecast is -0.92%p.

Optimize Performance Combination Model
It has been investigated whether assigning different weights to the RF and the AF in CombiModel A can further improve the forecast accuracy. The training data shows that a ratio of 0.4/0.6 for the RF and the AF, respectively, leads to the best performance, i.e. the lowest forecast error. Applying this ratio leads to a MdAPE of 15.24, which is lower than the MdAPE of the 0.5/0.5 model (15.31). The difference in MdAPE between the AF (16.43) and the 0.4/0.6 model (15.24) is again significant, V = 712,708,947, p-value < 0.001***, $r = 0.17$. By applying this ratio, the median difference in %p has also increased relative to the current AF. For those adjustment types addressed in this model, the median %p difference in APE between the current AF and CombiModel A increased from -1.65%p in the 0.5/0.5 model to -1.87%p in the 0.4/0.6 model. It is examined how CombiModel A (0.4/0.6) performs over the course of a year and the 13-week time horizon to determine whether this model actually works under various timing scenarios for adjustments. In Figure 6.1 and Figure 6.2, it can be seen that regardless of the month or week ahead, CombiModel A provides better forecast accuracy than the current AF.

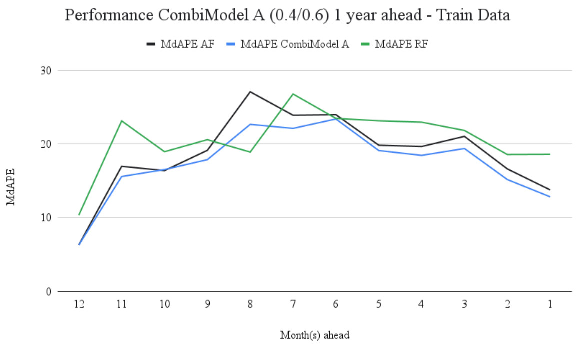

Figure 6.1: CombiModel A (0.4/0.6) 1 year - Train

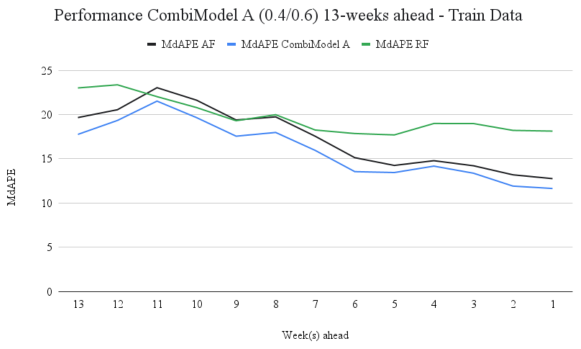

Figure 6.2: CombiModel A (0.4/0.6) 13 weeks - Train

Now that it appears that CombiModel A provides a better performance when applying a 0.4/0.6 ratio compared to the 0.5/0.5 ratio, it must be checked whether this also applies to the test data. The result in the test set does not show a significant improvement over the 0.5/0.5 model. The MdAPE of CombiModel A 0.4/0.6 is 16.54 where the MdAPE of CombiModel A 0.5/0.5 was 16.56. However, based on the test for a significance difference in MdAPE between the AF (17.11) and this new model (16.54), across the entire data set, it appears that the effect size increased slightly, V = 190,873,726, p-value < 0.001***, $r = 0.12$. In addition, the median %p difference increased from -0.92%p to -1.22%p. Therefore, a small improvement is visible.

To see how this model performs over the 1 year and the 13-week time horizon, Figure 6.3 and Figure 6.4 are shown. For the 13-week time horizon no clear improvement is visible. However, over a one year time horizon, it can be seen that by applying this model, benefits are mainly achieved in the long-term adjustments.

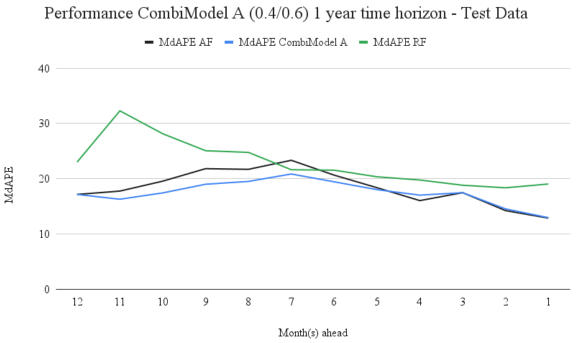

Figure 6.3: CombiModel A (0.4/0.6) 1 year - Test

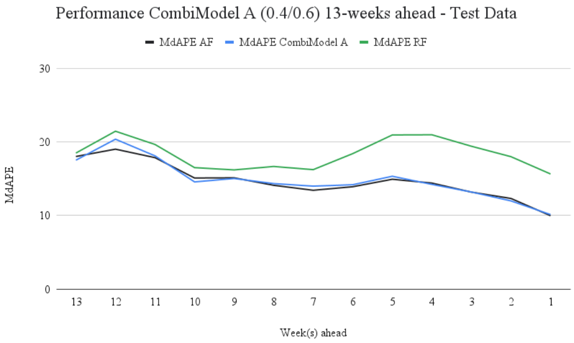

Figure 6.4: CombiModel A (0.4/0.6) 13 weeks - Test

6.3 Domain Specific Models

In the previous section, models were created based on the main results, with the largest effect sizes, and it was tested whether these lead to better forecast accuracy over the entire data set. These effects, however, were not the same across all domains, hence it is also examined which model performs best for each domain. In Table 6.2, is indicated per domain in bold which model performed the best in terms of producing the lowest error (MdAPE) for the training data set. Based on these results, it will be checked per domain, for the most suitable model, whether this yields a significant improvement, and whether these models can be optimized even further by assigning other ratios.

Table 6.2: Best performing model on train data per domain

Domain	RF	AF	1A	1B	2A	2B	3A	3B	Combi A	Combi B
CO	18.14	*17.10*	17.12	17.24	17.24	17.57	**16.96**	17.10	17.25	17.69
OD	27.21	*20.14*	20.42	**19.06**	20.54	20.28	20.20	20.09	21.06	19.99
PO	21.45	*14.91*	14.89	14.39	14.58	13.95	14.88	14.88	14.83	**13.83**
VR	18.32	*16.75*	15.67	16.03	14.63	15.63	16.28	16.47	**14.36**	15.39

Contacts Offered

For CO it appears that model 3A, with a MdAPE of 16.96 is the best-performing model based on the Non-Covid training data set (N = 9,478). Model 3A focuses on absolute adjustments and uses the average of the RF and the AF. The difference in MdAPE from the current AF (17.1) is significant, V = 1,814,096, p-value < 0.001***, $r = 0.05$. However, the effect size is smaller than the threshold value and therefore negligible. This means that the overall effect of model 3A, across all adjustments in CO, is negligible. Nevertheless, for the specific adjustment type addressed in this model, the absolute adjustments (N = 332), it does provide an improvement. For these adjustments, the median %p difference in APE between the current AF and the one of model 3A is -4.14%p. In order to determine whether the performance of model 3A can be improved, different weights have been assigned to the RF and the AF. It appears that this does not improve the performance of model 3A, and therefore the equal weighted model is applied to the Non-Covid test data set (N = 3,840). In the test data the same effect is seen. The difference in MdAPE for the AF (20.38) and model 3A (20.34) is not significant, which means that overall in CO there is no significant advantage. However, specifically for the absolute adjustments (N = 80) in CO, there is a median %p difference in APE of -2.8%p relative to the current AF by applying model 3A. In conclusion, the application of model 3A leads to a lower error for the absolute adjustments, however, the influence of this improvement is not large enough to cause a significant improvement in this domain at the total level.

Orders Delivered

For OD, based on the Non-Covid training data (N = 6,404), model 1B is the best performing model with a MdAPE of 19.06. This model focuses on subsequent upward adjustments using the dampening method where the AF is averaged with the AFM. The difference with the current AF (20.14) is significant, V = 6,380,957, p-value < 0.001***, $r = 0.35$. The median %p difference in APE for the adjustments incorporated in this model, the subsequent upward adjustments (N = 2,642), is -3.13%p relative to the current AF. To see whether the performance of model 1B can be further improved, different weights have been assigned to the AF and the AFM. Based on the performance in the training data, a ratio of 0.0/1.0 for the AF and the AFM, respectively, results in the best outcome. This means that subsequent absolute adjustments are eliminated completely. In the training data, model 1B with a ratio of 0.0/1.0 results in an MdAPE of 18.12. The difference with the AF (20.14) is significant, V = 6,130,926, p-value < 0.001***, $r = 0.31$. The median %p difference in APE for these adjustments is -5.71%p. It is examined how this model performs under various timing scenarios for adjustments, by plotting the MdAPE of this new model next to the RF and AF over the 1 year and the 13-week ahead time horizon, in Figure 6.5 and Figure 6.6, respectively. It can be seen that irrespective of the timing of an adjustment, model 1B with a ratio of 0.0/1.0 has a better performance than the current AF.

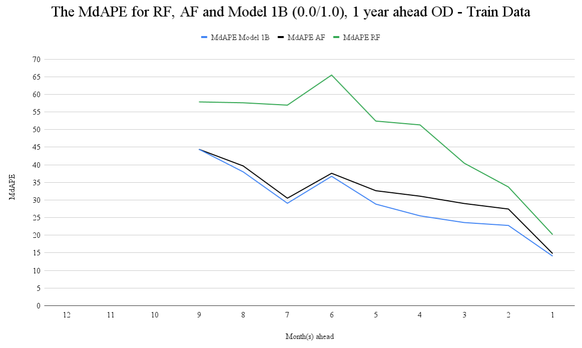

Figure 6.5: Model 1B, 1 year ahead OD - Train

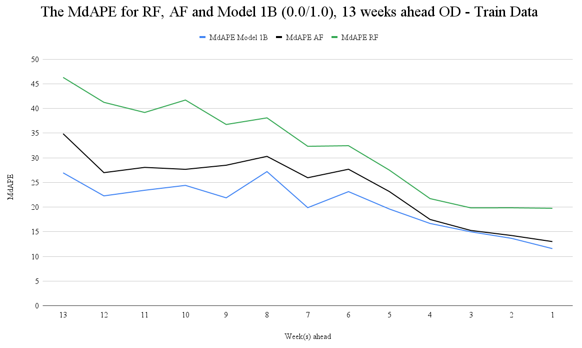

Figure 6.6: Model 1B, 13 weeks ahead OD - Train

Applying model 1B with the 0.0/1.0 ratio to the Non-Covid test data set (N = 5,193) results in a MdAPE of 17.22. The difference with the current AF in the test data set (20.98) is significant, V = 4,693,818, p-value < 0.001***, $r = 0.37$. With this 0.0/1.0 ratio, the median %p difference in APE relative to the current AF is -7.18%p for the subsequent absolute adjustments (N = 2,511). For the test data it was also investigated how this model performs under various timing scenarios for adjustments. In Figure 6.7 and Figure 6.8, it can be seen, for the one-year time horizon and the 13-week time horizon respectively, that model 1B with ratio 0.0/1.0 outperformed the current AF irrespective of the timing of adjustments. It can be concluded that in OD, forecasting performance is best when subsequent upward adjustments are eliminated.

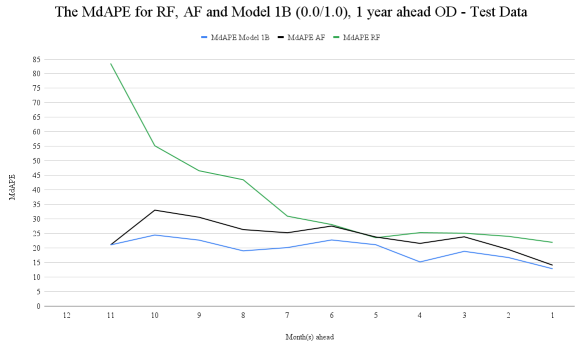

Figure 6.7: Model 1B, 1 year ahead OD - Test

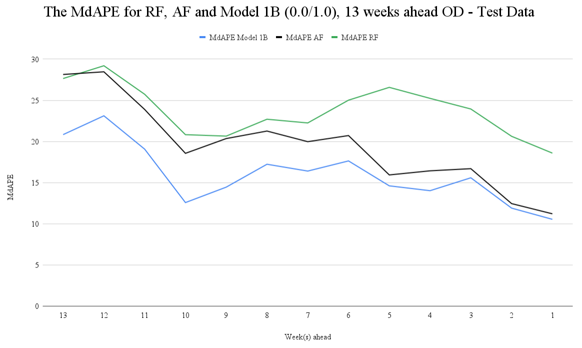

Figure 6.8: Model 1B, 13 weeks ahead OD - Test

Products Ordered

For PO, based on the Non-Covid training data (N = 22,972), CombiModel B is the best performing model with a MdAPE of 13.83. This model takes into account the subsequent upward, subsequent large and absolute adjustments, and is using improvement strategy B, where the AF and the AFM are averaged to dampen the generally harmful adjustments. The difference with the current AF (14.91) is significant, V = 92,937,455, p-value < 0.001***, $r = 0.17$. The median %p difference in APE for the adjustments incorporated in this model (N = 14,925), is -1.5%p relative to the current AF. In order to determine whether the performance of CombiModel B can be improved, different weights have been assigned to the AF and the AFM. It appears that this does not improve the performance of CombiModel B. In addition, the performance of CombiModel B is checked over time. It turns out that, regardless of the timing of the adjustments, CombiModel B outperforms the current AF, see Figure 6.9 and Figure 6.10 for the 1 year and 13 weeks time horizon, respectively.

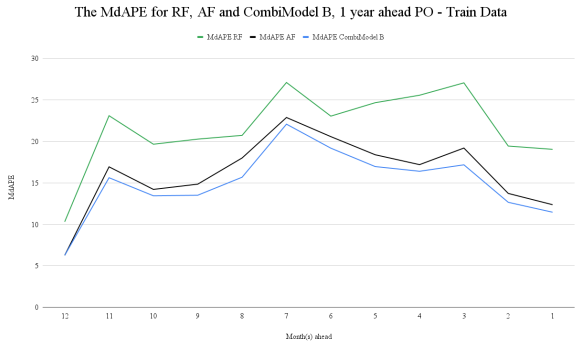

Figure 6.9: CombiModel B, 1 year PO - Train

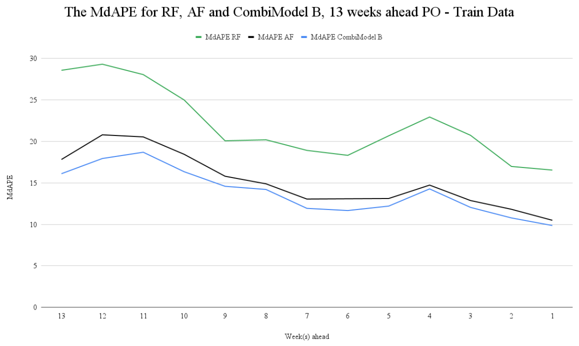

Figure 6.10: CombiModel B, 13 weeks PO - Train

The equal weighted model is applied to the Non-Covid test data set (N = 8,317). Here, it is seen that the difference in MdAPE for the AF (14.96) and CombiModel B (14.46) is significant, V = 11,818,853, p-value < 0.001***, $r = 0.09$. However, the effect size is smaller than the threshold value and therefore negligible. That means that overall in PO there is no significant advantage by applying CombiModel B, which can also be seen when looking at the performance over time, see Figure 6.11 and Figure 6.12. However, for those adjustments incorporated in this model, subsequent upward, subsequent large and absolute adjustments (N = 5,679) there is a median %p difference in APE of -1.18%p relative to the current AF. In conclusion, the application of CombiModel B leads to a lower error for the specific adjustments addressed in this model, however, the influence of this improvement is not large enough to cause a significant improvement in this domain at the total level.

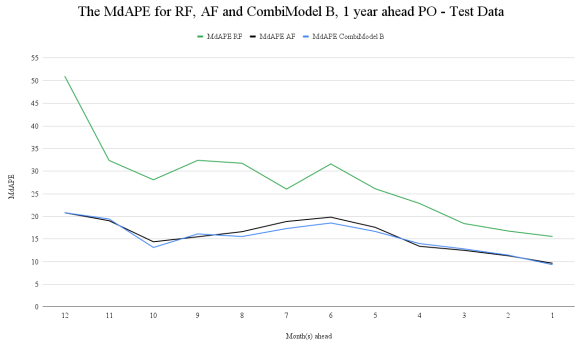

Figure 6.11: CombiModel B, 1 year PO - Test

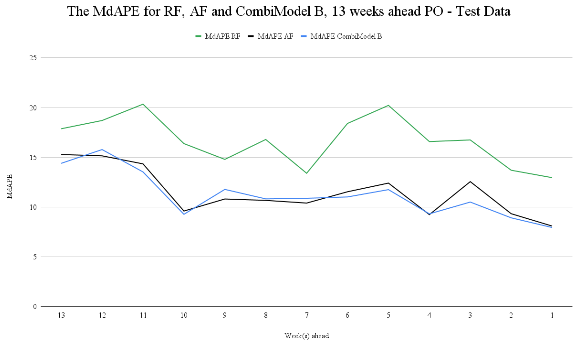

Figure 6.12: CombiModel B, 13 weeks PO - Test

Visitors Registered

For VR, based on the Non-Covid training data (N = 30,664), CombiModel A is the best performing model with a MdAPE of 14.36. This model takes into account the subsequent upward, subsequent large and absolute adjustments, and is using the combination of the RF and the AF as improvement strategy. The difference with the current AF (16.75) is significant, V = 134,105,681, p-value < 0.001***, r = 0.28. The median %p difference in APE for the adjustments incorporated in this model (N = 16,202) is -5.27%p relative to the current AF. In order to determine whether the performance of CombiModel A can be improved, different weights have been assigned to the RF and the AF. Based on the performance in the training data, a ratio of 0.6/0.4 for the RF and the AF, respectively, results in the best outcome. In the training data, CombiModel A with a ratio of 0.6/0.4 results in an MdAPE of 14.34. The difference with the AF (16.75) is significant, V = 1.29e+08, p-value < 0.001***, r = 0.24. The median %p difference in APE for these adjustments is -5.41%p. It is examined how this model performs under various timing scenarios for adjustments. In Figure 6.13 and Figure 6.14, it can be seen that CombiModel A (0.6/0.4) is outperforming the AF irrespective of the timing of adjustments.

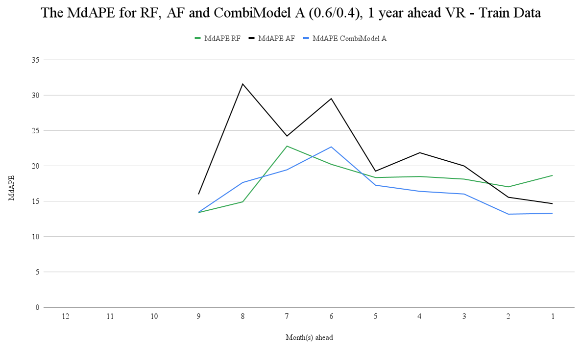

Figure 6.13: CombiModel A, 1 year VR - Train

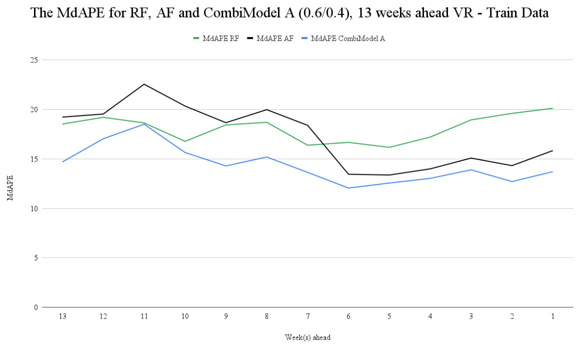

Figure 6.14: CombiModel A, 13 weeks VR - Train

CombiModel A with a 0.6/0.4 ratio for the RF and AF respectively, is applied to the Non-Covid test data set (N = 18,078). It turns out that the difference in MdAPE for the AF (16.44) and CombiModel A (13.95) is significant, V = 51,232,454, p-value < 0.001***, r = 0.26. With this 0.6/0.4 ratio, the median %p difference in APE relative to the current AF is -4.48%p for the adjustments incorporated in this model (N = 10,510). For the test data it was also investigated how this model performs under various timing scenarios for adjustments. In Figure 6.15 and Figure 6.16, it can be seen, for the one-year time horizon and the 13-week time horizon respectively, that the application of this model is particularly beneficial for long-term adjustments.

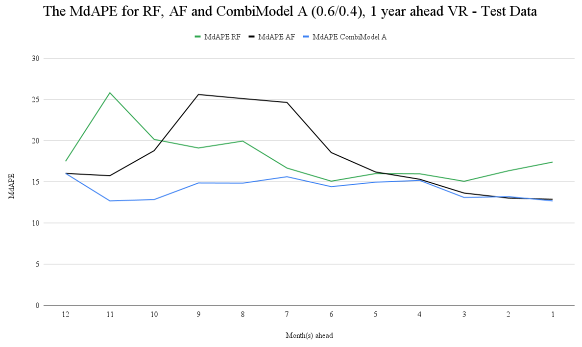

Figure 6.15: CombiModel A, 1 year VR - Test

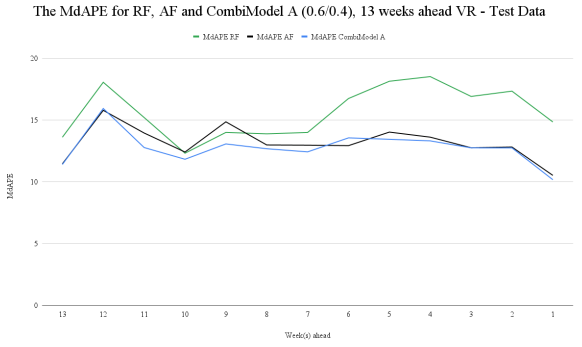

Figure 6.16: CombiModel A, 13 weeks VR - Test

6.4 Main Findings

The simulation results show that, when the entire dataset is considered, CombiModel A, which is targeted at subsequent upward, subsequent large, and absolute adjustments, with a weight of 0.4 for the RF and a weight of 0.6 for the AF, results in de best performance. A median %p decrease in forecast error of -1.22%p is obtained for the adjustment types addressed in this model. The greatest benefit of this model is obtained from the long-term adjustments. Which model performs best differs per domain. For CO, model 3A, which focuses on absolute adjustments, using equal weights for the RF and the AF results in the best performance. A median decrease of -2.8%p in forecast error is obtained for the absolute adjustments, however, on total level for this domain this model does not result in a significant improvement of the forecast accuracy. For OD, model 1B with a ratio of 0.0/1.0, which means subsequent upward adjustments are completely eliminated, results in a significant improvement of the forecast accuracy. A median %p decrease in error of -7.18%p is obtained for the forecasts concerned. For PO, CombiModel B, which takes into account the subsequent upward, subsequent large and absolute adjustments, using equal weights for the AF and the AFM, results in the best performance. Overall there is no significant improvement in terms of error reduction, however, for the adjustments incorporated in the model, a median %p decrease in forecast error of -1.18%p is obtained. Finally, for VR, a significant reduction of error is obtained by applying CombiModel A, targeting at subsequent upward, subsequent large and absolute adjustments and assigning a weight of 0.6 to the RF and a weight of 0.4 to the AF. For these adjustments this results in a median decrease in the forecast error of -4.48%p and this model is particularly beneficial for long-term adjustments.

7. Conclusion

In this section, answers will be given to the two research questions formulated in Section 1.4. First, in Section 7.1, the behavioral tendencies of analysts when making judgmental adjustments to forecasts on the expected workload for the operational domains are described. Second, in Section 7.2, it is described how the insights into the behavioral tendencies of analysts can be used to improve the operational forecasting procedure.

7.1 Behavioral Tendencies of Analysts

In this study, the strengths and weaknesses of analysts in making judgmental adjustments to operationally oriented forecasts are investigated with regard to the characteristics of adjustments such as the direction, size, and timing. It has been examined whether, when and how much certain types of adjustments increase or decrease the accuracy of operationally oriented forecasts. For this purpose, a data analysis was performed of all adjustments made over a time horizon of one and a half year for four different operational domains. A distinction has been made between Covid and Non-Covid data, and conclusions have been drawn based on the Non-Covid data.

First, the results reveal that analysts generally reduce the forecast error of operationally oriented forecasts relative to the raw forecast by making judgmental adjustments. Judgmental adjustments that were related to Covid had a greater positive impact than adjustments that were not related to Covid. Furthermore, exploratory research shows that in approximately half of the cases a subsequent adjustment reduces the forecast error. In addition, what was striking in terms of the types of adjustments is that absolute adjustments generally increased the forecast error, in contrast to percentage and override adjustments. The finding that judgmental adjustments contribute to an improved forecast accuracy is robust across domains.

In general, analysts proved competent in making adjustments in the right direction. In addition, analysts were more capable of choosing the right direction for a first adjustment to the pure raw forecast, than for subsequent adjustments. Moreover, closer to the forecast date, adjustments were made relatively more often in the right direction. Finally, the finding that in general the majority of adjustments is adjusted in the right direction, is consistent across domains.

Analysts made adjustments slightly more often downward than upward, however, this difference was not significant. In terms of the performance of upward and downward adjustments, it appeared that downward adjustments, in general, are more likely to decrease the forecasting error than upward adjustments. Subsequent upward adjustments generally increased the forecast error, and they were especially detrimental for OD and VR. The tendency for downward adjustments to be more inclined to lower the forecast error than upward adjustments is robust across the domains.

In terms of the performance of relatively large and small adjustments, it appeared that for first adjustments, relatively large adjustments are more likely to decrease the forecast error, and that for subsequent adjustments, relatively small adjustments are more likely to decrease the forecast error. Relatively large subsequent adjustments, in general, increased the forecast error, and this especially applies to VR. With regard to the combination of large, small, upward and downward, for first adjustments, small upward adjustments were the least likely to be beneficial, and especially large downward adjustments were useful in reducing the forecast error. For subsequent adjustments, small downward adjustments were, in general, the most advantageous, and in particular large upward adjustments were generally detrimental.

Finally, in terms of the timing of adjustments, it appeared that, in general, the forecast error of the adjusted forecast is decreasing with a decreasing time horizon. That means, the forecast error decreased as the forecast date got closer. As of 6 weeks before the forecast date, analysts were able to significantly improve the forecast accuracy relative to the raw forecast. In addition, it appeared that the forecast accuracy improvement relative to the raw forecast increased with a decreasing time horizon. Lastly, subsequent adjustments made within 6 to 3 weeks ahead generally lead to an improvement of the forecast, where in the final two weeks, the forecast is generally not improved anymore by subsequent adjustments.

In conclusion, by making judgmental adjustments, analysts contribute to a better accuracy of the operationally-oriented forecasts. They do this mainly with downward adjustments, depending on whether it is the first or a subsequent adjustment with large and small adjustments respectively, and mainly in the short term, i.e. from about 6 weeks before the forecast date. Analysts generally had a worse performance with absolute adjustments, subsequent upward, and subsequent large adjustments.

7.2 Improvements Adjustment Procedure

The second part of the study was focused on designing an adjustment procedure that takes into account the behavioral tendencies of the analysts to improve the accuracy of the operationally oriented forecasts. It can be concluded that, regarding the entire data set, the most effective strategy is to apply a model that is targeted at both subsequent upward, subsequent large, and absolute adjustments, where a weight of 0.4 is assigned to the raw forecast and a weight of 0.6 is assigned to the adjusted forecast. This results in a median %p decrease in forecast error for these adjustments of -1.22%p. This strategy is most beneficial for long-term adjustments.

The most effective procedure for adjusting forecasts differs per domain. For CO it appears that taking the average of the raw forecast and the adjusted forecast, in case of absolute adjustments, is most effective and results in a median %p decrease in forecast error of -2.8%p for these adjustments. However, the impact of lowering the forecast error after absolute adjustments is not powerful enough to gain an advantage in this domain at the total level. For OD, it appears that forecasting performance is best when subsequent upward adjustments are completely eliminated, regardless of the timing of these adjustments. By eliminating these adjustments there is a median %p decrease in forecast error of -7.18%p for the forecasts concerned.

For PO, damping both subsequent upward, subsequent large and absolute adjustments by assigning a weight of 0.5 to these adjustments results in the best performance. For these adjustments, a median %p decrease in forecast error of -1.18%p can be achieved, however, the impact of this improvement is not large enough for a significant improvement on the total level in this domain. Finally, for VR, a significant reduction of the error can be obtained by assigning a weight of 0.6 to the raw forecast and a weight of 0.4 to the adjusted forecast for those forecasts for which a subsequent upward, subsequent large and/or an absolute adjustment has been made. For these adjustments this results in a median decrease in the forecast error of -4.48%p and applying this strategy is particularly beneficial for long-term adjustments.

8. Discussion

In this chapter, the academic and practical implications of the research are discussed in Section 8.1. In addition, the limitations of the research will be highlighted in Section 8.2. Finally, recommendations will be given, and opportunities for further research will be presented in Sections 8.3 and 8.4, respectively.

8.1 Implications

8.1.1 Academic Implications

From an academic perspective, the purpose of this research was to contribute to the existing literature on judgmental forecasting. More specifically, the aim was to see whether the same characteristics of judgmental adjustments play a role when forecasting, and what the effects of these adjustments are on the accuracy of operationally oriented forecasts which are made and adjusted at an aggregated level. This, in contrast to most existing literature on judgmental adjustments, which is usually committed to sales forecasting at a highly disaggregated level. Therefore, the results of this research will be discussed, and there is reflected on the similarities and differences with, and the additions to the existing literature in this section.

Firstly, the results of this study, regarding the characteristics of judgmental adjustments and the effect of these adjustments on the accuracy of operationally oriented forecasts, are largely in line with what is seen in the sales forecasting context. Consistent with the sales forecasting literature, forecasters in this context also proved capable of making adjustments in the right direction, downward adjustments were, in general, more likely to decrease the forecasting error than upward adjustments, and when adjustments were made to the pure raw forecast, relatively large adjustments were more likely to decrease the forecast error than relatively small adjustments, especially the large downward adjustments. In contrast to the sales forecasting literature, where it was generally seen that adjustments were adjusted upward more frequently than downward, this study found an equal ratio of upward and downward adjustments. In addition, even though not much research has been done on the timing of adjustments and the effect of this on forecast accuracy, the results of this study are different from what has been seen in the literature so far. In this study, the performance of adjustments increased when they were made closer to the forecast date, whereas it was seen in the literature that the performance of adjustments deteriorated closer to the forecast date. The above-mentioned similarities and differences with the existing literature will be discussed in greater detail later in this section.

A topic on which only minimal research has been done is the characteristics and effects of subsequent adjustments. This research shows that in approximately half of the cases subsequent adjustments increase the forecast accuracy. Furthermore, it was seen that analysts chose the right direction relatively less often for subsequent adjustments than for first adjustments. Assuming that it is more difficult to determine the direction of an adjustment when the forecast is closer to the target, this may be a reason that subsequent adjustments were made less often in the right direction than adjustments to the pure raw forecast. Moreover, whether subsequent adjustments generally increase or decrease the forecast accuracy is highly dependent on the characteristics of such subsequent adjustments. Subsequent upward adjustments generally increased the forecast error significantly. A potential explanation for this is the possible asymmetric loss function and/or an optimism bias, causing upward adjustments, in general, being excessively high or wrongly directed, as can also be seen in the study by Fildes et al. (2009) and Syntetos et al. (2009). Another possibility, more focused on the fact that it concerns *subsequent* upward adjustments, is that the concept brought up by Fildes et al. (2009) of the confusion between a forecast and a decision plays a role here. The idea of this concept is that instead of the pure *forecast* (an estimation of what will happen in the future), a *decision* (an estimate that should be acted upon in order to minimize loss) is already made by setting a forecast at a higher level in order to avoid a shortage of employees. The risky thing about this situation, however, is that the decisions, what they actually are, but labeled as 'forecasts', are subject to misinterpretation, and that they could be adjusted even further upwards later on, i.e. when making a subsequent adjustment. Finally, the size of subsequent adjustments was generally smaller than that of first adjustments, and it turned out that the relatively large subsequent adjustments generally increased the forecast error. This applies in particular to subsequent large upward adjustments. This may again be due to a possible optimism bias among analysts causing them to make over-optimistic adjustments upwards.

Another addition to the existing literature is the difference in performance that was found for adjustments that were directly linked to a period with disruptions, the COVID-19 pandemic, and the adjustments that were not directly linked to disruptions. It appeared that the analysts had a greater positive impact on operational performance with disruption-related adjustments. A possible reason for this is that, compared to the algorithm, analysts had more information at their disposal about for example new measures. Additionally, they could respond more quickly to this new information since analysts can estimate the impact of these measures, but due to a lack of historical data, this is much more difficult to overcome with statistical techniques, as also argued by Fildes and Goodwin (2007).

Furthermore, one of the exploratory topics has focused on different types of adjustments, namely percentage, absolute, and override adjustments. The results show that there is a difference in effectiveness between the different types of adjustments. While percentage and override adjustments were generally effective in reducing the forecast error, absolute adjustments were generally detrimental to forecast accuracy. However, it is possible that the causes of specific types of adjustments — rather than the way in which they are made — are what ultimately contribute to the effects of these adjustments. Absolute adjustments are made, for example, when a new product is released or when an increase in the forecast is needed due to warm weather. Rather than the type of adjustment, it could be that the reasons for such adjustments are the cause of the generally negative effect of these adjustments.

In the following sections, the findings of this study will be further discussed in comparison with the existing literature. Firstly, an important topic of debate in the literature is whether judgmental adjustments to statistically generated forecasts are at all beneficial for forecast accuracy. In this research, the adjustments made by analysts, based on contextual information, appear to be generally beneficial for forecast accuracy, which is in line with the results of, for example, Fildes et al. (2009) and Syntetos et al. (2009). In addition, similar to previous studies, such as those of Mathews and Diamantopoulous (1990) and Petropoulos et al. (2016), it was found that the majority of judgmental adjustments was made in the right direction. It appeared that adjustments were made relatively more often in the right direction closer to the forecast date. A possible explanation for this, as suggested by Van den Broeke et al. (2019), is that near the forecast date, more contextual information becomes available, which can be utilized by analysts to decide on the direction of an adjustment.

Consistent with the results of, for example, Fildes et al. (2009), Baecke et al. (2017) and Van den Broeke et al. (2019), it was found that relatively large adjustments were more likely to decrease the forecast error than relatively small adjustments. However, this only applies if it was the first adjustment for a particular key and date. Especially the large downward adjustments were useful in reducing the forecast error, which was also seen in the study of Syntetos et al. (2009). A possible reason for this is that both downward and large adjustments are often based on reliable information (Baecke et al., 2017; Fildes et al., 2009). For first adjustments, small upward adjustments were the least likely to be beneficial. It could be that analysts make these small upward adjustments because of, for example, an illusion of control effect (Kottemann et al., 1994) or tinkering with the data (Fildes et al., 2009), instead of having really good reasons to do so. However, no conclusions can be drawn about this.

Regarding the timing of adjustments, it was found that, contrary to expectations, the forecast error of the adjusted forecast decreased with a decreasing time horizon, and that as of 6 weeks before the forecast date, analysts were able to significantly improve the forecast accuracy relative to the raw forecast. It could be that analysts, instead of overreacting to new information, are actually making effective use of this recent contextual information, as expected by Van den Broeke et al. (2019). From 6 weeks in advance, analysts probably had access to relevant contextual knowledge that the algorithm did not have. Striking is that the forecast error of the adjusted forecast increased in the last days before a forecast date. It could be that analysts do overreact to new information these last days.

Finally, the averaging method as introduced by Blattberg and Hoch (1990), using the complementary strengths and weaknesses of analysts and algorithm, as well as the dampening method as proposed by Fildes et al. (2009), and assigning different weights to forecasts, were found to be effective in improving the forecasting accuracy of the operationally oriented forecasts. This indicates that dampening, using the complementary strengths and weaknesses of analysts and algorithm, or even completely eliminating adjustments that are generally biased, can, according to this study, improve the forecast accuracy.

8.1.2 Practical Implications

From a practical perspective, the purpose of this research was to gain insight into the effects of the judgmental adjustments that analysts make to the algorithm-generated forecasts. More specifically, the goal was to gain insight into which adjustments to algorithm-generated forecasts, in terms of the direction, size and timing, increase or decrease the accuracy of the operationally oriented forecasts. The ultimate goal was to propose an improved forecasting procedure that ensures a lower forecast error and time savings. Concrete insights regarding its operationally oriented forecasting process are obtained for Coolblue, which will be discussed in this section.

Firstly, one of the most important insights is that analysts generally reduce the forecast error of the operationally oriented forecasts by making judgmental adjustments. However, the results show that, in terms of adjustment characteristics, some adjustments are effective in reducing the forecast error whereas other adjustments are detrimental for the forecast accuracy. Simulations with new adjustment procedures, showed that a more effective forecasting procedure could be realized by dampening and sometimes even completely eliminating generally detrimental adjustments, and utilize those adjustments which generally added value in terms of error reduction. The results show that the error of the operationally oriented forecasts can be decreased, and by eliminating specific types of adjustments, time spent on making adjustments could be reduced. The time saved can be used for other value-adding tasks. The reduced forecast error and therefore more accurate forecasts of the workload for the operational domains, enables more efficient employee scheduling. More efficient employee scheduling, in turn, can ensure that 1) not too few employees are scheduled, and 2) not too many employees are scheduled. Not scheduling too few employees, could have a positive effect on customer satisfaction, since scheduling enough employees will contribute to, for example, on-time delivery of products, being able to assist customers in the stores, and answer customer questions without too long waiting times. Not scheduling too many employees, could lead to cost reduction due to FTE (Full-Time Equivalent) savings. It would be interesting to be able to calculate what a certain error reduction can theoretically result in, in terms of cost savings, however, there is currently no clear translation available of the amount of error reduction in the operational forecasts to the number of FTEs that can be saved with this. This could be an interesting topic for future research. It should be taken into account, that planners can deviate from the forecast when deciding how many employees they eventually schedule. It is therefore important to find out whether the planners in the operational domains base their planning purely on the forecast as input or whether they still use interpretation and contextual knowledge to determine how many employees they are going to schedule. In other words, the correlation between the forecast and the number of employees scheduled must be checked. Only when this is known, and when this correlation is high, it can be said whether a reduction in the forecast error would actually lead certain cost savings in terms of FTE's.

It was seen that in some domains an improvement was achieved by applying the improvement models irrespective of the timing of adjustments. However, in other domains it was mainly the long-term adjustments that benefited from applying the models. These long-term adjustments and the resulting forecasts are focused on long-term capacity planning. This implies that the new models in these domains, that achieve higher forecast accuracy, can ensure timely action, i.e. making hiring and firing decisions, to ensure that capacity is at the right level for the future workload in the operational domains.

Furthermore, as mentioned, some adjustments are more effective than others in terms of reducing the forecast error. The results show that mainly the subsequent upward and the relatively large subsequent adjustments generally increase the forecast error. Regarding the type of adjustment, absolute adjustments have, in general, a relatively poor performance. These results imply that there is something causing these types of adjustments to generally have poor performance. It needs to be considered how these adjustments come about, i.e. when these types of adjustments are made, and whether the reasons for these kind of adjustments can be justified. For instance, absolute adjustments are often made for new product releases and heatwaves. It must be investigated whether, rather than the way of adjusting, these reasons for adjustment could underlie a deterioration in forecasting performance. Another example, subsequent upward adjustments are often made based on input from commercial teams. It should be considered whether this input is reliable enough, and/or whether analysts are not overreacting based on this input.

Moreover, there is an important practical implication with regard to the timing of adjustments. The results show that as of 6 weeks before any forecast date, analysts were able to significantly improve the forecast accuracy relative to the raw forecast. This implies that analysts are likely to have contextual information from that time onward, that will enable them to achieve this better performance. By knowing this, time and resources can be used in such a way that there is more focus on the adjustments in this time period and less in the time period before, in which apparently not enough contextual information is yet available for the analysts to actually achieve a significant improvement relative to the raw forecast.

The ultimate goal of this research was to increase the accuracy of Coolblue's operationally oriented forecasts and/or to reduce the time spent on making adjustments. The results show that this is possible by making effective use of the interaction between algorithm and analysts.

8.2 Limitations

8.2.1 Academic Limitations

In most of the existing literature, adjustments are made to a pure statistical forecast. As a result, a statement can be made about the performance of these adjustments based on a comparison with this statistical forecast. In this study, however, the first adjustment is made to the statistical forecast, but then, using the statistical forecast as a reference point, adjustments are made to a forecast which consists of a statistical forecast with all previous adjustments processed in it. This makes a direct comparison of the findings with regard to the performance of various types of adjustments in this study with those of existing literature more difficult.

Furthermore, in this study, the analysis is focused on adjustments made by five data analysts. This is a relatively small sample size and is therefore not representative for forecasting analysts in general. Additionally, all five analysts are responsible for (a part of) the forecasts of an operational domain. It could therefore be that the differences seen in terms of the characteristics and performance of adjustments in the various domains are not due to the differences in the domains, but to the personal behavioral tendencies of analysts when adjusting.

Moreover, in this study, a distinction is made between Covid-related adjustments and Non-Covid-related adjustments, in order to investigate whether there is a difference in the characteristics and performance of adjustments that are directly linked to a period with disruptions, and the adjustments that are not directly linked to disruptions. However, assigning the label Covid and Non-Covid is done on the basis of the category and justification of an adjustment.

This category and justification are string data fields and it is possible that there are adjustments that were related to Covid but did not have a clear Covid category or justification or that did not end up in the Covid data set due to many ways of spelling. In addition, the sample size of Covid-related adjustments was relatively small compared to the sample size of adjustments that were not Covid related, making it more difficult to compare the results.

In addition, in this study, the Median Absolute Percentage Error (MdAPE) was used as an error measure due to a high number of outliers in the data set. This error measure can generate valuable insights into a data set with a relatively high number of outliers. However, a disadvantage of this error measure, in contrast to the Mean Absolute Percentage Error (MAPE), is that it does not take all precise values of observations into account but only the most central value, and thus does not use all available data in the data set.

Another limitation of this study is the time-based data split made for the train and test data set. Although this is the most suitable split to perform the data analysis, since exactly the same time periods are analyzed, it is not ideal for the simulation models. Because, in theory, and what was also seen, more adjustments can be made for certain periods of time. Such a split then results in an unequal distribution of data. This could affect the training and testing of the models and thus the results in terms of performance of such models.

Finally, since this is case study, the results can be used for Coolblue, but generalisation of the results to other companies is not possible.

8.2.2 Practical Limitations

On the practical side, there are some data related limitations. First, the date an adjustment is made is not stored in the database. A workaround was used to obtain time-related information about the adjustments. Applying this workaround has two consequences, first of all, it caused some observations to be deleted. Despite the fact that it only concerns <1% of the data, this can be considered a data loss. Second, the creation date obtained by this workaround may differ by 1 day from the actual creation date. The consequence of this is that the raw forecast can also differ by 1 day, and so can the adjusted forecast morning since its calculated based on the raw forecast. The effect of this is small because it can be assumed that the raw forecast does not change much from one day to the next, however it needs to be mentioned.

In addition, a constraint for the creation date of an adjustment was applied to prevent adjustments that are made during the implementation phase of the new forecasting system to be in the data set. This causes that for the first few months in the data set, not all (long term) adjustments are included. Furthermore, the amount of data available for each domain was significantly different. For example, VR data accounted for nearly half (46.89%) of the total data set. Obviously, this ensures that the effects seen in this domain had a major influence on the effects seen in the total data set.

Finally, a completely realistic simulation for improving the forecast procedure was not possible due to a very dynamic forecasting system. Normally, after an adjustment, the total adjustment made so far for a particular day and key is calculated. This total adjustment is then used to calculate the adjusted forecast morning. However, in this study, the effect of the new models on previous adjustments has not been passed on in this total adjustment, and therefore also not in the adjusted forecast morning. This is not done because it would change the adjusted forecast morning on which analysts base their adjustments, and it cannot be assumed that they would have made the same decisions in terms of adjustments if this number is different. Because this study investigated analyst behavior, there is only focused on the data points that were available to the analysts at the time of adjustment. This leads to a suitable simulation of analyst behavior but not to a one-to-one simulation of what would happen in practice when applying such models.

8.3 Recommendations

First, based on the results of this research, it is important that the data analysts are informed about the insights that have been obtained. By providing them with domain-specific information about the performance of adjustments and the various effects of different types of adjustments, they can be made more aware of adjustment behavior and learn to recognize or overcome possible causes of bias. This can be done by presenting the results, a meeting or discussion. Nonetheless, these results are based on a one-time analysis at a point in time. A second recommendation, in line with this, is therefore that such an analysis should be repeated at more moments, over time, in order to obtain regular feedback on the adjustment performance. Since an R add-on exists for BigQuery, an R script similar to the one generated for this study could be run on a more frequent basis to check performance and get insights. However, a dynamic dashboard can also be built in, for example, Google Sheets, which is more in line with the currently used tools within Coolblue, in which the performance of (different types of) adjustments can be tracked. In addition to this, in line with what has already been mentioned, it is recommended that, based on the insights into the adjustments that have generally proved less effective, i.e. subsequent upward, subsequent large and absolute adjustments, a reconsideration is made as to whether the information sources for those adjustments are reliable (enough). Examples of this are the subsequent upward adjustments in OD, which are often based on input from commercial teams, and the absolute adjustments, which are often made for new product launches or heat waves.

Moreover, the insights about effective and generally ineffective adjustments, as well as the insights into the timing of adjustments, and the insights gained by applying the new models, i.e. the damping of certain adjustments but also the complete elimination of certain adjustments, should be used to allocate time and resources to the type of adjustments that have the biggest positive impact on forecast accuracy. The valuable time this saves can be spent on value adding tasks instead of spending time on adjustments that do not add significant value to the forecast accuracy. Again, this should be done based on up-to-date information, and so the insights need to be updated regularly.

One last note, in order to gain more and easier data insights in the future and to be able to check the performance of adjustments on a regular basis as described above, there are two more data-related recommendations. First, it is recommended to link the data field with the date of adjustment to the tables in the data base. In this way, it will be easier in the future to retrieve information about the timing of adjustments. Secondly, by applying a more structured way of using categories and justifications, insights can be obtained in the future into the relationship between a category and/or justification of an adjustment and the related performance of such adjustments. This can be done, for example, by organizing meetings and reaching a consensus about the categories that should be available and the way in which justifications are utilized. Especially for the category, a drop-down menu could be used instead of string data. Differences in language and spelling can then be prevented, which in turn can make analysis easier and can lead to valuable insights being obtained.

8.4 Future Research

For future research, it could be interesting to investigate how the magnitude of error reduction in the operational forecasts, caused by making judgmental adjustments, can be translated to the number of FTE's that can be saved by more efficient employee scheduling. This can provide an even better picture of the added value of analysts in the forecasting process, and what an effective collaboration of analyst and algorithm can lead to in economic terms.

In addition, the results of this study show that certain types of adjustments are more effective than others. The question that these results have raised is whether this is caused by the specific type of adjustment or by the reasons underlying specific types of adjustments. Therefore, it could be an interesting topic for future research, from both a practical as well as an academic perspective, to investigate whether there is a relationship between the reason for an adjustment and the performance. This could, for example, be done by a text analysis based on the category and the justification of an adjustment and might provide additional information on what causes the performance of particular adjustment types. This could also make clear whether there may be information sources that are not or less reliable than others. In addition, by investigating the relationship between the reason for an adjustment and the performance of particular types of adjustments, it may be possible to gain more knowledge about the sources of bias in particular adjustments and how this could be reduced.

References

Ali, Ö. G., Sayın, S., Van Woensel, T. & Fransoo, J. (2009). Sku demand forecasting in the presence of promotions. *Expert Systems with Applications*, *36*(10), 12340-12348.

Alvarado-Valencia, J., Barrero, L. H., Önkal, D. & Dennerlein, J. T. (2017). Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting*, *33*(1), 298-313.

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners* (Vol. 30). Springer.

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting*, *22*(3), 583-598.

Armstrong, J. S., Green, K. C. & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research*, *68*(8), 1717-1731.

Arvan, M., Fahimnia, B., Reisi, M. & Siemsen, E. (2019). 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 judgement into statistical demand forecasting systems. *International Journal of Production Economics*, *191*, 85-96.

Bendoly, E., Croson, R., Goncalves, P. & Schultz, K. (2010). Bodies of knowledge for research in behavioral operations. *Production and Operations Management*, *19*(4), 434-452.

Blattberg, R. C. & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, *36*(8), 887-899.

Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. (2015). *Time series analysis: forecasting and control* (5th ed.). Hoboken: John Wiley & Sons.

Bunn, D. & Wright, G. (1991). Interaction of judgemental and statistical forecasting methods: issues & analysis. *Management science*, *37*(5), 501–518.

Carbone, R., Andersen, A., Corriveau, Y. & Corson, P. P. (1983). Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment. *Management Science*, *29*(5), 559–566.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum Associates.

Davydenko, A. & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. *International Journal of Forecasting*, *29*(3), 510-522.

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. J., Simmons, J. P. & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Donohue, K., Özer, Ö. & Zheng, Y. (2020). Behavioral operations: Past, present, and future. *Manufacturing & Service Operations Management*, *22*(1), 191–202.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, *8*(1), 81-98.

Fildes, R. & Goodwin, P. (2007). Against your better judgment? how organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*(6), 570–576.

Fildes, R., Goodwin, P. & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, *42*(1), 351-361.

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*, *25*(1), 3-23.

Fildes, R., Goodwin, P. & Önkal, D. (2019). Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting*, *35*(1), 144-156.

Fildes, R., Nikolopoulos, K., Crone, S. F. & Syntetos, A. A. (2008). Forecasting and operational research: a review. *Journal of the Operational Research Society*, *59*(9), 1150-1172.

Fildes, R. & Petropoulos, F. (2015a). Improving forecast quality in practice. *Foresight: The International Journal of Applied Forecasting*, *36*, 5–12.

Fildes, R. & Petropoulos, F. (2015b). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692-1701.

Franses, P. H. (2011). Averaging model forecasts and expert forecasts: Why does it work? *Interfaces*, *41*(2), 177-181.

Franses, P. H. & Legerstee, R. (2009). Properties of expert adjustments on model-based sku-level forecasts. *International Journal of Forecasting*, *25*(1), 35-47.

Franses, P. H. & Legerstee, R. (2011a). Combining sku-level sales forecasts from models and experts. *Expert Systems with Applications*, *38*(3), 2365-2370.

Franses, P. H. & Legerstee, R. (2011b). Experts' adjustment to model-based sku-level forecasts: does the forecast horizon matter? *Journal of the Operational Research Society*, *62*(3), 537–543.

Gardner, E. S. (2006). Exponential smoothing: The state of the art—part ii. *International Journal of Forecasting*, *22*(4), 637-666.

Gardner, E. S. & McKenzie, E. (1985). Forecasting trends in time series. *Management science*, *31*(10), 1237–1246.

Goodwin, P. (1996). Statistical correction of judgmental point forecasts and decisions. *Omega*, *24*(5), 551-559.

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega*, *30*(2), 127-135.

Goodwin, P. & Lawton, R. (1999). On the asymmetry of the symmetric mape. *International journal of forecasting*, *15*(4), 405–408.

Goodwin, P. & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting*, *9*(2), 147-161.

Grossmann, W. & Rinderle-Ma, S. (2015). *Fundamentals of business intelligence.* Springer.

Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Harlow, Essex: Pearson.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes*, *63*(3), 247–263.

Haselton, M., Nettle, D. & Andrews, P. (2005). The evolution of cognitive bias. In *The handbook of revolutionary psychology* (1st ed., p. 724-746). Hoboken: John Wiley & Sons Inc.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5-10. (The paper is a reprinted version of the 1957 report)

Huang, L.-T., Hsieh, I.-C. & Farn, C.-K. (2011). On ordering adjustment policy under rolling forecast in supply chain planning. *Computers & Industrial Engineering*, *60*(3), 397-410.

Hyndman, R. & Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts.

Hyndman, R. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679–688.

Hyndman, R., Koehler, A. B., Ord, J. K. & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach.* Springer Science & Business Media.

Jones, D. R. & Brown, D. (2002). The division of labor between human and computer in the presence of decision support system advice. *Decision Support Systems*, *33*(4), 375-388.

Jones, D. R., Wheeler, P., Appan, R. & Saleem, N. (2006). Understanding and attenuating decision bias in the use of model advice and other relevant information. *Decision Support Systems*, *42*(3), 1917-1930.

Khosrowabadi, N., Hoberg, K. & Imdahl, C. (2022). Evaluating human behaviour in response to ai recommendations for judgemental forecasting. *European Journal of Operational Research*.

Kottemann, J. E., Davis, F. D. & Remus, W. E. (1994). Computer-assisted decision making: Performance, beliefs, and the illusion of control. *Organizational Behavior and Human Decision Processes*, *57*(1), 26–37.

Lawrence, M., Goodwin, P., O'Connor, M. & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting*, *22*(3), 493-518.

Lawrence, M. & O'Connor, M. (1996). Judgement or models: The importance of task differences. *Omega*, *24*(3), 245-254.

Lawrence, M. & O'Connor, M. (2005). Judgmental forecasting in the presence of loss functions. *International Journal of Forecasting*, *21*(1), 3-14.

Lawrence, M., O'Connor, M. & Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research*, *122*(1), 151-160.

Leitner, J. & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information – a review of the literature. *European Journal of Operational Research*, *213*(3), 459-469.

Makridakis, S. & Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451-476.

Mathews, B. P. & Diamantopoulos, A. (1986). Managerial intervention in forecasting. an empirical investigation of forecast manipulation. *International Journal of Research in Marketing*, *3*(1), 3-10.

Mathews, B. P. & Diamantopoulous, A. (1990). Judgemental revision of sales forecasts: Effectiveness of forecast selection. *Journal of Forecasting*, *9*(4), 407–415.

McNees, S. K. (1990). Man vs. model? the role of judgment in forecasting. *New England Economic Review*(Jul), 41–52.

Mitchell, T. M. (1997). *Machine learning.* McGraw-hill New York.

Montgomery, D. C. & Runger, G. C. (2018). *Applied statistics and probability for engineers* (7th ed.). John Wiley & Sons.

Moon, M. A., Mentzer, J. T. & Smith, C. D. (2003). Conducting a sales forecasting audit. *International Journal of Forecasting*, *19*(1), 5-25.

Nahmias, S. & Olsen, T. L. (2015). *Production and operations analysis.* Waveland Press.

O'Connor, M., Remus, W. & Griggs, K. (1993). Judgemental forecasting in times of change. *International Journal of Forecasting*, *9*(2), 163–172.

Perera, H. N., Hurley, J., Fahimnia, B. & Reisi, M. (2019). The human factor in supply chain forecasting: A systematic review. *European Journal of Operational Research*, *274*(2), 574-600.

Petropoulos, F., Fildes, R. & Goodwin, P. (2016). Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research*, *249*(3), 842-852.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K. & Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management*, *60*, 34-46.

Petropoulos, F., Makridakis, S., Assimakopoulos, V. & Nikolopoulos, K. (2014). 'horses for courses' in demand forecasting. *European Journal of Operational Research*, *237*(1), 152-163.

Sagaert, Y. R., Aghezzaf, E.-H., Kourentzes, N. & Desmet, B. (2018). Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research*, *264*(2), 558-569.

Sanders, N. R. (1992). Accuracy of judgmental forecasts: A comparison. *Omega*, *20*(3), 353-364.

Sanders, N. R. & Manrodt, K. B. (1994). Forecasting practices in us corporations: survey results. *Interfaces*, *24*(2), 92–100.

Sanders, N. R. & Manrodt, K. B. (2003). Forecasting software in practice: Use, satisfaction, and performance. *Interfaces*, *33*(5), 90–93.

Sanders, N. R. & Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting. *Journal of Behavioral Decision Making*, *5*(1), 39–52.

Sanders, N. R. & Ritzman, L. P. (2001). Judgmental adjustment of statistical forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (p. 405-416). Boston, MA: Springer US.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A. & Kamaev, V. A. (2013). A survey of forecast error measures. *World applied sciences journal*, *24*(24), 171–176.

Syntetos, A. A., Kholidasari, I. & Naim, M. M. (2016). The effects of integrating management judgement into out levels: In or out of context? *European Journal of Operational Research*, *249*(3), 853-863.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R. & Goodwin, P. (2009). The effects of integrating management judgement into intermittent demand forecasts. *International Journal of Production Economics*, *118*(1), 72-81.

Tong, J., Feiler, D. & Larrick, R. (2018). A behavioral remedy for the censorship bias. *Production and Operations Management*, *27*(4), 624–643.

Trapero, J. R., Pedregal, D. J., Fildes, R. & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting*, *29*(2), 234-243.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van den Broeke, M., De Baets, S., Vereecke, A., Baecke, P. & Vanderheyden, K. (2019). Judgmental forecast adjustments over different time horizons. *Omega*, *87*, 34-45.

Webby, R. & O'Connor, M. (1996). Judgemental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting*, *12*(1), 91-118.

Webby, R., O'Connor, M. & Lawrence, M. (2001). Judgmental time-series forecasting using domain knowledge. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 389–403). Boston, MA: Springer US.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*, *6*(3), 324–342.

A. Statistical Tests

In this section, the statistical tests that are used for hypothesis testing are discussed.

A.1 Wilcoxon signed-rank test

To compare the medians of two related samples, the paired samples Wilcoxon signed-rank test can be used as a non-parametric alternative to the paired-samples t-test (Montgomery & Runger, 2018). In this case, the null hypothesis states that the median difference between pairs of data is zero. The null hypotheses (H_0) and the alternative hypotheses (H_1) are:

$H_0 : m = 0$ $H_1 : m \neq 0$ (two-sided)

$H_0 : m \leq 0$ $H_1 : m \ > 0$ (one-sided)

$H_0 : m \geq 0$ $H_1 : m \ < 0$ (one-sided)

The Wilcoxon signed-rank test takes into account the plus and minus signs of the differences between observations of two paired samples, as well as the size of these differences. First, the differences are computed. After that, the absolute differences are ranked ascending. Then, the sign of the related difference is assigned to the rankings. Finally, the negative and positive ranks are summed and the minimum of those two is picked as the test statistic (Montgomery & Runger, 2018).

$$w = \min\left(w^{+}, w^{-}\right) \tag{A.1}$$

Where,
w^{+} is the sum of the positive ranks
w^{-} is the absolute value of the sum of the negative ranks

If the p-value that results from this test is less than 0.05, it is considered statistically significant. In that case, the null hypothesis can be rejected and the alternative hypothesis can be accepted. For the two-sample test, it means that there is a significant difference between the medians of the two samples. For the Wilcoxon signed-rank test on paired samples, the differences between the two samples should be distributed approximately symmetrically around the median. In case the assumption is violated for a particular test, a sign test can be performed. The sign test does not assume a symmetrical distribution of the differences.

A.2 Kruskal-Wallis Test

The Kruskal-Wallis rank-sum test is a non-parametric alternative to the one-way ANOVA test and is an extension of the two-sample Wilcoxon test for a more than two groups context. This test can be used to find out whether there are significant differences in medians between the different groups. The null hypotheses (H_0) and the alternative hypotheses (H_1) are:

H_0: The medians across all groups are equal.
H_1: At least one of the medians is different from the others.

The following formula is used to compute the test statistic:

$$H = (N-1)\eta^2 \tag{A.2}$$

Where,
H = distributed as χ^2 with df = groups -1
N = sample size
η^2 = eta-squared

If the p-value generated by this test is less than 0.05, the results are considered statistically significant. In that case, the null hypothesis can be rejected and the alternative hypothesis can be accepted, meaning that at least one of the medians is significantly different from the other medians.

A.2.1 Pairwise Wilcox Test

If the Kruskal-Wallis test is significant, that means that at least one of the medians is significantly different from the other medians. To find out if there are specific groups where there is a significant difference in medians, multiple pairwise comparisons can be performed using the pairwise Wilcoxon rank-sum test. If there is one or more p-value(s) that is/are less than 0.05, that means that for that pair(s), the difference in medians is statistically significant.

A.3 Chi-square goodness-of-fit test

To compare an observed distribution to an expected distribution, for two or more categories in the data, the goodness of fit test can be used. This test is based on the chi-square distribution. A frequency histogram with k bins is used to organize observations from a sample of size n. The observed frequency in the ith bin is denoted by O_i. The expected frequency in the ith bin (E_i), is calculated using the hypothesized probability distribution (Montgomery & Runger, 2018). The null hypothesis (H_0) and the alternative hypothesis (H_1) are:

H_0: There is no significant difference between the observed and the expected distribution.
H_1: There is a significant difference between the observed and the expected distribution.

The following formula is used to compute the test statistic:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{A.3}$$

Where,
O_i = the observed frequency
E_i = the expected frequency

If the p-value that results from this test is less than 0.05, it is considered statistically significant. In that case, the null hypothesis can be rejected and the alternative hypothesis can be accepted, meaning that there is a significant difference between the observed and the expected distribution.

A.4 Effect Sizes

In case the null hypothesis is false for a certain statistical test, it is false to some particular degree. This is called the effect size (ES) (Cohen, 1988). The effect size highlights the importance of a research finding in terms of practical significance. In contrast to a small effect size, which suggests limited practical applicability, a large effect size denotes the practical significance of a research finding. Some test statistics for the effect size will be discussed below.

Wilcoxon Effect Size *r*

The effect size for the Wilcoxon signed-rank test (r) can be calculated by dividing the Z-statistic by the square root of the sample size, see Equation A.4.

$$r = \frac{Z}{\sqrt(N_{obs})} \tag{A.4}$$

Where,
Z = Z-statistic
N_{obs} = sample size

The range for the effect size r for the Wilcoxon signed-rank test is from 0 to 1 and can be interpreted using Table A.1.

Table A.1: Wilcoxon signed-rank effect size

Effect Size	***r***
Small	0.10 – < 0.30
Medium	0.30 – < 0.50
Large	≥ 0.50

Cramer's v

A statistic that can be used to measure the effect size of the chi-square goodness-of-fit test is the Cramer's v. The square root of the chi-square statistic is divided by the sample size and the minimal dimension minus 1 to get Cramer's v. The formula for Cramer's v:

$$v = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{A.5}$$

Where,
χ^2 = Pearson chi-square statistic
N = the sample size
k = min(c - 1, r - 1), where r = rows, c = columns

The p-value determined by the Pearson's chi-squared test is the same one used to determine the significance of v. The Cramer's v statistic generates a value between 0 and 1 and can be interpreted using Table A.2.

Table A.2: Cramer's v effect size

Effect Size	***v***
Small	0.10 – < 0.30
Medium	0.30 – < 0.50
Large	≥ 0.50